A PAIR-WISE FRAMEWORK FOR COUNTRY ASSET ALLOCATION USING SIMILARITY RATIO

TAY SWEE YUAN

BSc (Hons) (Computer & Information Sciences), NUS

MSc (Financial Engineering), NUS

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF MATHEMATICS

NATIONAL UNIVERSITY OF SINGAPORE

Towering genius disdains a beaten path.

It seeks regions hitherto unexplored.

Abraham Lincoln

Acknowledgements

I would like to thank Dr Ng Kah Hwa, Deputy Director (Risk Management Institute, National University of Singapore) who encouraged me to take up the challenge of pursuing another Master's degree. He has also kindly volunteered himself to be my supervisor for my research project. His comments and feedbacks have been very invaluable to me.

My sincere thanks go to Dr Sung Cheng Chih, Director (Risk and Performance Management Department, Government of Singapore Investment Corporation Pte Ltd), for giving me his full support and for showing faith in me that I am able to cope with the additional commitments required for the Master's degree.

During the process of research work, I have benefited from discussions with colleagues and friends. I am especially indebted to Dr David Owyong for endorsing the approach for my empirical studies. My team mates in Equities Risk Analysis (EqRA) also deserve special thanks. The EqRA folks have to shoulder a lot more work due to my commitments to this research. I am glad to have them around and that allows me to work on my research without the worry that the team's smooth operations will be jeopardized.

This thesis would not have been possible without my family's full support and encouragement. My wife, Joyce, have to spend more time with the housework and the kids, especially during weekends, to let me work on this research; she is also constantly encouraging me and reminding me not to give up. The hugs and kisses from the five-year old Xu Yang, and the one-year old Xu Heng also never failed to cheer me up. Xu Yang's words truly warm my heart and keep me going, "Papa, I know you don't know. It's OK; just do your best *lah*. Don't give up *huh*."

Last but not least, those whom have helped me in one way or another, a big THANK YOU to all of you.

Table of Contents

Summary

There is no evidence that pair-wise modeling is widely applied in active portfolio management. Many forecasting models are constructed to predict individual asset's returns and the results are used in stock selection. To apply pair-wise strategies in portfolio construction requires forecasts of the relative returns of all possible pair-wise combinations in the investment universe. As the number of forecasts required is a small fraction of the total number of forecasts generated, it means a good measure of quality would be required to select the best set of forecasts.

In assessing the quality of relative returns forecast, the most important criterion is the accuracy of predicting the direction (or sign) of the relative returns. The quality of forecasts is reflected in the model's ability in predicting the sign. However, there is a lack of research and effort in designing a scoring measure that aims to quantify the forecasting model in terms of directional accuracy. The commonly accepted measure is Information Coefficient or the number of correct sign-predictions expressed as a percentage of the total number of predictions (Hit Rate).

This thesis presents a pair-wise framework to construct a country allocation portfolio and a measure called the Similarity Ratio as the confidence score of each forecasting model. In essence, the framework recommends that one should customize a model for each asset pair in the investment universe. The model is used to generate a forecast of the relative performance of the two assets, and at the same time, calculate the Similarity Ratio.

The Similarity Ratio is used to rank the pair-wise forecasts so that only forecasts with the best quality is used in portfolio construction. The Similarity Ratio is a distance-based measure that is innovative and intuitive. It emphasizes on directional accuracy and yet able to make use of the magnitudes of the forecasts as tie-breaker if the models have the same directional accuracy.

We provide extensive empirical examinations by constructing various country allocation portfolios using the pair-wise framework and Similarity Ratio. We show that the portfolios delivered better risk-adjusted performance than top quartile managers who have similar mandates. The global, European and Emerging Asia portfolios generated Information Ratios of 1.15, 0.61 and 1.27 respectively for the seven-year period from 2000 to 2006. We also find empirical evidences that show the portfolios constructed using Similarity Ratio out-performed all other portfolios constructed using other scoring measures, such as Information Coefficient and Hit Rate.

List of Figures

List of Tables

1 Introduction

There is no wide application of pair-wise strategies in active portfolio management. This is probably why this is a lack of effort in finding a scoring measure that quantifies the quality of models that forecast relative returns or direction of asset returns. The more commonly used approaches are Information Coefficient and Hit Rate, which is the number of correct sign-predictions expressed as a percentage of the total number of predictions.

This thesis presents a pair-wise framework to construct country allocation portfolio in a systematic and objective manner. The framework comprises of two parts, first it recommends contextual forecasting models should be built to predict asset pairs' relative returns, with emphasis on the sign-accuracy. Next it presents Similarity Ratio as an ideal scoring measure to quantify the quality of pair-wise forecasting models. Similarity Ratio is an innovative and intuitive measure that emphasizes on directional accuracy and yet able to make use of the magnitudes of the forecasts as tie-breaker if two sets of data have the same directional accuracy.

The focus of the research is not to find the best model that forecast relative returns or the best way to put these forecasts together. The emphasis is to build forecasting models to predict relative returns and the use of Similarity Ratio to measure the quality of such predictions. In this chapter, we start by examining the empirical results using perfect forecasts. We then review the problems related to pair-wise modeling that we have observed based on current practices and research. We conclude the chapter with the contributions of this research and outline of the report.

1.1 Interesting Results from an Empirical Study Using Perfect Forecasts

Before we go further, there are some interesting results that we observed with the use of *perfect forecasts*. Using perfect forecasts means we assume we have the perfect foresight of future asset returns, that is, we used the actual asset returns as our forecasts. While this is not realistic in real life, it does allow us to ignore the quality issue of forecasting models and focus on the drivers of portfolio performance. We fed the perfect forecasts into the Black-Litterman formula to obtain an expected return vector. This vector was then used together with Mean-variance Optimization to construct the test portfolios.

1.1.1 Portfolio Constructed Using Relative Returns Performed Better

We constructed two global country portfolios: one used perfect forecasts of the individual returns of each country in the benchmark universe, and the other used the relative returns of every country pair. We call these two portfolios the *Individual Model* and *Pair-wise Model* respectively. Both portfolios were benchmarked against the world equity index. The portfolios were held for three months and rebalanced at the end of the holding period. We tracked the portfolios performances for seven years:

- Individual Model's annualized value-added is 5.11% compared to the Pair-wise Model's 7.75%
- Individual Model's Information Ratio is 2.43 compared to the Pair-wise Model's 4.83
- Individual Model out-performed the benchmark returns 89.3% in the seven-year period. The Pair-wise Model out-performed the benchmark returns in every quarter over the seven-year period.

The results suggest that there are some merits in using a pair-wise approach to portfolio construction. The intuition behind is likely due to the fact that in constructing a portfolio with a given set of investment universe, it is the trade-off between each asset pair that helps to decide which asset to be allocated more weights and hence a good set of relative returns forecasts help to make such trade-off decision. This makes the pair-wise approach a more natural and intuitive portfolio construction approach.

1.1.2 Directional Accuracy Drives Investment Performance

The previous test suggests that portfolios which are constructed based on pair-wise returns are likely to generate better performances. In the next test, we want to test how much is the superior portfolio performance coming from directional accuracy of the relative returns forecasts. We constructed a *Perfect Direction* portfolio in which we preserved only the signs of the perfect forecasts. The magnitudes of the forecasts were set to be the historical average. For example, if the relative return is "predicted" to be positive, we will use the average positive relative returns over the previous five years as the magnitude. Like before, we put these forecasts into the Black-Litterman framework and used optimizer to find the weight of the holdings.

The results clearly point to the fact that it is "directional accuracy" that matters most:

- Perfect Direction delivers an annualized value-added of 7.23% as compared to the 7.75% of the Perfect Forecast.
- Perfect Direction's Information Ratio is 4.41 as compared to Perfect Forecast's 4.83
- Perfect Direction out-performed the benchmark returns every quarter for the seven-year period, just like the Perfect Forecast.

Considering the fact that the Pair-wise Model used *Perfect Forecasts* (i.e. perfect direction and magnitude) and only marginally better than the Perfect Direction portfolio, it is clear that bulk of the out-performance is actually driven by directional accuracy and not magnitude.

Impact of Forecasts' Magnitudes on Portfolio Performances

To see the impact of the magnitudes of the forecasts on the perfect direction portfolio, we constructed different perfect direction portfolios by varying the magnitudes of the forecasts. We started with a 1% relative performance and increase it to 50%. The annual value-added for this group of portfolios ranges from 6.40% to 8.20%, while the Information Ratio ranges from 3.91 to 5.15. We see that even the worst portfolio is still able to generate a performance that is about 80% of the perfect forecasts portfolio:

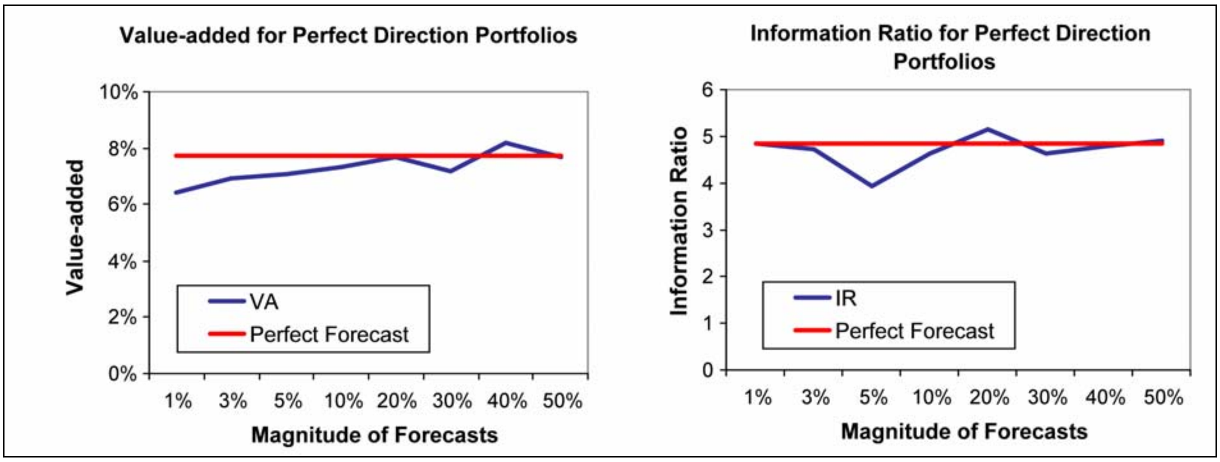

Figure 1-a: Impact of Magnitudes of Forecasts on Perfect Direction Portfolios

While this clearly supports our belief that it is direction accuracy of the relative returns forecasts that drives the portfolio performances, it also shows that magnitude of forecasts does have some impacts on the portfolio performance. The impact may be small but certainly not negligible.

1.1.3 Magnitude of Forecasts Determines Bet Size

Consider two pair-wise forecasts: (1) asset *A* to out-perform asset *B* by 10%, and (2) asset *C* to out-perform asset *D* by 1%. Assuming all other considerations are the same, including the level of confidence in the forecasts, the first forecast is more likely to result in a larger bet than that of the second since we expect it to generate a larger out-performance.

Hence we see that it is “direction” that determines the type of trades (*buy* or *sell*, *long* or *short*), it is the “magnitude” that determines the size of the trades. In other words, “direction” decides whether the trade ended up in profit or loss, and “magnitude” determines the size of the profit or loss.

1.2 Directional Accuracy Drives Investment Profitability

This simple empirical study using perfect forecasts highlight three important points:

1. Accuracy of pair-wise relative returns is more important than accurate returns of individual assets in the quantitative approach in constructing portfolio

2. Directional accuracy is the driving factor behind the performance of the portfolios

3. Magnitude of forecasts do have some values in determining portfolio performance

The key observation from the empirical study using perfect forecast is the importance of directional accuracy in portfolio construction. Numerous studies were also done to confirm the importance of predicting correct direction in the area of financial forecasting, for example, Yao and Tan (2000), Aggarwal and Demaskey (1997), Green and Pearson (1994) and Levich (1981) all supported the claim with empirical studies.

While most of these studies focus on the direction- or sign-prediction of individual assets, we believe that the focus should be the direction- or sign-prediction of the relative returns of asset pair. In constructing portfolio with the expectation to beat a benchmark, the ability to know which of any two assets will perform better is important, that is, the ability to pick the *winner* for any two assets. This is because it allows the investor to make the “trade-off” decision relating to the two assets. This decision is determined by the direction of the relative returns forecast.

While directional accuracy is the main driver for portfolio performance, we also see that forecast's magnitude should not be ignored completely. Thus we think that when assessing the quality of pair-wise forecasts, the ideal scoring measure should take into consideration the magnitude of forecasts. The pair-wise framework and Similarity Ratio are built with these two ideas as the underlying concept.

1.3 Observations from Current Practices and Research

1.3.1 Modeling of Individual Asset Return is not Necessary the Best Approach

Modeling of financial assets' expected returns has been the cornerstone of conventional quantitative approach to construct an equity portfolio. Researchers and practitioners aim to find the set of factors that best model the behavior of assets' returns. Individual forecasting models are constructed based on the selected set of factors to predict future returns for individual asset or market. Information Coefficient[1] (IC) is often used as a gauge to assess the quality of the forecasting model.

Intuitively, we know that it is the asset pairs' relative returns that help one to decide which asset to over-weight and at the expense of which asset. This was further confirmed by our empirical study using perfect forecasts. This raises doubts that if modeling of individual asset return is the best way forward in constructing a country asset allocation portfolio.

1.3.2 Pair-wise Modeling is Rarely Used in Portfolio Management

The idea of pair-wise strategies is to look at the assets a pair at a time. The most common form of implementing such strategy is the pair-trading of stocks, or relative value trading. Typically, the trader will first select a pair of stocks, for example, based on the stocks' co-integration, then *long* the security that (*he thinks*) will out-perform and *short* the other.

[1] IC is defined as the correlation coefficient between the forecasts and the actual returns over time.

There seems to be limited usage of the pair-wise strategies to construct active portfolio. Among the limited literature found, Alexander and Dimitriu (2004) constructed a portfolio consisting of stocks in the Dow Jones Index based on the co-integration of the index and the constituent stocks. However, this approach still does not take into consideration any potential relationship between the stock-pairs. Also, the scope of the implementation is restricted to Dow Jones' 30-stock universe.

Qian (2003) was another one who suggests the use a pair-wise strategy in tactical asset allocation. In his paper, he proposes the use of Pair-wise Information Coefficient (Pair-wise IC) as the means to influence asset weights. In the case where there is no significant Pair-wise IC, his model reverts back to use the Information Coefficient of individual asset model. However, there is no extensive empirical evidence provided on this pseudo pair-wise approach.

In general, pair-wise strategies are uncommon in active portfolio management and country asset allocation. This is possibly due to the lacking of a good scoring measure to help picks the right set of pair-wise relative returns forecasts as, and we will show, that the commonly used IC or Pair-wise IC may not work under a pair-wise framework.

1.3.3 No Known Scoring Measure that Emphasizes on Directional Accuracy

With successful pair-selection playing an important role in a pair-wise framework, it is important that we have a scoring measure that emphasizes on direction accuracy. Given the limited application, if any, of pair-wise modeling, it is not surprising to find that there is no scoring measure designed specifically for such purpose. The most commonly used measures are hit rate and distance-based measures (e.g. IC). Qian (2003) suggest Pair-wise IC but this measure also does not take into consideration the directional accuracy. In addition, it is susceptible to the existence of outliers, as we will show in 3.2.3. If the generally accepted Information Coefficient is not an ideal measure in the pair-wise world, what would be?

1.3.4 Regression-based Forecasting Model Commonly Used in Individual Model Construction

Regression is not a popular choice in academia research on forecasting model, and it is often used as a inferior alternative compared to a more complex approach such as neural network and support vector machine. Despite this, regression remains a popular approach in the finance industry because its simplicity makes it easier to interpret and explain the models.

We have used regression to generate the forecasts required for our empirical study. Although the focus of our research is not to find the best forecasting model, we want to study if the commonly used forecasting approach still works in a pair-wise environment. While regression model is commonly used in forecasting of individual asset returns, does it also work in the pair-wise framework?

1.4 Contributions of this Research

Against the backdrop of a lack of application of pair-wise strategies in active portfolio management, our research works provide:

- a generic framework to implement pair-wise strategies
- an innovative scoring measure that emphasizes on directional accuracy of relative returns forecasts
- a comparison of linear forecasting model built using robust regression against other classification techniques to predict signs of relative returns.

Forecasting model is often proprietary and highly guarded by investment houses. The model is a reflection of institutional research and creativity. We emphasize that the pair-wise framework does not depend on the type of forecasting models used to predict relative returns or the way to transform the forecasts into a portfolio. What is important is that we recommend that forecasting model should be built to predict relative returns, and use Similarity Ratio as a scoring measure to quantify the forecasting quality. Thus the framework allows investment firms to retain their respective competitive advantage in generating forecasts and portfolio construction while incorporating the pair-wise framework that we propose.

1.4.1 A Framework to Implement Pair-wise Strategies

The outline of the framework to generate pair-wise forecasts is as follows:

1. Identify a set of factors to use
2. Decides the criteria to be used to determine the predictive power of the factors
3. For each asset pair, choose the best factor and construct the model
4. When all the forecasts for the possible asset pairs have been generated, use Similarity Ratio to rank the pairs.
5. Select the pairs to construct the portfolio

We can see the framework as consists of the two stages:

- **Stage 1**: For each of the possible pair of assets, construct a contextual model that forecast the relative returns.
- **Stage 2**: Screen the best models based on Similarity Ratio and used the selected forecasts to construct the portfolio.

The results of the first stage will be the pair-wise forecasts for all possible combinations of two assets. In an *N*-asset investment universe, there will be $N \times (N-1)/2$ possible combinations. In order to construct a portfolio with a view on each asset in the investment universe, we will need only a maximum of $(N-1)$

pair-wise forecasts. Thus the pair-wise model leads to a large redundancy of forecasts. As the number of assets gets larger, more forecasts will be redundant. Thus how stage 2 being carried out will have a significant impact on the success of the pair-wise model. This hinges on the choice of a scoring measure to select the right set of pairs.

1.4.2 Innovative Scoring Measure that Emphasizes on Directional Accuracy

As acknowledged by Alexander and Weddington (2001), the "*selection process is perhaps the hardest but most important part*" in forecasting. We designed a scoring measure to quantify the forecasting quality of a model. The scoring mechanism embedded in Similarity Ratio uses directional accuracy as the main consideration when assigning a score, and supplement with the magnitude of forecast.

For an actual-forecast pair (a, f), Similarity Ratio is defined as:

$$\text{Similarity Ratio } (a, f) = \begin{cases} 0 & \text{if } a^2 + af \le 0 \\ \dfrac{|f+a|}{|f+a| + |f-a|} & \text{otherwise} \end{cases}$$

Similarity Ratio for a model will be the average Similarity Ratio for every actual-forecast pair generated from the model.

The intuition of Similarity Ratio can be seen geometrically; let's consider a Cartesian plane with the x-axis as the actual values (*a*) and the y-axis as the forecast (*f*) values, then each actual-forecast pair, (*a*, *f*), can be plotted on the Cartesian plane. The line $y = x$ contains all the points where the forecasts matched the corresponding observed values. The line $y = -x$ contains all the forecasts that are directly opposite of the observed values.

Intuitively, a point that is nearer to the $y = x$ line is *good* in terms of the accuracy of the forecast. On the contrary, if a point is near to the $y = -x$ line is *bad*. Thus we called the lines $y = x$ and $y = -x$ as the *Good* line and *Bad* line respectively. Conventional distance-based measures (e.g. Information Coefficient) rely on the projections to the Good line but we recognized that this alone is not sufficient. Similarity Ratio was derived based on the distances of the orthogonal projections of each actual-forecast pair to **BOTH** the *Good* and *Bad* lines.

Not only did the portfolios constructed using Similarity Ratio out-performed other scoring measures empirically, Similarity Ratio also exhibits important characteristics of an ideal scoring measure:

- Model with a higher accuracy in forecasting the direction of relative returns (hit rate) has a higher score.

- If there are two models having the same hit rate, higher score will be given to the model with more accurate forecasts, that is the deviations from actual values observed is smaller.

- The model's score is not susceptible to the presence of outlier in the sample.

We surveyed and tested various scoring measures such as Information Coefficient, Theil's Forecast Accuracy Coefficient, etc. We found that none of them meets all the criteria we have specified for the ideal measure. Thus we propose to use Similarity Ratio as the scoring measure for our pair-wise framework.

1.4.3 Comparison of Regression Model with Classification Techniques

We conducted extensive empirical study to find evidences to support that regression approach still works in the Pair-wise framework. The empirical results provide invaluable evidences to show that the portfolios constructed using regression-based forecasting models out-performed those constructed using classification techniques such as Neural Network, Discriminant Analysis, etc. In addition, this study also shows that it is possible to incorporate different forecasting models into the framework and this supports our claim that the framework is model-independent.

1.5 Outline of this report

The rest of this report is organized as follows:

- Describes the two stages in contextual pair-wise framework: (1) forecasts generation for all asset pairs and (2) pair-selection

- Reviews different scoring measures and provides a full description of the Similarity Ratio

- Presents the approach and implementation details for the empirical study to test the pair-wise framework and Similarity Ratio.

- Lists the evaluation criteria that will be used to measure the performances of the test portfolios

- Presents the results of the various empirical tests, including the performances of the model portfolios.

- Reviews the application of classification techniques in the pair-wise framework. Different classification techniques were used to construct test portfolios and the results were presented in this chapter.

2 Contextual Model in the Pair-wise Framework

The motivation behind the pair-wise framework is that we recognize it is the "directional accuracy" of an investment decision that determines its profitability. We feel that the best approach to obtain better "directional accuracy" in forecasting is to model relative returns instead of absolute returns. We propose the main objective of forecasting should be to predict the "direction" or sign of relative returns. Relative returns forecasts should not be a by-product of asset returns modeling, as in many of the forecasting models.

In addition, we found that each asset pair is different and that one should not be tempted to use a single model to forecast the relative returns for all asset pairs. Instead, one should have a contextual model for each pair. The pair-wise framework consists of two stages; first is to have a contextual model for each asset pair, the second stage is to select the best forecasts to be used in constructing portfolios. This chapter describes the contextual modeling and the framework. We also cover the details on how we implemented the two-stage framework for our empirical study.

2.1 The Need for a Contextual Model

2.1.1 What if there is no Contextual Modeling?

In constructing forecasting models, two questions that need to be answered are:

1. What factors should we use to construct the forecasting model for each asset pair?
2. How often should we review the efficacy of these models?

To answer these questions, one will need to do extensive back-testing. However, that does not mean that a set of factors that performed well during back-testing will translate to outstanding portfolio performance going forward. The problems could be due to data snooping, structural breaks, etc.

Even if the models do work during the initial period, one will need to decide on when the models should be reviewed. Had one waited for the model to break before reviewing, it may be too late as the damage to the portfolio performance may be too great by then. Without contextual modeling, it is analogous to having an "expiry date" for the models and yet no one knows when it is until it is expired – this may mean great financial loss.

Since no one knows when a model will actually "expired", it makes sense to consider it as only *good-for-one-use*, and that each time one needs a model to generate forecast, reconstruct the model building process. This is what contextual model proposes.

2.1.2 Empirical Study on Indicator's Predictive Power

To find empirical evident to support our claim of a contextual model, we made use *momentum-based indicators*[1]. We identified three sets of indicators commonly used in the publications related to momentum-based indicators:

- Short term indicators = 1-, 2-, 3-month returns
- Medium term indicators = 6-, 9-, 12-month returns
- Long-term indicators = 24-, 36-, 60-month returns

We classified "short term" and "medium term" indicators as "positive momentum", and we expect indices that exhibit these characteristics to continue their current trend. We considered "long term" as an indication that the market is positioned for a reversal in current trend. We expect the relationship between the reversal indicators and future returns to be negative hence we reversed the sign of the indicators.

[1] A detailed review of the momentum indicators can be found in the appendix

We denote the indicators as I_m, where $m \in \{1, 2, 3, 6, 9, 12, -24, -36, -60\}$. We use the negative sign to distinguish the momentum and reversal indicators. When referring to the indicator value for a particular time period, t, we add a subscript to it. That is, $I_{m,t} = sign(m) r_{t-|m|}$. For example, $I_{-60,t} = -r_{t-60}$ and $I_{3,T} = +r_{T-3}$, where $r_{T-t} = \frac{p_T - p_t}{p_t}$ and p_T is the price level at time T.

Definition of Predictive Power of Indicators

To determine if an indicator I_m is to be included as an input factor to the forecasting model, it needs to have explanatory power to predict the direction of the returns *h*-period ahead. We define the predictive power of the indicator as the number of times the indicator correctly predicted the direction and expressed it as a percentage to the total number of predictions made. We call this percentage *Classification Hit Rate* for the indicator, or simply *Hit Rate*.

The direction of the relative returns is indicated by the sign of the returns. A positive sign indicates an out-performance while a negative sign indicates an under-performance. Hence if the signs of two values, say *x* and *y,* are the same, their product will be greater than zero.

From a data and implementation perspective, we defined *Sign Test* as follows:

$$SignTest\ (x, y) = \begin{cases} 1 & if\ (x)(y) > 0 \\ 0 & Otherwise \end{cases}$$

Where $x, y \in R$

Lastly, we define:

$$\text{(Classification) Hit Rate for indicator } I_m = \frac{1}{T}\left\{\sum_{i=1}^{T} SignTest\left[I_{m,i}, r_{i+3}\right]\right\}$$

Where T = number of sample data

Hit Rate is simple, effective and intuitive. We believe that it is a good measure for the predictive power of the indicators. Some other names used for Classification Hit Rate are *Directional Change* (Refenes 1995), *Hit Ratio* (Huang et al 2005) and *Directional Symmetry* (Tay and Cao 2001).

Result 1: Indicator's Predictive Power Changes over Time

The historical hit rates of each indicator in predicting the direction of the relative returns of the EU/ AP pair are plotted in Figure 2-a.

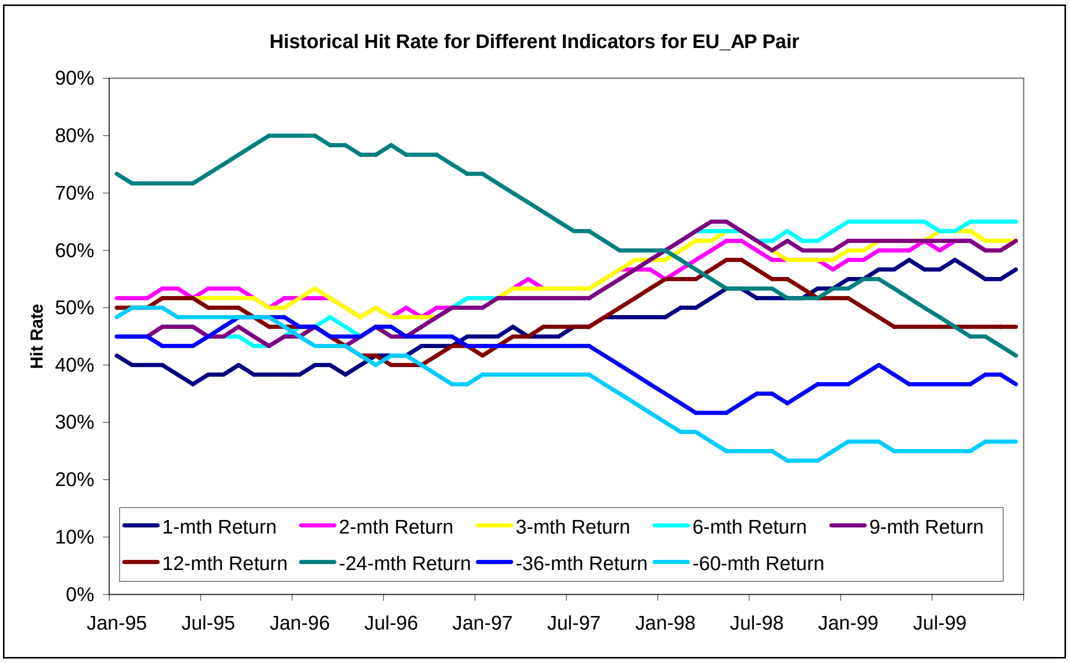

Figure 2-a: Historical Hit Rates for Various Indicators for EU_AP Pairs

We see that while the indicators generally stay relevant and effective for a certain period, the length of the periods vary from indicator to indicator. For example, we see that the predictive power of *I*-$_{24}$ diminishes over time while I_6 gets better in later part of the sample period. This suggests that having a fixed period to review the efficacy of the indicators may not be able to capture failing indicator, or include the more effective indicator in the forecasting model. This study supports the need to have a contextual model where the most appropriate indicators are used at the time when investment decision is to be made.

Having a fixed period to review the efficacy of the indicators in forecasting is not necessary a good idea. In fact, imagine an investor did some back-testing at end of 1996 and found that I_{-24} to be an excellent indicator to use in constructing factor model. If he had not review the predictive power of the indicators regularly, he will be hurt by the poor performance of the indicators.

Result 2: Different Indicators Work for Different Asset Pairs

There is no universal set of indicators that is suitable for all the asset pairs. Figure 2-b shows the historical hit rates of I_6 for all asset pairs. While it is consistently a good indicator in predicting the direction of North America's returns relative to Japan, it is obviously not a suitable candidate for the JP/AP pair:

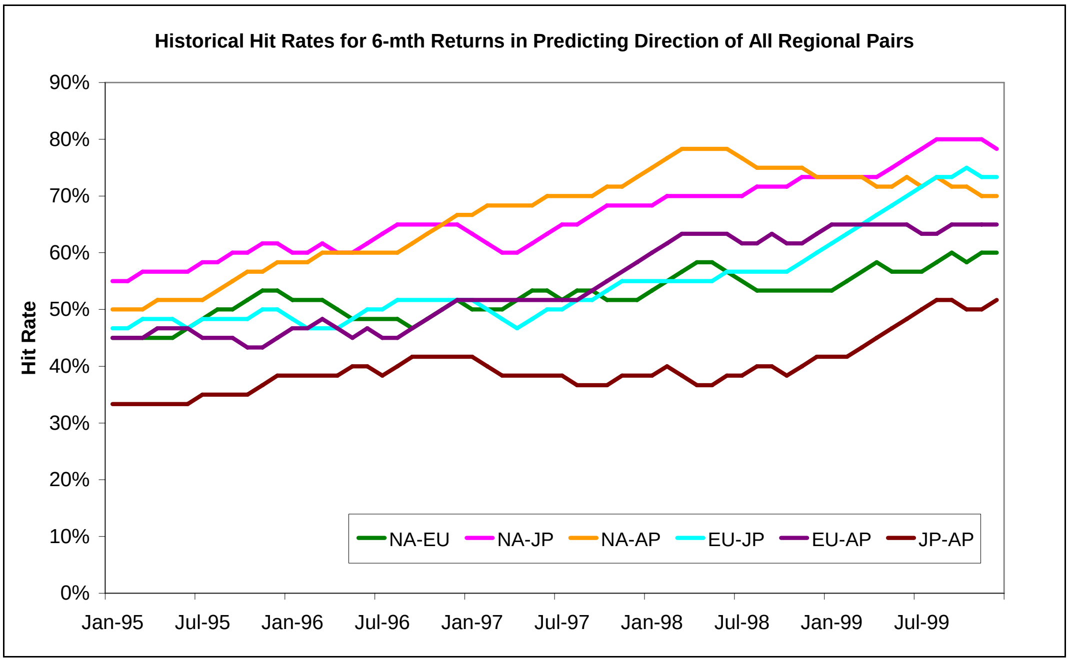

Figure 2-b: Historical Hit Rates for 6-mth Returns in Predicting Direction of All Regional Pairs

This brings us to the point where we need to model each asset pair independently. For each asset pair, choose the set of indicators that has the best descriptive power of the pair's relative returns, and use these indicators to construct the forecasting model.

2.1.3 Contextual Model Uses the Most Appropriate Set of Indicators for Each Asset Pair

From our preliminary study, we observe the following characteristics of the momentum and reversal indicators:

1. The indicator's predictive power changes over time. There is no one indicator continue to stay the best over a long period

2. Each asset pair is unique and hence different indicators are needed to model the pair's relative returns

Contextual model requires that one always applies the indicators that are most appropriate at the decision time. Every previous model is discarded each time a forecast is to be generated and re-start the models construction process and used data available up to that point in time. By doing this, there is no "expiry date" to each model. This makes the models less susceptible to data snooping and manipulation. More importantly, this can become an on-going process without having to set a review period to re-test the efficiency of the indicators.

2.2 Pair-wise Framework is a Two-stage Process

In our pair-wise framework, we place emphasis on the modeling of relative returns to determine the winning asset of each possible pair-wise combination. We recommend using the most appropriate set of indicators available at decision time to construct the forecasting model. A set of evaluation criteria should be used to identify the indicators that best explain the pair-wise relative returns.

We mentioned earlier that with the set of factors and the criteria to determine the predictive strength of these factors decided, the pair-wise framework is a two-stage process:

1. Generate forecasts on the relative performance for each possible pairs of assets in the investment universe

2. Select the right set of views to be included in the market view

In Stage 1, a contextual model will be built for each asset pair. All the forecasting models will undergo a series of fitness tests and only those deemed “fit” enough will be used in forecasting. Stage 2 of the pair-wise model is to determine among the fit models, which are the ones that should eventually be fed into the optimizer. The next two sections describe both stages. By detailing the implementation considerations we have adopted in our research also shows how the framework can be implemented in an investment process.

2.3 Stage 1 – Build Contextual Model for All Possible Pairs

The pair-wise framework proposed a contextual model to be constructed for each asset pair. The following steps outline the approach in which the contextual models are constructed:

- Select the indicators to use

- Construct forecasting model to predict relative returns

- Validate the model

- Computes the confidence score for the model

2.3.1 Select the Indicators to Use

One needs to decide on a set of evaluation criteria to select a set of indicators. The important point to note is that in the selection of indicators, the emphasis is to choose the set of indicators that best predicts the direction of the pair-wise relative returns. In our implementation, we have been using hit rate as a measure but we don’t rule out other possible evaluation criteria.

For each of the indicators, we compute the hit rate of its accuracy in predicting the winner. We used five-year monthly data, which gives us a 60-period sample size. Recalled that our set of indicators range from 1-mth to 60-mth price returns; using a 60-period sample size effectively requires we have at least 10-year of monthly data. In addition, we have also reserved seven-year of data for out-of-sample empirical studies, thus we are using a full 17-year of data for our analysis.

The number of data points needed in a regression study depends on the size of pool of potentially useful explanatory variables available. Kutnet et al (2005) suggest that the rule-of-thumb in deciding the sample size and number of variables is that there should be at least 6 to 10 cases for every variable in the pool. To ensure that we have sufficient data points to capture the relationship between the predictor and the dependent variable, we recommended 20 points per variable. With an in-sample size of 60, we used Kutnet’s rule-of-thumb to back out the maximum number of independent variables we can have: $60 \div (2 \times 10) = 3$. However, we will only consider indicator with a hit rate of at least 50%.

2.3.2 Construct a Forecasting Model

Pair-wise framework does not specify the forecasting approach to use. What it does specify is that the forecasting models should emphasize on sign-prediction of asset’s relative returns. In our implementation, we build a linear model and estimate the factor loadings using robust regression. We have also explored other forecasting techniques, such as Neural Network, Discriminant Analysis, etc. The empirical results of these classification techniques can be found towards the end of this report. Here we focus only on the regression model.

Supposed the indicators selected are I_1, I_3 and I_6. Then the linear model is as follows:

$$\hat{Y} = \beta_0 + \beta_1 I_1 + \beta_2 I_3 + \beta_3 I_6$$

Where

$\hat{Y} =$ Forecast of relative returns

Robust Regression is used to Estimate Factor Loadings

Financial markets can experience sudden moves and deviate from the structure observed from the past data. This results in outlier data which affects the effectiveness of the regression model. The figure below displays an error bar plot of the confidence intervals on the residuals from a least squares regression of Japanese Yen one-month volatility forecast with historical and implied volatilities as the predictors. The plot represents residuals as error bars passing through the zero line. The outliers, those that lie outside the 95% confidence intervals, are marked in red.

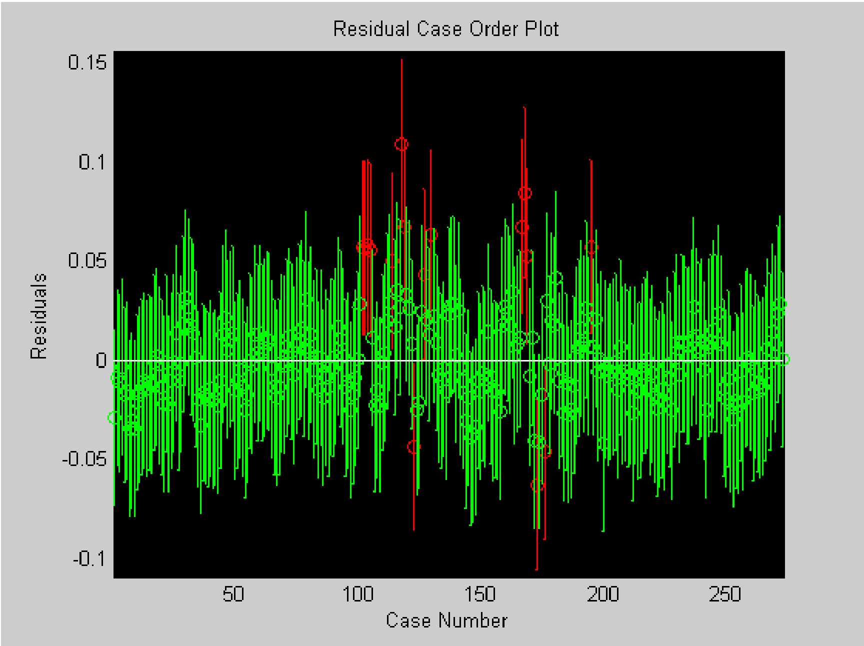

Figure 2-c: Error bar plot of the confidence intervals on the residuals from a least squares regression of daily FX returns

To down-weigh the undesired impact of these outliers in the observations, robust regression is used to estimate the regression coefficient of the contextual model. Robust regression techniques are an important complement to classical least squares regression. While least squares regression weights all observations equally, a robust method applies a weighting function to reduce the impact of outliers in the observations.

Robust regression techniques available are least trimmed squares (LTS) regression, least median squares (LMS) regression, least absolute deviations (L1) regression, and M-estimates of regression. The different regression techniques are briefly described in the appendix (for a more detailed discussion, refer to Draper and Smith 1998). In this paper, Andrew's wave function is used for the M-Estimation, with the tuning constant set to 1.339.

2.3.3 Validate the Model

Each constructed model is put through a series of *fitness* tests to ensure that it is stable and that the forecast generated is good for use. This *quality assurance* process will reject a model if it failed any of the tests.

The list of statistical tests used to determine the model's fitness is as follows. The statistical tests are conducted at a 95% confidence level:

- Test the stationarity of dependent variable
- Test the significance of the regression coefficients
- Test the normality and independence of model residuals

Test the Stationary of Dependent Variable

The foundation of time series analysis is stationarity, either *strictly* or *weakly* stationarity. A process that is weakly stationary implies that the time plot of the data would show that the values fluctuate with constant variation around a constant level. Because it only requires the first and second moments to be constant with respect to time, weak stationarity is also called *covariance stationarity*. In financial modeling, it is common to assume asset returns to be weakly stationary, for example, Balvers and Wu (2005).

Even if weak stationarity is an accepted assumption, Tsay (2002) recommends one should check this assumption empirically to ensure forecasting model's stability. We implemented the KPSS[1] test, as Pretorius et al (2007) did, to ensure that relative returns series that violates the stationarity assumption will not be included.

[1] Kwaitowski et al (1992).

In the KPSS test, we compute the test statistics, $t = \frac{1}{\sigma^2 T^2} \sum_{t=1}^{T} \left[\sum_{j=1}^{t} \left(r_i - \bar{r} \right) \right]^2$

Where

r_i = Asset pair relative returns for time period *i*

σ^2 = Estimate of the long-term population variance of the asset pair relative returns

$\bar{r}$ = Mean asset pair relative returns

T = Number of data period

Intuitively, the sequence of relative returns measures recursively the volatility of the relative returns series. High values of the statistics imply that the estimated volatility is unstable, an indication of non-stationarity.

Test the Significance of the Regression Coefficients

The regression coefficients were tested with the null hypothesis that they are zero against the alternative hypothesis that they are not. The t-stats for each regression coefficient are being calculated and only models with significant predictors are considered as suitable.

Test the Normality and Independence of the Regression Residuals

For regression model, the error terms are assumed to be independent and normally distributed, with zero mean and constant variance σ^2. We apply the following tests to the residuals of each model:

- Jarque-Bera test – test for normality
- t-test – test for zero mean
- Durbin Watson test – test for independence

2.3.4 Generate Confidence Score for the Model

The quality of a forecast is dependent on the quality of the forecasting model that generated the forecast. Qian (2003) correctly pointed out that information gained from the forecasting process should be used in the portfolio construction process. This can be achieved if the scoring measure is constructed based on the out-of-sample track record of the forecasting models. The best way to assess the quality of a model is indeed the model's track record of out-of-sample forecasting accuracy. Based on the track record, which is a series of model's predictions and the corresponding observed values, we construct a confidence score that is ranges from "0" to "1", indicating "completely no confidence" and "perfectly certain" respectively.

Methods of Data Partitioning

We shall call the data used for obtaining this track record the *semi*-out-of-sample. In our implementation, we considered a few options in data partitioning:

Cross-validation

Cross-validation is the most common way of partitioning the data set into in-sample and *semi*-out-of-sample. In its most elementary form, cross-validation consists of dividing the data into *m* samples. Each sub-sample is predicted via the classification rule constructed from the remaining (*m-1*) sub samples, and the set of predictions and the corresponding observations formed the semi-out-of-sample set. Stone (1974) describes cross-validation methods for giving unbiased estimates of the error rate. As for what should m be set to, it will depend on the context of work, for example, Qian and Rasheed (2007) set *m* to 5 when building multiple classifier for stock prediction.

"One-Shot" Train-and-test

The commonly seen procedure of splitting the sample data into *training* and *testing* set is a special case of cross-validation with *m* = 1. Michie et al (1994) termed this as the "one-shot" train-and-test.

This procedure requires large samples. Michie et al (1994) suggest that one should only use this method if they have more than 1000 data points in the training set.

Leave-one-Out or Holdout Method

At the other extreme of cross-validation, *m* can be set to the size of the sample set minus one. This is the *leave-one-out* method of Lachenbruch and Mickey (1968), or the *Holdout* method described in Rencher (2002). The Holdout method is an improved version of the sample-splitting procedure. In the holdout procedure, all but one observation is used to compute the classification rule, and this rule is then used to classify the omitted observations.

Generating the Confidence Score

We adopted the Holdout method so that we can make full use of the sample data we have. Supposed we need to construct a model at decision time T (that is, the time period where the decision is to be made), our sample period is from $(T - S - h + 1)$ to $(T - h)$,

Where

S = number of sample points

T = decision time

h = holding period, or look-ahead (forecast) period

Using the Holdout method, we exclude one set of data values (which include the actual value of dependent variable and a set of values for the independent variables) and use the rest of the data points to estimate the coefficients. If there are a total of S sample points, then the coefficients are estimated using the rest of the $(S-1)$ data points. The indicator values from the excluded point were use to generate a forecast, and together with the actual observed value for the dependent variable in the first data period, we obtained the one forecast-actual pair. The procedure is repeated until each data point had been excluded once, and at the end of the procedure, we have generated *S* pairs of forecast-actual values. These data are then used to calculate the confidence score of the final model. To generate the final forecast, all the sample data are used to construct the final model. Throughout the whole process, the same set of momentum and reversal indicators is used.

2.4 Stage 2 – Select the Pair-wise Forecasts to Use

2.4.1 Probability of Selecting Right Pairs Diminishes with Increasing Number of Assets

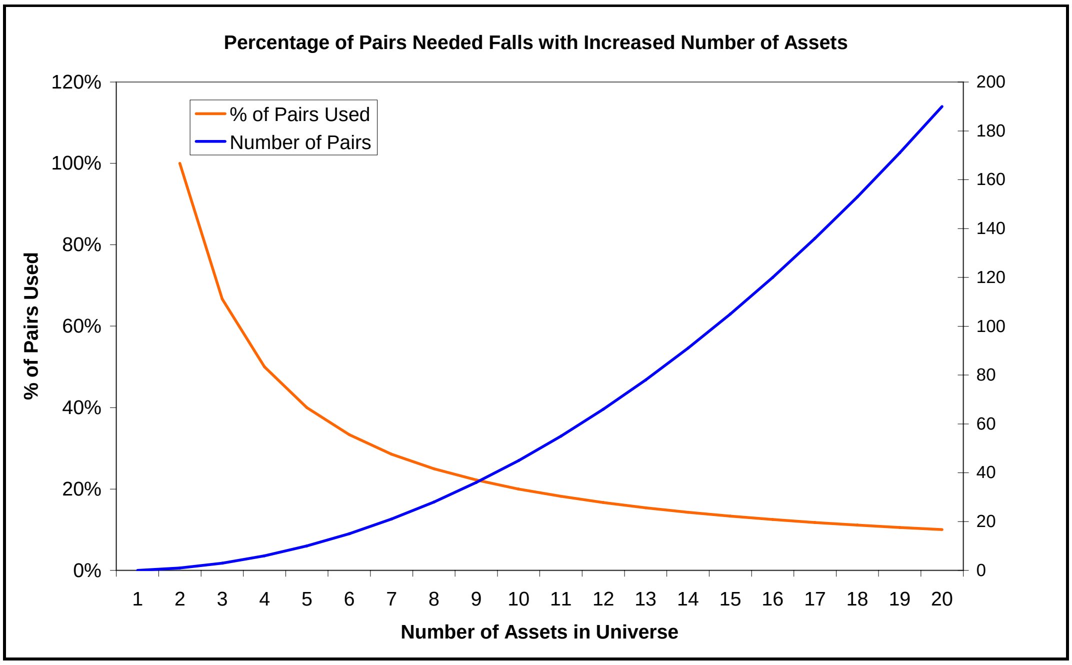

Figure 2-d: Pair-wise Forecast Usage Level for Different Number of Assets in Universe

The maximum number of views needed in an *N*-asset investment universe is *N-1*, out of the possible $N \times (N-1)/2$ pairs. The number of pairs increased exponentially with every additional asset in the universe. This difference in *growth* rate means that sieving the correct views becomes more difficult as the number of assets in the universe increases. Figure 2-d shows that while the number of pairs increased exponentially with each additional asset, the proportion of these pairs needed to formulate the market view decreased exponentially as well. For example, for a 10-asset investment universe, we need only about 20% of the number of pair-wise views generated. In other words, we have a 20% chance of picking the right pairs if we pick them randomly.

MSCI World has 23 countries, and MSCI Europe is formed by 16 countries – this means that our chances of picking the right pairs for our country allocation portfolios is about than 10%. This highlights the importance of having a good scoring measure that allows us to rank the relative returns forecasts for all the pairs.

2.4.2 Pairs Selection Consideration and Algorithm

Pairs Selected to Avoid Cycles

The selection process is done based on the confidence score of each view. We start screening the forecasts from the highest to the lowest. The screening process stopped if any of the following conditions is met:

1. All of the assets in the investment universe are being *represented*.
2. There is no more forecast that has a score higher than 0.5

Each forecast is the relative returns of one asset over another asset, thus we say that an asset is being *represented* if it is one of the two assets that were in a selected forecast. Supposed *L* contains the list of assets after each iteration of screening, then if the pair of assets in the current iteration is not a sub-set of the set *L*, it means that at least one of the two assets are not an element of set *L*, thus we can include the asset in set *L*. By doing so, we avoid having "cycles" or "deadlock" in our market view. Having cycles would make the market view meaningless. For example, consider the three views are those with highest confidence scores:

- View 1: Asset A out-performs Asset B
- View 2: Asset B out-performs Asset C
- View 3: Asset C out-performs Asset A

In addition, we impose a hurdle rate of 0.5 (that is, 50% confidence level) for the confidence score to prevent our process from including non-contributing forecasting model. However, this would mean not all assets in our investment universe will have a view representing its relative performance to another asset.

Algorithm to Select Pairs

The following algorithm outlines the selection process to select the set of forecasts that is required to represent the views of the relative performance of all the assets:

```
L = Empty Set
U = Set of assets in the investment universe
V = Array of forecasts sorted in descending order of confidence score
W = Empty Set
P = Array of asset pairs with corresponding score in V
FOR each view w in the array V
      IF L = U THEN
            EXIT FOR
      END IF
      P = the two assets in the pair in view w
      IF Score of w ≥ 0.5 THEN
            IF P ⊄ L THEN
                  L = L ∪ P
                  W = W ∪ w
            END IF
```

```
        END IF
    END FOR
```

At the end of the process, set *W* contains the list of views that were selected to form the market view, and set *L* contains the list of assets from the investment universe that are being represented in the views set *W*.

Minimum Number of Assets in the set of Selected Views

As one can see from the algorithm above, there is no requirement that there must be at least one view or forecast to be selected. In the scenario where no view has a confidence score that is higher than 50%, it will result in an empty set being return. There is no part of the algorithm that enforces all the assets must be represented in the set of views selected -- again the selection is determined by the confidence score.

2.5 Critical Success Factor to the Pair-wise Framework

The contextual modeling stage of the pair-wise framework generates relative returns forecasts for all the possible combinations of asset pairs in the investment universe. As we do not require all the forecasts, hence we require the second stage to select the best set based on our confidence to all the models.

As we based our selection on the confidence score, the implicit assumption is that the confidence score is a good reflection of the quality of the forecasting model. By selecting the views generated by models with highest score possible is selecting those best views that we can use to formulate our market view. The reliance on this assumption makes the choice of scoring measure the most important part of this research. The next chapter is devoted for the derivation of the scoring measure used in the pair-wise process.

3 Similarity Ratio Quantifies Forecast Quality

We have seen the importance of having a good scoring measure in the pair-wise model to select the right set of pair-wise forecasts. This score reflects our confidence in the forecasting model, and hence our belief of the likelihood of the forecast being correct. This chapter reviews some potential candidates for the scoring measure. We then proposed a new measure that was based on comparing the distances of two orthogonal projections. We called this innovative measure Similarity Ratio, and we will present the full derivation of this measure in this chapter.

3.1 Scoring Measure for a Forecasting Model

3.1.1 Assessing Quality of a Point Forecast

Suppose there are three models and each forecasted the performance of asset *A* relative to asset *B* for a holding period of three-month ahead:

Model	Relative Returns Forecast	Interpretation
1	6%	Asset *A* out-performed asset *B* by 6%
2	3%	Asset *A* out-performed asset *B* by 3%
3	-0.5%	Asset *A* under-performed asset *B* by 0.5%

At the end of the holding period, we observed that the asset *A* out-performed asset *B* by 1%. With this information at hand, we can assess the forecast quality of each of the three models. Clearly models 1 and 2 had done better than model 3 because both models correctly predicted that asset *A* will out-perform asset *B*. If we have to choose between the two models then we will have to compare to see which forecast is closer to the actual value observed. Model 2's forecast of 3% out-performance is much closer to the actual value of 1%, as compared to model 1's forecast, thus it is clear that model 2's forecast is more accurate.

Let's look at the differences in the magnitude between each forecast and the observed value. Model 1's difference is 5%, which is more than three-times the difference of model 3's forecast and actual value. But we know that trade executed based on model 1's forecast will result in profit of 1% simply because it had correctly predicted the winner of the two assets.

Trade executed based on model 3's forecast will result in a loss of -1.5%. However, the difference in magnitude suggested that model 3 is better than model 2 since the difference between the forecast and actual of 1.5% is less than 2% of model 2. This is the main reason why most distance-based measures do not work well in the pair-wise framework is that the focus is on measuring the magnitude of deviation and not directional accuracy.

The above observations lead to the following conclusions on assessing the quality of forecast:

- Correctly predicted direction is more important than having small deviation between forecast and actual value but with the direction wrong.

- The size of the difference between forecast and actual value can be used to distinguish two forecasts of the same direction (e.g. both correctly predicted the direction of the relative returns or both got it wrong)

3.1.2 Assessing Quality of a Collection of Point Forecasts

To assess the quality of a forecast model, we can look into the track record of the model's out-of-sample forecasts. From the forecast/observations pairs (samples) pooled over time we can get a reliable verification statistics. This can be done by treating the sample as a whole and study the statistics of the sample, for example, correlation between the forecast and observed values. Alternatively, one can also get pooled results by aggregating verification statistics over each sample point.

Scores that Treats the Sample as a Whole can be Skewed by Outliers

The problem with treating the sample of actual-forecasts pairs as a whole to obtain a statistical measure, e.g. correlation, is that it can be skewed by the presence of outlier. This is illustrated in Figure 3-a where the scatter plots were plotted using exactly the same set of data except for one point. The correlation for the first sample (the scatter plot on the left hand side) is 0.13. The presence of an outlier moved the correlation to 0.65, as can be seen from the upward trending line in the scatter plot on the right hand side.

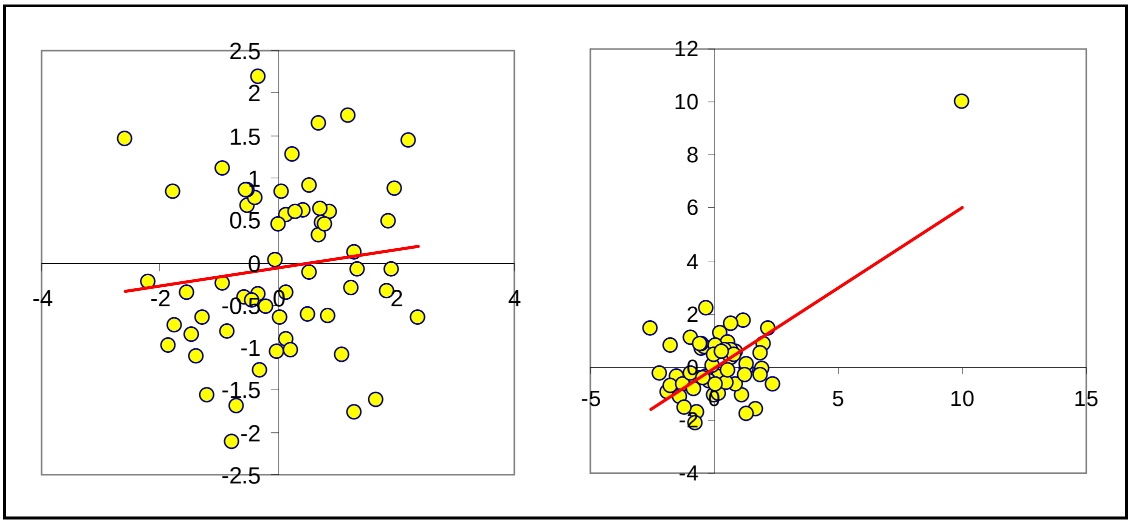

Figure 3-a: Illustration of the Impact of Outlier had on Correlation

Averaging Individual Score Not Affected by Outliers

Suppose we are able to generate a score between 0 and 1 for each of the actual-forecast pairs, and obtained the score for the sample by averaging the score of each point, we are able to by-pass the problem of outliers. For illustration, let’s assume we have a 60-pair sample, each with a score of 0 or 1. The average score of the sample increases linearly with the number of perfect scores present in the sample:

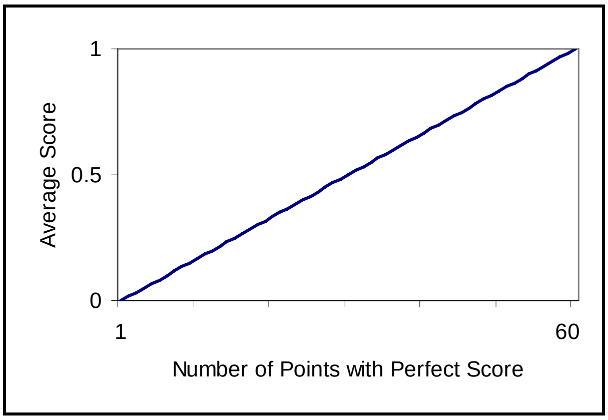

Figure 3-b: Illustration of Number of Points with Perfect Score had on Average Score

Observations

From the above, we see that for a scoring measure to be less susceptible to the presence of outliers, it would be necessary for the scoring measure to be able to treat each sample individually. To obtain an aggregate for a set of samples, the best way is to get a simple average of individual sample's score.

3.1.3 Properties of an Ideal Scoring Measure

The ideal scoring measure that quantifies the quality of a forecasting model should have the following properties:

- Single score that is range bound between 0 and 1.
- Model with a higher accuracy in forecasting direction, that is has a higher hit rate, should have a higher score.
- If two models have the same hit rate, higher score will be given to the model with more accurate forecasts, that is the deviation from actual value observed is smaller.
- Able to score each pair of forecast/ observed values.

- The model's score is not susceptible to the presence of outlier in the sample.

Based on the above, we identified three evaluation criteria for our scoring measure:

- P1. Measures directional accuracy of relative returns forecasts.

- P2. Uses magnitude of forecast only if directional accuracy of direction is not sufficient in distinguishing the quality of the two models

- P3. Has an aggregated score that it is not susceptible to outlier

This set of criteria will be used for us to evaluate the suitability of each candidate of the scoring measure. Ideally, we would want to have the scoring measure to be able to incorporate the above properties into a single score.

3.2 Review of Currently Available Scoring Measures

Now that we have identified the properties of an ideal scoring measure, we can start our search for potential candidates. Forecasting is a common tool used in many areas, and there are different measures used for measuring the forecast quality. The most common measure in the financial industry is the Information Coefficient, which is the correlation between the actual and forecast series. We also looked beyond financial industry and we found possible candidates such as Theil's Forecast Accuracy Coefficient (sales & marketing research) and Anomaly Correlation (climatology). In subsequent sections, we state the definition of these measures and briefly describe their strengths and weaknesses.

Notation

f_i = Forecast for i^{th} sample

a_i = Actual observed value for i^{th} sample.

Φ = A collection of actual-forecasts over the test period, $\{(a_1, f_1), (a_2, f_2), \ldots, (a_T, f_T)\}$

T = Size of the data sample, or number of (a_i, f_i) pairs.

For simplicity, we will drop the subscript *i* if it is not necessary, that is, we will use (a, f) instead of (a_i, f_i).

3.2.1 R^2

Alexander (2001) points out that statistical-based measure tends to favor models that are more stable and have less wild fluctuations. However, more stable predictions usually mean more rational forecasts to the investors and this helps to make them feel more confident. For this reason, we included a candidate from this family of measures in our study as a means for comparison even Alexander (2001) cautions that using it as the sole evaluation criteria in unlikely to distinguish the winning model.

Since regression is used to estimate the coefficient of our forecasting models, we have selected the goodness-of-fit measure, R^2, as one of the candidates. It is commonly used in regression analysis and it has a range of 0 and 1, which matches that of the confidence score which we are looking for. Thus we can set the score to the value of R^2:

Score = R^2

3.2.2 Hit Rate

Hit Rate is simply the ratio of right "hits" to the total number of predictions made. In our context, it measures the accuracy of a model's ability to forecast the direction of the relative returns. This is the most important criteria and certainly it meets property P1 of our ideal scoring measure. There are some other fancy names given to Hit Rate, such as *Fraction Same Sign* (Wolberg 2000), *Hit Rate* (Huang et al 2005) and *Directional Symmetry* (Tay and Cao 2001). Throughout this paper, we will use the term "Hit Rate".

Recalled that we have defined *SignTest* as follows:

$$SignTest\ (x, y) = \begin{cases} 1 \quad if\ (x)(y) > 0 \\ 0 \quad Otherwise \end{cases}$$

Then, we can define Hit Rate as follows:

$$\text{Hit Rate} = \frac{1}{T}\left\{\sum_{i=1}^{T} SignTest[a_i, f_i]\right\} \in [0,1]$$

Since Hit Rate is already ranges between 0 and 1, thus we can set the confidence score to be the Hit Rate itself:

Score = Hit Rate

While Hit Rate measures the correct direction, it completely ignores the differences between the magnitudes of forecast and actual values. This means that it does not measure the quality of the model in terms of the size of the directional bet, that is, it does not satisfy property P2. Wolberg (2000) also highlighted this weakness of Hit Rate and points out the fact that Hit Rate treats values that are close to zeros the same way as it treats those that are far from the zeros may undermine the usefulness of Hit Rate for non-dichotomous variables.

Refenes (1995) shares the same concerns and observes that it is relatively easy to obtain a high value in a trending market, thus suggested that this statistics should be used with care, or normalize it with its standard deviation over at least 30 test runs with different cross-validation sets and/ or initial conditions.

3.2.3 Information Coefficient (IC)

Information Coefficient (IC) is the most commonly used metric in assessing the forecast quality of financial models and managers' performance. For example, Guerard (2006) applies IC and regression on consensus analysts forecast and revisions in selecting stocks in Japan and US. Kung and Pohlman (2004) recommended using IC to assess managers' abilities when it comes to exploring alpha opportunities. Qian (2003) termed the correlation coefficient between the differences of two forecasts and the difference of two actual returns as Pair-wise Information Coefficient. This is consistent with our definition hence we will simply use the term "Information Coefficient" in this thesis.

IC is defined as the correlation between the observed values and the forecast.

$$\text{IC} = \frac{\sum_{i=1}^{T}\left(f_i - \overline{f}\right)\left(a_i - \overline{a}\right)}{\sqrt{\sum_{i=1}^{T}\left(f_i - \overline{f}\right)^2 \sum_{i=1}^{T}\left(a_i - \overline{a}\right)^2}} \in [-1,1]$$

Where $\left(a_i, f_i\right) \in \Phi$.

Because correlation coefficient ranges from -1 to 1, we need to normalize IC to range between 0 and 1 using the following transformation:

Score = (IC + 1) ÷ 2

Despite the fact that IC is commonly cited in financial modeling application, Wolberg (2000) points out that high degree of correlation may not necessary imply the predicted values are useful. This is particularly critical in the pair-wise framework where directional accuracy is the main driver for portfolio performance and yet not being catered for in IC. In addition, we have also seen that IC can be skewed when there is outlier, as illustrated in Figure 3-a.

3.2.4 Un-centered Information Coefficient (UIC)

One possible improvement to the IC-based score is to use un-centered correlation. UIC is defined as the un-centered correlation between the observed values and the forecast.

$$\text{UIC} = \frac{\sum_{i=1}^{T}(f_i)(a_i)}{\sqrt{\sum_{i=1}^{T}\left(f_i - \overline{f}\right)^2 \sum_{i=1}^{T}\left(a_i - \overline{a}\right)^2}} \in [-1,1]$$

Where $(a_i, f_i) \in \Phi$.

With the denominator always greater than zero, the sign of each pair is solely determined by the numerator. This means that pairs with forecast and actual are of the same sign will contribute positively to the score while those with different signs will subtract from it, this satisfies property P1.

To convert UIC to a score that ranges between 0 and 1:

Score = (UIC + 1) ÷ 2

3.2.5 Anomaly Information Coefficient (AC)

The anomaly correlation coefficient has been widely used to verify model predictions in numerical weather forecasting models in meteorological centers (e.g. Miyakoda et al 1986; Tracton et al. 1989; Palmer et al. 1990). The anomaly correlation coefficient is defined as

$$AC = \frac{\sum_{i=1}^{T}(f_i - c)(a_i - c)}{\sqrt{\sum_{i=1}^{T}(f_i - c)^2 \sum_{i=1}^{T}(a_i - c)^2}}$$

Where *c* represent the climatological values for the points of a grid at a given time (Radok and Brown 1993) and $(a_i, f_i) \in \Phi$.

To apply AC in our context, we set *c* = 0. The resulting formula is similar to the *Un-centered Information Coefficient*. We define Anomaly Information Coefficient as follows:

$$AC = \frac{\sum_{i=1}^{T}(f_i)(a_i)}{\sqrt{\sum_{i=1}^{T}(f_i)^2 \sum_{i=1}^{T}(a_i)^2}} \in [-1,1]$$

Where $(a_i, f_i) \in \Phi$.

Wolberg (2000) called this special case the *Coefficient through the Origin* (CCO).

Like IC, AC also ranges from -1 to 1. Hence we apply the same conversion approach:

Score = (AC + 1) ÷ 2

3.2.6 Theil's Forecast Accuracy Coefficient (UI)

Forecasting plays an important role in the area of sales and marketing research, and one measure that is commonly used to assess the quality in marketing forecasts is the Theil's Forecast Accuracy Coefficient. It was proposed by Theil in the *Economic Policy and Forecast*:

$$UI = \frac{\left[\frac{1}{n}\sum_{i=1}^{T}(a_i - f_i)^2\right]^{1/2}}{\left[\frac{1}{n}\sum_{i=1}^{T}(a_i)^2\right]^{1/2} + \left[\frac{1}{n}\sum_{i=1}^{T}(f_i)^2\right]^{1/2}} \in [0,1]$$

Where $(a_i, f_i) \in \Phi$.

In case of perfect forecast, $a_i = f_i$, for all i, UI = 0. In the event of "maximum inequality" or "worst forecast", we have UI = 1. To map UI to our confidence score, we reverse UI so that we get a "1" for perfect forecast and a "0" for anything worse than "maximum inequality":

Score = 1 – UI

The main problem with Theil's Forecast Accuracy Coefficient is the same as with that of Information Coefficient (IC) – it does not take into consideration the directional accuracy of the forecasts.

3.2.7 Who is the Winner?

Recalled that our evaluation criteria tell us that a scoring measure is considered better than another if it

- P1. Measures directional accuracy of relative returns forecasts.
- P2. Uses magnitude of forecast only if directional accuracy of direction is not sufficient in distinguishing the quality of the two models
- P3. Has an aggregated score that it is not susceptible to outlier

We mapped the characteristics of various measures against the desired properties that we have specified for an ideal scoring measure but we see that there is no one measure that satisfies all three properties:

Scoring Measure	**Range of Original Definition**	**Scaling**	**P1**	**P2**	**P3**
R^2	0 to 1	Score = R^2	N	N	N
Hit Rate	0 to 1	Score = Hit Rate	Y	N	Y
IC	-1 to 1	Score = (IC + 1)/2	N	N	N
UIC	-1 to 1	Score = (UIC + 1)/2	Y	N	N
AC	-1 to 1	Score = (AC + 1)/2	Y	N	N
UI	1 to 0	Score = 1 – UI	N	N	N

Table 3-a: Summary of Characteristics of Alternative Scoring Measures

At this point, Hit Rate is the most promising candidate with UIC and AC come in close as well. However, we see that there is not a single measure that is able to consider the direction as the first criteria and only uses magnitude as a tie-breaker (this latter is property P2 of our ideal scoring measure).

What we are going to propose in the next section, we termed it as Similarity Ratio. The Similarity Ratio naturally combines both the direction and magnitude of each pair of a_i and f_i by taking two orthogonal projections. We will see that the Similarity Ratio satisfies all the three properties of an ideal scoring measure.

3.3 Definition and Derivation of Similarity Ratio

3.3.1 The Worst Forecast and Maximum Inequality

We considered the "worst forecast" as the forecast that has not only predicted the direction of the relative returns wrongly, but also has a magnitude that is significantly different from the actual value observed.

To quantify the correctness of the direction is easy as one can simply compare the signs of the two values. The tougher question remains on how much we should deem the forecast is *significantly different* from the actual value. The "tolerance level" or threshold should be relative to the actual value. That is, if the actual value is large, then it is natural to have a higher tolerance to the forecasted value. On the other hand, if the value is small, the threshold should be small.

Like Theil did in *Economic Policy and Forecast*, who specifies the idea of "maximum inequality", we have limited the error to be one-time the magnitude of the actual value, but with an opposite sign. For example, if the actual value is 3%, any forecast that is less than or equal to -3% is considered as worse than the *maximum inequality* and should not be receiving any score.

3.3.2 Similarity Ratio for a Point Forecast

Suppose we have a forecast with the value of *f*, and when the actual value became available, we observed that the value is *a*. By comparing the values of *f* and *a*, it gives an indication of the quality of the forecast. Similarity Ratio for a point is a measure that compares the pair of values:

$$\text{Similarity Ratio } (a, f) = \frac{|f+a|}{|f+a|+|f-a|}$$

We also wanted to give a score of "0" to any forecast that is worse than the *Worst Forecast* or *Maximum Inequality*:

If $a > 0$ and $f < 0$, and $f \le -a$, then Similarity Ratio $(a, f) = 0$

If $a < 0$ and $f > 0$, and $f \ge -a$, then Similarity Ratio $(a, f) = 0$

The two sets of inequalities can be combined into one as follows:

If $a > 0$ and $f < 0$, and $f \le -a$, then $af \le -a^2 \Rightarrow af + a^2 \le 0$

If $a < 0$ and $f > 0$, and $f \ge -a$, then $af \le -a^2$ since $a < 0$, thus we have $af + a^2 \le 0$

Thus we can define Similarity Ratio for the point (a, f) as follows:

$$\text{Similarity Ratio } (a, f) = \begin{cases} 0 & \text{if } a^2 + af \le 0 \\ \dfrac{|f+a|}{|f+a|+|f-a|} & \text{otherwise} \end{cases}$$

Similarity Ratio(a, f) measures how "similar" is the forecast to the actual value. Naturally, a good forecast should have a higher "similarity" to the actual value observed as compared to a bad forecast. The perfect score of 1 occurred only if *a* = *f*.

3.3.3 Similarity Ratio for a Collection of Point Forecasts

We can compute the Similarity Ratio for each actual-forecast pair to obtain a score that is ranged between 0 and 1. To obtain the score for the whole sample, we average the Similarity Ratio for each sample point:

$$\text{Score} = \frac{1}{T}\sum_{i=1}^{T} SimilarityRatio(a_i, f_i) \in [0,1]$$

Where $(a_i, f_i) \in \Phi$.

3.4 Characteristics of Similarity Ratio

To illustrate the characteristics of the Similarity Ratio and show how the score varies with different quality of forecasts to an observed value, let's assumed the actual value be a positive number and plot the Similarity Ratio of a range of forecasted values. The similar plot for a negative actual value will be a mirror image along the vertical axis hence we shall not include it in our report.

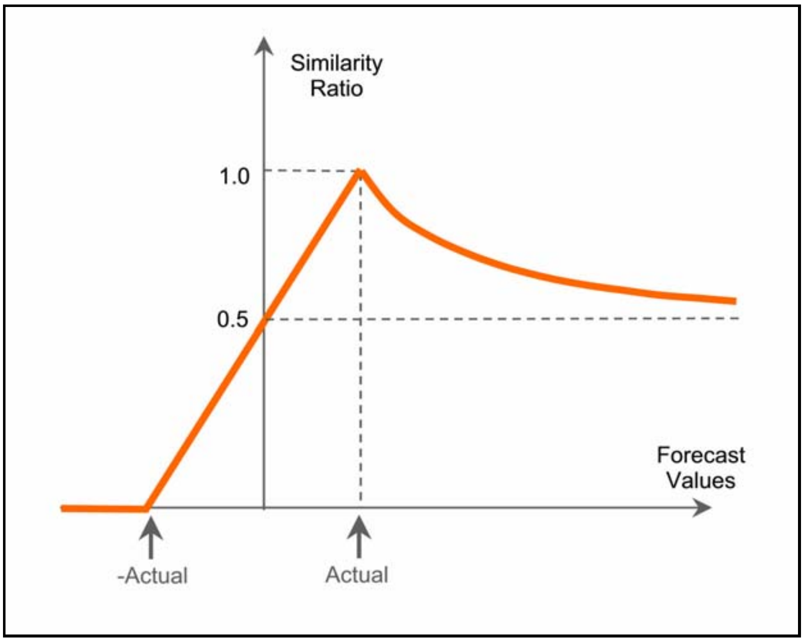

Figure 3-c: Distribution of Similarity Ratio Scores with Different Forecast Values

The graph shown in Figure 3-c displays a few important characteristics:

- If two forecasts both predicted the sign of the actual value correctly, the forecast that is closer to the actual value has a higher Similarity Ratio
- If a forecast correctly predicted the direction of the actual value, the Similarity Ratio is at least 0.5
- If a forecast correctly predicted the direction of the actual value, but over estimated the actual value, its Similarity Ratio approaches 0.5 as it deviates further
- Any forecast that is at “maximum inequality” or worse, the score is zero.

These characteristics show that Similarity Ratio satisfies all the three properties of an ideal scoring measure:

- P1. Forecast with correct direction is always higher than that of forecast that got the direction wrong.

- P2. The score converges to "1" as the forecast gets closer to the actual value means that magnitude of forecast is used in the scoring

- P3. Since we average the score of individual actual-forecast pairs to obtain the score of the sample, the aggregated score is not susceptible to outliers.

Asymmetric Nature of Similarity Ratio

The Similarity Ratio is asymmetric; this means that forecasts that are of equal deviation from the actual value will get different scores. The Similarity Ratio is biased towards deviation that is on the *right* side of the forecast. That is, if the actual is positive, the forecasts that are larger than the actual will have higher scores as compared to those forecasts that are smaller than the actual, but with equal deviation from the actual values. The opposite is true for actual values that are negative.

The intuition of this asymmetric property arises from the fact that if a forecast approaches the actual value from the *wrong* side, there is still a possibility that the model being incorrect in the directional forecast. However, if it approaches from the *right* side, it is at least sure of getting the direction correct in the forecast. As directional accuracy is the emphasis in our forecasting framework, we have designed Similarity Ratio to give more weights to directional accuracy of the forecasting models and penalized wrong directional forecast or more likely to make a wrong directional forecast.

The asymmetric property of Similarity Ratio can be proved mathematically; Supposed we have an actual value, *a*, and two forecasted values, $(a-\Delta f)$ and $(a+\Delta f)$, where Δf is some small positive deviation. If *a* is positive, we should expect Similarity Ratio for $(a, a+\Delta f)$ to be higher.

Similarity Ratio $\left(a, a+\Delta f\right) = \dfrac{|a+a+\Delta f|}{|a+a+\Delta f|+|a-a-\Delta f|} = \dfrac{|2a+\Delta f|}{|2a+\Delta f|+|\Delta f|}$

Similarity Ratio $\left(a, a-\Delta f\right) = \dfrac{|a+a-\Delta f|}{|a+a-\Delta f|+|a-a+\Delta f|} = \dfrac{|2a-\Delta f|}{|2a-\Delta f|+|\Delta f|}$

The difference between the two Similarity Ratios =

Similarity Ratio $\left(a, a+\Delta f\right)$ - Similarity Ratio $\left(a, a-\Delta f\right)$=

because $|2a+\Delta f| > |2a-\Delta f|$, we have

$$\frac{|2a+\Delta f|}{|2a+\Delta f|+|\Delta f|} - \frac{|2a-\Delta f|}{|2a-\Delta f|+|\Delta f|} > \frac{|2a+\Delta f|}{|2a-\Delta f|+|\Delta f|} - \frac{|2a-\Delta f|}{|2a-\Delta f|+|\Delta f|}$$

Since $a > 0, \Delta f > 0$, we have $|2a+\Delta f| - |2a-\Delta f| > 0$, thus

$$\frac{|2a+\Delta f| - |2a-\Delta f|}{|2a+\Delta f|+|\Delta f|} > 0$$

This follows that the difference between the two Similarity Ratios is greater than zero.

3.5 Derivation of Similarity Ratio

If we plot the forecast and actual values on a Cartesian plane, with the vertical axis representing the range for forecast and the horizontal axis representing the actual values observed, we get a point on the plane with the coefficient $\left(a, f\right)$. The Similarity Ratio is derived by comparing the distances of orthogonal projections of the point $\left(a, f\right)$ onto two lines: the *Good* and the *Bad* lines.

3.5.1 Definition of the *Good* and *Bad* Lines

Notation

We shall use the following notation:

- Let (x, y) be a point on the X-Y Cartesian plane,

  Then we let $(x, y)|_{y=f(x)}$ denotes the minimum distance between the point (x, y) and the line denoted by the equation $y = f(x)$. That is, the length of the orthogonal projection from the point to the line.

The two special cases that we considered in the derivation of Similarity Ratio are the orthogonal projections to the lines $y = x$ and $y = -x$.

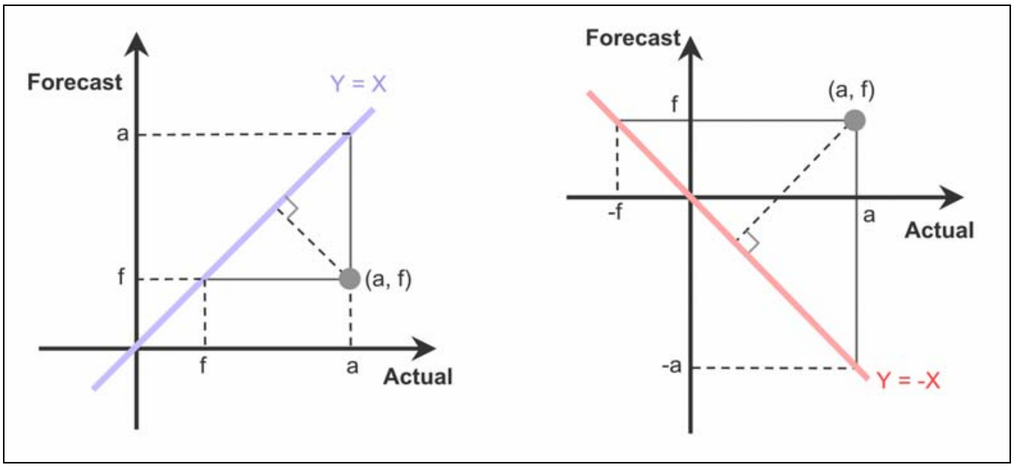

Figure 3-d: Orthogonal Projections to the *Good* and *Bad* Lines

With the application of Pythagoras Theorem, we have

$$(a, f)|_{y=x} = \frac{\sqrt{(f-a)^2 + (f-a)^2}}{2} = \frac{|(f-a)|}{\sqrt{2}}$$

and

$$(a, f)|_{y=-x} = \frac{\sqrt{(f+a)^2 + (f+a)^2}}{2} = \frac{|(f+a)|}{\sqrt{2}}$$

The fundamental idea of the Similarity Ratio is the orthogonal projections onto the *Good* and *Bad* Lines.

Points that are on the line $y = x$ would mean the forecasted values matched the observed actual values exactly. From a financial forecasting model perspective, this is the best scenario and hence the $y = x$ line is termed the *Good* line. Similarly, points that are on the $y = -x$ line would mean that the observed values are exactly opposite to the predictions, which can be disastrous if investment decisions were made based on the forecast, as such, we termed the $y = -x$ as the *Bad* line.

Using our notation, we say that the distance of the orthogonal projection from a point (*a, f*) to the *Good* line and the *Bad* line are $(a, f)|_{y=x}$ and $(a, f)|_{y=-x}$ respectively.

3.5.2 Orthogonal Projection to the Bad Line

On the *Good* line, the forecasts matched the actual values exactly thus the projections to the *Good* line measure how close the forecasts are to the actual observed values. While it is intuitive to see why we wanted to have a projection to the *Good* line, it is less clear why we need to have an orthogonal projection to the *Bad* line. However, it is this second orthogonal projection that makes Similarity Ratio different from other scoring measures. To illustrate the point on why just the projection to the *Good* line alone is not good enough, let's consider three sets of actual-forecasts which we plotted on the Cartesian plane:

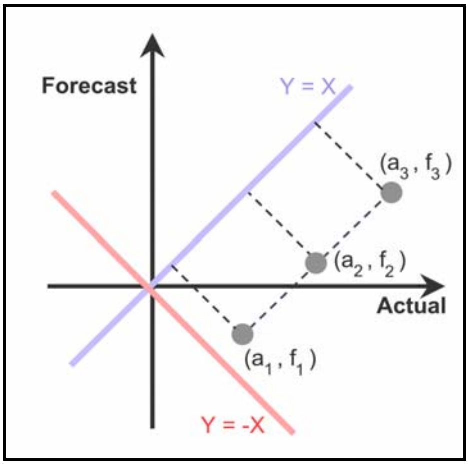

Figure 3-e: Points with Same Distance from the *Good* Line but Different Distances from the *Bad* Line

All these points have the same distance from the *Good* line but clearly we are not going to assign the same score for all the three points. The case for (a_1, f_1) is clear as the forecast and the actual value are of different signs but the argument for other two points are less obvious. With the three points having the same distance from the Good line, the magnitude of the difference between the actual values and forecasts is the same. Thus if the forecasted value is large, this difference in magnitude is smaller relative to the forecast. This represents an increasing forecast-to-observation ratio that approaches 1:

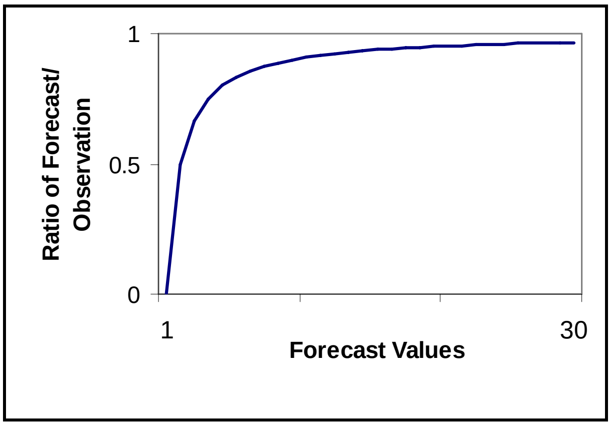

Figure 3-f: Points with Same Distance from the *Good* Line have Different Forecast-to-Observation Ratios

Forecast-to-observation ratio is a measure of "closeness" of the forecast to the observed value. Clearly we would like the forecasts that are larger to receive a higher score under this circumstance. To incorporate this in our scoring measure, we can make use of the points' distances of their orthogonal projection to the *Bad* line. The intuition is that if the points are further away from the Bad line, it is better. We will see the proof for this under proposition 3 in next section.

3.5.3 Propositions Implied in Similarity Ratio

This section reiterates the propositions of Similarity Ratio and provides the relevant proofs. The proofs will be presented assuming the case of $a > 0$. The case where $a < 0$ can be proven using similar approach, hence the proofs will not be repeated in this thesis.

Proposition 1: If a forecasts f_1 correctly predicted the direction of the actual value a, the Similarity Ratio is at least 0.5.

Similarity Ratio for (a, f_1) is $\dfrac{(a+f_1)}{(a+f_1)+(a-f_1)} = \dfrac{a+f_1}{2a} = \dfrac{1}{2} + \dfrac{f_1}{2a}$

Since $f_1 > 0$ and $a > 0$ then $\dfrac{f_1}{2a} > 0$. Thus Similarity Ratio for (a, f_1) is greater than 0.5

Proposition 2: If a forecast, f_1 correctly predicted the direction of the actual value a, the forecast f_1 over-estimated the value of a, its Similarity Ratio approaches 0.5 as it deviates further.

Because $f_1 > a$, hence $|a - f_1| = f_1 - a$, hence

Similarity Ratio for (a, f_1) is $\dfrac{(a+f_1)}{(a+f_1)+(a-f_1)} = \dfrac{a+f_1}{2f_1} = \dfrac{a}{2f_1} + \dfrac{1}{2}$

As f_1 deviates away from a, we have

$$\lim_{f_1 \to \infty}\left(\frac{a}{2f_1} + \frac{1}{2}\right) = \frac{1}{2}$$

Thus we have proved that if a forecast correctly predicted the direction but the magnitude of the forecast is wildly off, it can only get a score of 0.5.

Proposition 3: If forecasts have the same distance from the Good line, the one that is furthest from the Bad line has the higher Similarity Ratio.

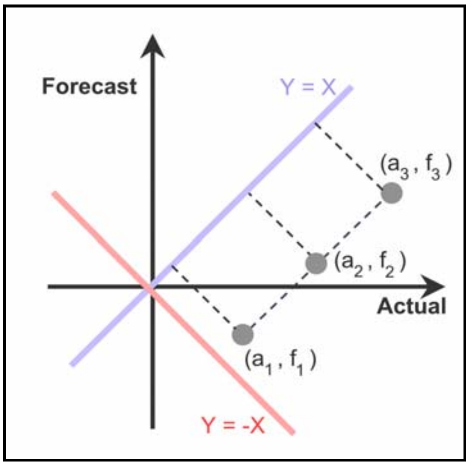

Figure 3-g: Points with Same Distance from the *Good* Line but Different Distances from the *Bad* Line

Since (a_1, f_1) has got the wrong direction in the forecast, Proposition 1 tells us that its score will be less than 0.5, which will be less than the other two sets of forecast-actual. Thus we just have to prove the case involving (a_2, f_2) and (a_3, f_3).

Since the two points are of the same distance to the *Good* line, we have the following equation:

$$(a_2, f_2)|_{y=x} = (a_3, f_3)|_{y=x}$$

$$\frac{|f_2 - a_2|}{\sqrt{2}} = \frac{|f_3 - a_3|}{\sqrt{2}} \Rightarrow |f_2 - a_2| = |f_3 - a_3|$$

Also, since (a_3, f_3) is further from the Bad line as compared to (a_2, f_2), we have:

$$(a_3, f_3)|_{y=-x} > (a_2, f_2)|_{y=-x}$$

$$\frac{|a_3 + f_3|}{\sqrt{2}} > \frac{|a_2 + f_2|}{\sqrt{2}} \Rightarrow |a_3 + f_3| > |a_2 + f_2|$$

The difference between the two Similarity Ratios,

Similarity Ratio (a_3, f_3) - Similarity Ratio (a_2, f_2)=

$$\frac{|a_3 + f_3|}{|a_3 + f_3| + |a_3 - f_3|} - \frac{|a_2 + f_2|}{|a_2 + f_2| + |a_2 - f_2|}$$

because $|a_2 + f_2| = |a_3 + f_3|$, the above can be re-written as

$$\frac{|a_3 + f_3|}{|a_3 + f_3| + |a_3 - f_3|} - \frac{|a_2 + f_2|}{|a_2 + f_2| + |a_3 - f_3|}$$

given that $|a_3 + f_3| > |a_2 + f_2|$

$$\frac{|a_3 + f_3|}{|a_3 + f_3| + |a_3 - f_3|} - \frac{|a_2 + f_2|}{|a_2 + f_2| + |a_3 - f_3|} > \frac{|a_3 + f_3|}{|a_3 + f_3| + |a_3 - f_3|} - \frac{|a_2 + f_2|}{|a_3 + f_3| + |a_3 - f_3|}$$

$$\frac{|a_3 + f_3|}{|a_3 + f_3| + |a_3 - f_3|} - \frac{|a_2 + f_2|}{|a_3 + f_3| + |a_3 - f_3|} = \frac{|a_3 + f_3| - |a_2 + f_2|}{|a_3 + f_3| + |a_3 - f_3|} > 0$$

Hence we proved that the difference between the two Similarity Ratios is always greater than zero.

The above proof shows that for points which have the same distance to the *Good* line, those points that are further away from the *Bad* line will have a higher score.

Now does this proposition make economic sense? Let's assumed the pair of assets in which relative returns that we are forecasting is asset A and B, that is, if a forecast is positive, it indicates that asset A will out-perform asset B. Now consider the simple investment decision involving the two assets. Since f_3 is larger than f_2, it implies that asset A's relative performance to asset B is higher than that implied by f_2. This should result in a bigger bet being placed on asset A in the expense of asset B because the optimization process is usually set to maximize a portfolio's risk-adjusted returns.

With the outcomes of a_3 larger than a_2, and all else being equal, the investment decision implied by forecast f_3 will result in a higher profit, thus it merits to distinguish the two forecasts and that a good scoring mechanism should recognize that forecast f_3 is more superior that forecast f_2.

3.5.4 Similarity Ratio and Canberra Metric

Even we had derived the Similarity Ratio from the first principles, we continued to research into various distance related metrics to see if there is any measure that is similar to Similarity Ratio, or have taken a similar approach. To the best of our knowledge, there is no literature found on any distance related measures being derived from the orthogonal projections, and nothing that is the same as the Similarity Ratio.

The closest we can find is a relatively unknown distance measure, called Canberra Metric or Canberra Distance. A few variations in its definition were found in literature:

Wolfram (2007), Scheidat et al (2005):

$$\text{Canberra Metric Variation 1} = \sum_{i=1}^{T} \frac{|a_i - f_i|}{|a_i| + |f_i|}$$

Johnson and Wichern (1998), Emran and Ye (2001), Rieck et al (2006):

$$\text{Canberra Metric Variation 2} = \sum_{i=1}^{T} \frac{|a_i - f_i|}{(a_i + f_i)}$$

Androutsost et al (1998):

$$\text{Canberra Metric Variation 3} = \sum_{i=1}^{T} \frac{|a_i - f_i|}{|a_i + f_i|}$$

Ignoring the summation, one would notice some similarities between Similarity Ratio and Canberra Metric. In fact, if we re-write the variation (3) of Canberra Metric using our notation, we see that it indeed includes the distances of the two orthogonal projections we have described:

$$\text{Canberra Metric Variation 3} = \sum_{i=1}^{T} \frac{(a_i, f_i)|_{Y=X}}{(a_i, f_i)|_{Y=-X}}$$

It is unclear whether were orthogonal projections considered when Canberra Metric was first derived. We found that there is relatively little literature on the subject and in general, few research on the derivation of various distance measures and which one to use when performing individualized analyses. However, considering there are three variations found on Canberra Metric, it is hard to believe that it had originated from orthogonal projections' distances because the *projection distance component* certainly did not show up in the other two variations. In fact, most who adopted Canberra Metric cautioned that it should only be applied to non-negative x_i and y_i values, for example, Johnson and Wichern (1998) and Androutsost et al (1998).

Interestingly, Androutsost et al (1998) use variation (3) of the Canberra Metric and if they do have the idea of orthogonal projections in mind, comments that the Canberra Metric is only applicable to non-negative x_i and y_i would seem unnecessary.

Thus we have every reason to believe that the idea of using two orthogonal projections to derive Similarity Ratio is innovative and intuitive in our context, so is the application of Similarity Ratio in creating a scoring measure for financial forecasting.

3.6 Similarity Ratio as the Scoring Measure for Pair-wise Framework

The choice of scoring measure is critical to the success of the pair-wise framework as it is the indication of the confidence of the forecasts for relative returns for all the possible pair-wise combinations in the investment universe. These scores will be used to sieve out the best set of pair-wise forecasts to be used in the portfolio construction stage.

In this chapter, we have surveyed six currently available scoring measures. The result of the initial qualitative assessment is that none of these measures exhibits all the characteristics of an ideal scoring measure. We went on to present Similarity Ratio and propose it to be used together with the pair-wise framework.

We see from the propositions implied in Similarity Ratio that the measure does exhibit all the important characteristics of an ideal scoring measure. While we claim this innovative scoring measure is ideal for our pair-wise framework, we will need to support with empirical evidences. The next few chapters will go into the details on how the empirical tests to be carried out and the results will be presented.

4 Testing the Framework and Similarity Ratio

We describe the pair-wise framework in the previous chapter and propose to use Similarity Ratio to be the scoring measure to select pairs. We need to provide the empirical evidences that support our proposal. As the framework aims to help construct country allocation portfolios in a systematic and objective manner, the best approach to test the framework would be to evaluate the performances of the portfolios constructed using the framework.

The output of the framework is a set of pair-wise forecasts that represents the *views* of the market. We find that Black-Litterman asset allocation model is the most appropriate and natural way for us to implement this set of views. The fact that Black-Litterman framework is able to incorporate views on asset returns or relative returns makes it a perfect choice for us to implement our pair-wise forecasts. In addition, the starting point of the Black-Litterman model is the market equilibrium returns, and this means that if the pair-wise framework does not generate any forecast that meets the minimum confidence level, we are still able to obtain an expected returns vector for use in the optimizer.

Since the objectives of our empirical studies are to test the feasibilities of the pair-wise model and Similarity Ratio, we will not elaborate the technicalities of the Black-Litterman framework in this report. Instead, we will focus on the key ideas and emphasize only those details that are required for our implementation of the Black-Litterman framework.

4.1 Black-Litterman Framework

The Black-Litterman framework was introduced in Black and Litterman (1990), and expanded in Black and Litterman (1991, 1992). It has been well received and well documented, for example, Bevan and Winkelmann (1998), He and Litterman (1999). The Black-Litterman asset allocation model is a sophisticated portfolio construction method that accounts for the uncertainty associated with the inputs and overcomes the problem of unintuitive and highly-concentrated portfolios caused by the use of optimizers.

The Black-Litterman model uses a Bayesian approach to combine the subjective views that an investor has on the expected returns of one or more assets with the market equilibrium vector of expected returns (the prior distribution). The resulting new vector of expected returns (the posterior distribution) leads to intuitive portfolios with sensible portfolios weights.

4.1.1 Market Implied Expected Returns

The use of the market implied equilibrium expected returns allows one to avoid the need to estimate the expected returns of the assets in which the investor has no view on. The Black-Litterman framework uses the set of expected returns that clears the market as a neutral starting point. This set of returns, termed Equilibrium Returns, will be augmented with the investor's subjective views to produce the expected returns vector for the optimizer. Intuitively speaking, the equilibrium returns in the Black-Litterman model are used to "center" the optimal portfolio on the market portfolio.

Reversed Optimization

The equilibrium expected returns, denoted by Π, is derived using the reversed optimization process in which the vector of excess equilibrium returns is implied by the current market weights of the assets. The formula

$$w = (\lambda\Sigma)^{-1}\mu \qquad \text{(4a)}$$

is the solution of the unconstrained quadratic maximization problem:

$$\max_{w}\left(w'\mu - \frac{\lambda w'\Sigma w}{2}\right) \qquad \text{(4b)}$$

Where

w = weight vector

Σ = covariance matrix of excess returns

λ = risk aversion coefficient

μ = vector of expected excess returns

Equation (4a) suggests that with vector of excess returns implied by the market weights (w_{mkt}) can hence be computed as follows:

$$\Pi = (\lambda\Sigma)w_{mkt} \tag{4c}$$

Where

w_{mkt} = current market weight vector

Risk Aversion Coefficient

The risk aversion coefficient (λ) characterized the expected return tradeoffs – it is the rate at which an investor will forego expected returns for each unit of risk undertaken. In the reverse optimization process described above, λ acts as a scaling factor for the reverse optimization process to estimate the excess returns. That is, the larger λ is, the higher is the resulting expected returns vector, Π .

4.1.2 Views Matrices

The Black-Litterman model allows subjective views to be expressed in either absolute returns or relative returns. That is, an investor is allowed to make a prediction on the amount of an asset returns, alternatively, he can also forecast how much one asset will out-perform another asset.

Supposed there are K views being constructed for a universe of N assets, the set of matrices used in the Black-Litterman framework to represent the views are as follows:

- $P =$ $K \times N$ matrix that identifies the assets involved in the views.

- $Q =$ $K \times 1$vector that contains the returns (absolute or relative) of each view.

- $\Omega =$ $K \times K$ diagonal matrix that represents the uncertainty of each view.

Each row in matrix P corresponds to one view. The entries in each row identify the list of assets associated with that view. Each asset in the universe corresponds to a column in matrix P so a non-zero value in column i^{th} indicates that asset *i* is represented in the view. In absolute view, the asset identified will have "1" in the corresponding column. In a relative view that involves only two assets, the asset that is over-weight will have a value of "1" while the other will have a value of "-1".

For example, supposed there are five assets (A, B, C, D and E) in the universe (*N* = 5), and we have two views (*K* = 2): Asset A out-performed asset C by 1.18%, and asset D out-performs asset A by 2.12%. Assuming column 1 is for asset A, column 2 is for asset B and so on, we can map the five assets into the five columns in matrix P, in the order of A, B, C, D and E. Matrices P and Q can be represented as follows:

$$P = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \end{bmatrix} \qquad Q = \begin{bmatrix} 1.18\% \\ -2.12\% \end{bmatrix}$$

If the relative views involved more than two assets, e.g. Asset *A* out-performs asset *B* and asset *C*, and then there are a few approaches one can choose to set up matrix P. Idzorek (2004) proposes a market capitalization weighting scheme, Satchell and Scowcroft (2000) suggest an equally weighting scheme, and Litterman (2003) prefers to assign a percentage value.

While each of these approaches handle multiple-assets differently, their representation for a relative view containing only two assets are the same. In our pair-wise model, we construct views with a pair of assets each time. This means that our relative views are always involving only two assets; this allows us to by-pass the issue of having to choose one of the three approaches to handle multiple-asset relative views.

4.1.3 Black-Litterman Formula

From the implied market equilibrium returns (Π) and the views expressed as P, Q and Ω, a new combined return vector can be constructed:

$$E[R] = \left[(\tau\Sigma)^{-1} + P'\Omega^{-1}P\right]^{-1}\left[(\tau\Sigma)^{-1}\Pi + P'\Omega^{-1}Q\right] \tag{4d}$$

Where

$E[R]$ = the new combined return vector

τ = a scalar

w = weight vector

Σ = covariance matrix of excess returns

λ = risk aversion coefficient

μ = vector of expected excess returns

P = matrix that identifies the assets involved in the views.

Q = vector that contains the returns (absolute or relative) of each view.

Ω = diagonal matrix that represents the uncertainty of each view.

Π = equilibrium expected returns

τ is a scalar that was to be used in the fine-tuning process after the new combined return vector is being computed. This parameter is ambiguous and difficult to estimate. However, with the proposed modification suggested by Idzorek (2004), which we will be implementing, the need to specify a value for τ is removed.

4.1.4 Uncertainty in Views

The diagonal elements of Ω represent the uncertainties of the views. This makes Ω the most abstract mathematical parameter of the Black-Litterman model. However, this is one common question without a "universal answer", as noted by Idzorek (2004). Herold (2003) adds that the major difficulty of the Black-Litterman model is that it forces the user to specify a probability density function for each view, which makes the model only suitable for quantitative managers.

Idzorek (2004) proposes a method to construct the error matrix using implied confidence levels in the views, which can be coupled with an intuitive user-specified confidence level (ranges from 0% to 100%). The user-specified confidence levels are used to determine the values of Ω and in the process of doing so eliminates the need of specifying the value for the scalar τ .

The details of the algorithm were presented in Idzorek (2004) hence we only provide an outline of the process. The process described below is applied on a view-by-view basis, thus the whole process has to be repeated *K* times, where *K* = the total number of views:

- Assume 100% confidence in the view
- Calculate the new combined return vector based on perfect confidence in view
- Using equation (4a), obtain the weight vector corresponds to the perfect confidence combined return vector
- Obtained the deviations of the weight vector from the market weights
- Scaled the deviations by the user-specified confident levels
- Calculate the adjusted weights by adding the scaled deviations to the market weights

- Find the diagonal value of Ω that corresponds to the view such that this value minimizes the squared difference between the adjusted weights and the weights that we obtained from the perfect confidence combined return vector.

Similarity Ratio as the Confidence Level of Views

For the above procedure to work, one would need to define a confidence level for each view. The confidence we have on each view is associated with the scoring measure we used to select the pair-wise forecasts to form the views. Thus the Similarity Ratio can be used as a proxy to the confidence we have in each forecast. It will be used to construct the error covariance matrix that is a critical input to the Black-Litterman framework.

4.2 Portfolio Construction with Black-Litterman Model

For our empirical study, we apply the Black-Litterman model with Similarity Ratios as the confidence level for all the relative returns forecasts to generate the expected returns vector. Using the return vector, we can run the mean-variance optimizer to generate the optimal country mix that meets the investment constraints we have set, e.g. holdings weights constraints.

4.2.1 Optimize to Maximize Risk Adjusted Returns

The unconstrained solution to the Black-Litterman model is stated in equation (4a). When there are constraints are to be imposed, Idzorek (2004) states that mean-variance optimization (MVO) can be used to find the optimal solution to the following objective function stated in equation (4b):

$$\max_{w}\left(w'\mu - \frac{\lambda w'\Sigma w}{2} \right)$$

We aim to find the optimal portfolio that maximizes risk-adjusted returns so we will also use the optimizer to construct our optimal portfolio that satisfies the objective function. In our implementation, we have included additional controls such as checking if the optimization process converged to a solution. If there is no convergence for any period, we will avoid taking any active position and will rebalance our test portfolio to be the market portfolio.

4.2.2 Problems with Mean-Variance Optimal Portfolios

Optimal portfolios obtained from the mean-variance optimization are often accused of being unintuitive and highly-concentrated. They are sensitive to inputs and tend to maximize the error in expected returns forecasts. For example, a large over-forecast of an asset's expected returns will result in the *optimal* portfolio to have a large over-weight of the asset, which is maximizing the error embedded in the over-forecast of the asset's expected returns. These issues are confirmed by extensive empirical studies and are well documented, for example, Idzorek (2004), Green and Hollifield (1992) and Best and Grauer (1992).

Furthermore, to use MVO one would need to provide an expected return vector, which consists of the expected returns of all the securities in the investment universe that an optimal weight will be computed. This forces users of MVO to estimate the expected returns of those securities that they did not have a view on. This simply translates to more erroneous estimates going into the portfolio construction process.

4.2.3 Dealing with the Problems of MVO

MVO portfolios were also reported to be out-performed by simple allocation strategies such as the equally-weighted portfolio (Jobsen and Korkie 1981) or the global minimum variance portfolio (Jorion 1991). Fabozzi et al (2006) noted that despite all the observed problems of MVO portfolios, these are not signs that the MVO does not work; rather it is the sensitivity of the modern portfolio theory framework to accuracy of the inputs that is causing the problems. This is the result of the optimizers' assumptions that the inputs are deterministic and are known with certainty. In reality, this is rarely the case as the inputs are statistical estimates and are uncertain.

Fabozzi et al (2006) suggested that if the estimation errors can be taken into account in applying MVO, there can be further improvements to the results of MVO portfolios. Three broad approaches were suggested to reduce the impact of estimation errors in the MVO portfolios:

1. Robust estimation framework e.g. shrinkage estimators, Bayesian estimators

2. Incorporate estimation errors directly into optimization process

3. Implement portfolio weights constraints in optimization process

The first two approaches aim to account for the estimation errors in the portfolio construction process while the third approach is an attempt to manage the risk of having over-concentrated portfolio which may bring about unnecessary *ex-post* risk.

The Black-Litterman model is an implementation that uses the first approach. In addition, we will also impose weights constraints to the optimization process as an additional means to ensure sufficient diversification in the portfolio

4.2.4 Long-only and Other Weights Constraints

Grinblatt and Moskowitz (2004) find that a large part of the gains associated with momentum-based strategies are due to the short positions in small and illiquid stocks, which may result in higher transaction costs and in some cases, impossible to implement as there may not be stocks available for shorting. To avoid the situation where the empirical results skewed by short positions, we will implement a long-only constraints in our empirical tests.

When short-selling is prohibited, we will need to add a constraint to make sure that the asset weights are always greater than or equal to zero. This is a commonly implemented constraint as many institutional active managers are not allowed to short due to restriction in clients' mandates. Griffin et al (2005) find that much returns of the momentum-based portfolios come from long-positions thus suggest such strategies will still be profitable without having to short any assets.

Upon consulting experts from the Compliance Risk area, we have set the active weights constraints of 3%, 5% and 10%, for our implementation. The allocation of limits will depend on the market capitalization of the assets:

Market Capitalization	Active Weights Constraints
Less than 10%	±3%
Between 10% and 30%	±5%
More than 30%	±10%

Table 4-a: Weights Constraints Implemented for Optimization

For example, for an asset with a market capitalization of 13% will be allowed to have an asset weight of that ranges from 8% to 18% (i.e. 13±5%). However, the long-only constraint will have the over-riding power in setting the lower bound, this means that the lower bound of active weight should be greater than or equal to zero. For example, if the asset weight is 2%, then the optimal asset weight should be from 0% to 5% instead of a range from -1% to 5%.

4.2.5 Implementation Software

The empirical tests were conducted in MATLAB[1] environment. MATLAB R2006a was used for the development with the following toolboxes: Statistics Toolbox, Neural Network Toolbox, Optimization Toolbox and Bioinformatics Toolbox

We store the data and results in a MATLAB workspace (MAT) file to allow fast retrieval and batch processing. This allows us to conduct the empirical tests in batches. Some of the MATLAB codes used can be found in the appendix.

4.3 Portfolio Implementation

The output of the optimizer is an optimal country allocation mix that will give the best risk-adjusted performance based on the pair-wise forecasts we have generated, bounded by the investment constraints we have put in. Since we only aim to see if the country mix obtained using the pair-wise framework is indeed able to generate value-added, hence we assume that the portfolio is implemented using index derivatives or ETFs without slippages. That is, index derivatives or ETFs track the performance of the underlying index perfectly. We then evaluate the performance of the country allocation portfolio based on the criteria detailed in the next chapter.

---

[1] MATLAB is a high-level language and interactive environment that is able to perform computationally intensive tasks faster than with traditional programming languages such as C, C++, and FORTRAN. For more information, see www.mathworks.com

5 Evaluation of Portfolios Performances

The primary objective of financial modeling and forecasting is financial rewards – either in earning a profit or avoiding a financial loss. Hence in evaluating the success of our pair-wise model and Similarity Ratio, it is inevitable to test the effectiveness and profitability in out-of-sample trading. However, having a profitable portfolio in out-of-sample empirical results is not the only criteria when one evaluates the feasibility of the process. There are other evaluation criteria that are worth considering:

- Does the amount of returns justify the risk taken?
- Are there excessive transactions that make the portfolio too costly to implement?
- Can the strategy withstand different market cycles and continue to perform?
- Is the profit coming from one or a few lucky trades?

These are practical considerations in evaluating the portfolios; we proposed a set of performance indicators to be used to evaluate our country allocation portfolios. Some of these indicators are commonly use in the asset management industry. The same set of indicators will also be used to compare different test portfolios. Generally, we consider a portfolio is more superior if it has:

- Higher value-added contribution
- Higher Information Ratio
- Higher percentage of the number of out-performing quarters during the test period
- Lower average portfolio turnover
- Lower correlation of its value-added to the benchmark returns

- Lower contribution of each trade to the total portfolio returns

- Higher trading edge or expected value-added

5.1 Contribution of Asset Allocation Decision to Portfolio Value-added

5.1.1 Portfolio Return and Value-added

Active managers are engaged to out-perform a specific benchmark, thus their performance is measured by how well they performance relative to the benchmark. The difference between the portfolio returns (R) and the benchmark returns ($\overline{R}$) is called *excess returns* or *value-added*:

Value-added, $VA = R - \overline{R}$

Waring and Siegel (2003) highlight the importance of evaluating portfolio performance using value-added as opposed to portfolio returns. They point out that portfolio return may be a result of the market delivering good returns and hence one may not get the value-added (pure alpha) that the active managers are paid for. By not separating the performance of benchmark return from the portfolio return will also result in incorrect apportioning of credit and blame. Portfolio manager should not take the blame for poor absolute return when the manager's asset class delivers a poor policy return. His performance should be determined by the value-added he had achieved.

5.1.2 Top-down and Bottom-up Approach to Generate Value-added

In equity investment, the investment universe can be partitioned into sub-groups, for example, a US portfolio can be partitioned into sectors[1], while a portfolio with a global mandate can be partitioned into regions or countries. These sub-groups represent groups of securities in the investment universe that are exposed to similar business or macroeconomic risks. Portfolios can be constructed by a *top-down* or *bottom-up* approach.

In a top-down approach, the investment process involves the explicit segmentation of bottom-up and top-down decisions. The investment manager first determined the allocation to the sub-groups and pick stocks from each sub-group to fulfill the target allocation mix. Since there are active bets in both asset allocation and stock selection, value-added can be generated from both sets of decision. At the same time, it also means that there is also a risk that poor stock selection may offset the value-added generated by the asset allocation resulting in an under-performance for the portfolio, or vice versa.

In a bottom-up approach, stocks are picked based on their attractiveness relative to each other. There is no explicit consideration or preference on the desired top-down allocation mix. The top-down allocation is a by-product of the stock selection. In such an approach, the value-added is generated purely from the results of the stock selection.

Top-down approach may seem to be riskier as there are additional risk factors that affect the sub-groups that the portfolio managers had to grapple with. For a given level of risk, the larger number of stock picks in a bottom-up approach facilitates diversification and improves the probability of out-performance.

[1] The commonly accepted classification is the Global Industry Classification Standard (GICS) as defined by MSCI Barra. The GICS methodology has been widely accepted as an industry analysis framework for investment research, portfolio management and asset allocation. The GICS structure consists of 10 sectors, 24 industry groups, 67 industries and 147 sub-industries.

However, many empirical studies have found that a top-down approach generates better results. For example, Fanelli and Urias (2002) conclude their study by claiming that "*resources permitting, the challenge and importance of top-down asset allocation suggests that a dedicated top-down effort is warranted*", *and that the allocation mix should not be a by-product of bottom-up stock selection.*"

5.1.3 Brinson Performance Attribution

Brinson et al (1986) propose a way to assess the skill sets of the active managers by attributing the value-added into top-down asset allocation, bottom-up stock selection and interaction effect:

Value-added = Asset Allocation effect + Stock Selection effect + Interaction effect

Where

Asset Allocation effect = $\sum_i \left(w_i - \overline{w_i}\right)\left(\overline{r_i}\right)$ (5a)

Stock Selection effect = $\sum_i \left(\overline{w_i}\right)\left(r_i - \overline{r_i}\right)$

Interaction effect = $\sum_i \left(w_i - \overline{w_i}\right)\left(r_i - \overline{r_i}\right)$

and

w_i = portfolio weight for "asset" *i*

$\overline{w_i}$ = benchmark weight for "asset" *i*

r_i = portfolio return for "asset" *i*

$\overline{r_i}$ = benchmark return for "asset" *i*

This form of attribution is often called "decision-based performance attribution" because it attributes the value-added to the decisions made by the manager. The word "asset" here means a sub-group (e.g. country) or an attribute where the stocks can be grouped together (e.g. trade currency).

Positive contribution to a portfolio's value-added by asset allocation can come from the following:

- Over-weight a sub-group and the sub-group out-performed the overall portfolio's benchmark, and
- Under-weight a sub-group and the sub-group under-performed the benchmark.

The "stock selection" effect measures the out-performance generated due selecting the *right* stocks in each country. Here, a *right* stock would be the one that out-performed the weighted average returns of all other stocks in the same country.

This leaves us with an "interaction" effect between stock selection and asset allocation. This is harder to explain than the other two effects. Briefly, the Brinson model produces a positive interaction effect when the manager over-weights countries that beat the benchmark, or under-weights countries that under-perform his benchmark. This is due to the fact that the holdings of the stocks are affected by both allocation to the country and the manager's preference for the stock that results in the stock being selected to fulfill the country allocation. Thus to understand the interaction effect, one needs to look at the results country by country. For more details, one can refer to Laker (2000).

5.1.4 Modified Brinson Performance Attribution

Combining Stock Selection Effect with Interaction Effect

Due to the un-intuitiveness of the interaction effect, and that asset allocation is usually implemented as an overlay portfolio on top of a bottom-up stock selection portfolio; it is an acceptable practice to combine interaction with the stock selection effect. The combined effect can still be broadly termed as stock selection effect:

Stock Selection effect = $\sum_i w_i \left(r_i - \overline{r_i}\right)$

It can be interpreted that if the stocks that the manager picked had out-performed those benchmark stocks with the same attribute, then these stocks' contribution to the portfolio value-added is scaled by the amount the manager owned.

Making the Asset Allocation Effect more Intuitive

To make the asset allocation effect more intuitive, we can re-write equation (5a) in the following manner:

$$\text{Asset Allocation effect} = \sum_i \left(w_i - \overline{w_i}\right)\left(\overline{r_i} - \overline{R}\right)$$

Where $\overline{R}$ = Benchmark return for the portfolio

The new form of the asset allocation effect is made up with of two terms:

- $\left(w_i - \overline{w_i}\right)$ = Portfolio weights of assets that has the attribute *i* relative to the proportion in the benchmark
- $\left(\overline{r_i} - \overline{R}\right)$ = Performance of benchmark assets with attribute *i* relative to the benchmark

Thus there is a positive contribution when the manager over-weighted the assets that out-performed the portfolio benchmark returns, or under-weighted those that under-performed.

With the two proposed modifications, the decision-based performance attribution formula can be re-written as follows:

Value-added = Asset Allocation effect + Stock Selection effect

Where

$$\text{Asset Allocation effect} = \sum_i \left(w_i - \overline{w_i}\right)\left(\overline{r_i} - \overline{R}\right)$$

$$\text{Stock Selection effect} = \sum_i w_i \left(r_i - \overline{r_i}\right)$$

5.1.5 Performance Measure to be used

To measure the performance of our country allocation portfolio, we have selected to use the Asset Allocation effect. The reason behind the selection is that there may be different ways to implement the country allocation portfolios, for example, full replication of benchmark stocks, index futures, or exchange-traded funds (ETFs). The choice of instruments and the corresponding operational efficiency will distort the portfolio performance which may skew our conclusion drawn from our empirical study. We want to see the real impact of our pair-wise model in constructing country allocation portfolios and this is best done by analyzing the value-added contributed by the asset allocation effect. For the rest of this thesis, the term *value-added* will refer to the Asset Allocation effect, unless otherwise stated.

5.2 Information Ratio – Risk-adjusted Performance

5.2.1 Tracking Error

Waring and Siegel (2003) suggest that clearly defined the value-added and active risk targets will avoid any confusion and unfairness in performance evaluation caused by mixing benchmark returns and true alphas. Since the objective of an active manager is to out-perform the benchmark returns, one good risk measure is to measure the degree of which the portfolio returns deviate from those of the benchmark. This deviation is termed *Tracking Error* or *Active Risk*. In other words, tracking error measures the consistency or volatility of excess return.

Mathematically, (ex-post) tracking error is the annualized standard deviation of the value-added. Thus a manager with a lower tracking error signifies that the portfolio “tracks” the benchmark index more closely than a portfolio manager with a higher tracking error. Tracking error computed using historical value-added series is called the *ex-post* tracking error.

5.2.2 Information Ratio

With value-added and tracking error, we can now compare the performances of two managers on common ground by looking at the returns of the portfolio in perspective with the amount of risk taken, that is, the returns per unit of risk taken. The Sharpe Ratio is commonly used to compare the absolute performances of portfolios. In the active management world, the equivalent is the Information Ratio:

Information Ratio = Annualized VA ÷ Annualized ex-post Tracking Error

As Information Ratio measures the reward for each unit of active risk taken, a portfolio with higher Information Ratio is preferred to one that has a lower Information Ratio. As Siegel (2003) correctly pointed out, active managers should seek to maximize their information ratio and that the delivery of Information Ratio is the only reason why the active managers should be paid an active fee.

The following table shows the Information Ratio of median and top quartile fund managers who were given MSCI World as their benchmark as at end 2006. As one can see, it is more difficult to achieve a high Information Ratio over a longer time period, for example, the top quartile managers only delivered an Information Ratio of 0.6 over a ten-year period while for a shorter period of three years, the Information Ratio is 1.05:

Period	**3-Year**	**5-Year**	**10-Year**
Median IR	0.36	0.47	0.32
Top Quartile IR	1.05	0.99	0.6
Number of Managers	192	155	80

Source: Mercer Global Investment Manager Database (GIMD)[1]

[1] GIMD is a global database on investment managers for institutional investors. The information in this database supports Mercer's global research on investment managers and the manager searches that Mercer performs for their clients. Mercer has over 610 full-time consultants advising over 2,700 clients with assets in excess of $3.5 trillion. For more information, see www.mercer.com

Table 5-a: Information Ratios of Median and Top Quartile Managers with Global Mandates

This statistics serves as a good comparison for us to measure our country allocation portfolio. We will compute the Information Ratio for our model portfolio to see if we are able to out-perform the top quartile managers.

5.3 Proportion of Out-performing Quarters

While it is important to look at the portfolio cumulative performance over a long time period, it is as important to have the portfolio out-performs the benchmark on a regular basis. Thus knowing the number of periods the portfolio out-performed is an important.

In our study, the portfolios are rebalanced on a quarterly basis. With the seven-year sample period, we have constructed 28 portfolios. A good measure for the success of the portfolio construction process is that out of the 28 portfolios, how many quarters did the portfolio generated positive value-added. We will express the number of quarters that out-performed as a percentage of the total number of quarters.

5.4 Turnover

As excessive transactions typically translate to high costs and hurt the profits of the investment strategies, a portfolio with lower transaction costs would be preferred to one with a higher transaction costs, assuming all other performance statistics are the same for the two portfolios. Turnover is an indication on the amount of transactions and hence the lower the turnover, the lower the expected transaction cost.

We define *turnover* as the average sum of the absolute value of the rebalancing trades across the N available assets and over the $T-M$ trading quarters:

$$\text{Turnover} = \frac{1}{T-M}\sum_{t=1}^{T-M}\sum_{j=1}^{N}\left|w_{j,t} - w_{j,t+1}\right|$$

Where $w_{j,t}$ is the portfolio weight in asset *j* chosen at time *t* before rebalancing at $t+1$, and $w_{j,t+1}$ the portfolio weight for asset *j* at time $t+1$. This definition is commonly used in similar empirical studies including DeMiguel et al (2005), Balvers and Wu (2005).

5.5 Correlation of Performance with Market's Performance

Active managers are expected to out-perform their benchmark regardless of the market conditions. This can be measured by the correlation between the portfolio performances to the market returns. Having a low correlation between the value-added and benchmark return is an indication of a good portfolio construction process. Thus when evaluating the portfolios, we will compare only the magnitude of the correlation values.

5.6 Cumulative Contribution Curve

One key challenge in portfolio performance analysis is to determine if an out-performance was a result of superior skills and process or the manager's luck. Quantification of *luck* is difficult and it is often inferred from other performance indicators. In fact, the performance indicators that we have presented so far attempt to dissect the portfolio performance to give some indication whether *luck* is the main driver.

In attempt to quantify the element of *luck*, Refenes (1995) proposes using the ratio of the profit from the largest profitable trade to the overall profit, and termed this ratio as the *Luck Coefficient* (*l*). Intuitively, a large value of *l* signals the portfolio's dependency on a single and probably non-recurring trade. Refenes

(1995) suggests that the idea can be extended to consider profit contributed by *k* largest profit-making trades to overall profits.

Brabaxon and O'Neil (2006) suggest that it is useful to have a visual examination of how the portfolio returns are being generated to give a different perspective than just inspect the risk-adjusted profits. Drawing inspiration from this comment, together with Refenes' Luck Coefficient, we designed the Cumulative Contribution Curve. This curve has a convex shape and is constructed using all the data instead of just the *k* largest points. With the curve, one can see how each trade period (e.g. quarter) contributes to the overall performance. The curve is constructed as follows:

1. Calculate percentage contribution to the portfolio value-added of each trade period
2. Sort the periods by the percentage contribution in descending order
3. Calculate the cumulative contribution for each period
4. Plot the cumulative contribution

5.6.1 Interpreting the Cumulative Contribution Curve

Cumulative Contribution Curves provide a visual comparison for portfolios out-performances. To draw a conclusion from the curves, we compare the curvatures (the flatness), the starting points and the turning points.

The Flatter, the Better

The curvature of the curve indicates the marginal contribution of each trade period. A flat curve indicates low marginal contribution of each trade period hence implies the performance is less volatile. Thus when comparing the Cumulative Contribution Curves of two portfolios, the one that is flatter is considered to be less dependent on *luck*.

The Lower the Starting Point, the Better

The starting point (the *y*-intercept point) of the Cumulative Contribution Curve is analogous to the Luck Coefficient as defined by Refenes (1995). When comparing the starting points of two Cumulative Contribution Curves, the one having a lower starting point is considered better as the profit from the best performing quarter is contributing less to the overall profit.

The Further the Turning Point, the Better

The curve is constructed by cumulating the performance one period at a time in a descending order of performance, thus the curve will continue to rise as long as the trade registers a positive contribution. So when the curve reaches the maximum point and starts to turn downwards, the losing quarters start to be included. Hence we would want the curve maximum point to curve after the half-way point, and the further the better.

Illustration on Interpreting Cumulative Contribution Curve

In the figure below, we put the Cumulative Contribution Curves of two portfolios together:

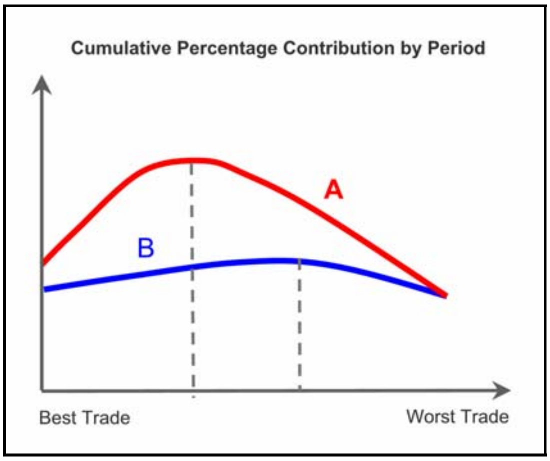

Figure 5-a: Illustration on Interpreting the Cumulative Contribution Curve

We say that portfolio B is considered to be more superior compared to portfolio A because:

- B has a flatter Cumulative Contribution Curve
- B has a lower starting point (indicated by the *y*-intercept of the curves)
- B's maximum point is further than that of A

5.7 Trading Edge or Expected Value-added

Covel (2007) points out that in trading, the hit rate or probability of positive profit counts for nothing if the magnitude of under-performance is large compared to the amount of out-performance. Covel (2007) proposes using expected performance as a measure for evaluating investment strategies. This expected performance is termed as *Trading Edge*, or simply *Edge*. Adapting it to our context, we defined *Edge* as follows:

$$\text{Trading Edge or Edge} = (P_w)(VA_+) - (1 - P_w)(VA_-)$$

where

P_w = Probability of out-performing

VA_+ = Average amount of positive value-added

VA_- = Average amount of negative value-added

We will use the percentage of out-performing quarters as a proxy for the probability of out-performing. Since the trading edge measures the expected value-added for the strategies for each quarter, a portfolio with a higher trading edge is preferred over one with a lower edge.

5.8 Summary of Portfolio Performance Evaluation

We have identified a set of comprehensive indicators that covers various dimensions for performance evaluation. The following table summarizes the performance indicators that we have described and their interpretations for a good portfolio performance.

Performance Indicator	Interpretation of Good Performance
PI1 – Asset Allocation Value-added	Higher
PI2 – Information Ratio	Higher
PI3 – Proportion of Out-performing Quarters	Higher
PI4 – Turnover	Lower
PI5 – Correlation with Market	Lower
PI6 – Cumulative Contribution Curve	Graphical Interpretation: • Flat Curve • Low Starting Point • Far Turning Point
PI7 – Trading Edge	Higher

Table 5-b: List of Performance Evaluation Criteria and Interpretation

We will use this set of performance indicators in our empirical studies to evaluate the performances of our test portfolios.

6 Empirical Results

We tested the pair-wise framework in constructing country allocation portfolios for the following benchmarks: MSCI World, MSCI Europe, MSCI Emerging Asia and MSCI Europe ex-UK. Correspond to these benchmarks, the four country-allocation model portfolios created are:

- Global (World developed equities)
- Europe (Developed Market)
- Asia (Emerging Market)
- Europe Excluding UK (Developed Market)

These portfolios were constructed using the approach detailed in chapter 4 and they are evaluated against the set of performance indicators listed in the previous chapter. This chapter presents our empirical findings.

6.1 Empirical Test Design

6.1.1 Test Objectives

The objectives of our empirical tests are to confirm the following hypotheses:

- Portfolios constructed by relative returns forecasts (pair-wise models) performed better than those based on individual asset returns forecasts (individual model)
- Portfolios constructed by using Similarity Ratio as the scoring measure in pair-selection performed better

The first objective is the test for "Choice of Relative Returns Forecast or Individual Asset Returns Forecast" while the second is a test for "Choice of Scoring Measure". To isolate the impact of choice of models and choice of scoring measure, we kept other parts of the portfolio construction process the same in each test. For example, in constructing the test portfolios using different scoring measures, the portfolio construction process is the same with only the scoring measure being different and only step 3 of the five-step process stated in **Error! Reference source not found.** was replaced.

6.1.2 Data set

Our empirical study uses 17-year (from 1990 to 2006) of monthly country index values in USD that we obtained from FactSet. We used data from 1990 to 1999 to construct the five-year indicators' returns series from 1995 to 1999 (the first data point for the 60-month returns time series starts in 1995). The performance of the constructed portfolios will be tested using data from 2000 to 2006. This gives us a five-year in-sample period and a seven-year out-of-sample period.

The constructed portfolios were held for one quarter (three months) before being rebalanced. US-dollar returns were used as we have assumed the portfolio to be currency hedge. By doing so, we have leave out the impact of currency in our analysis to isolate the impact of the pair-wise model on country allocation.

6.1.3 Empirical Results Presentation

We will present the detailed results of our model global country allocation portfolio. We will evaluate its performance based on the set of performance indicators we have identified earlier. Although model portfolios in other markets were also constructed, we will present only the detailed results for the global portfolio and show only the summary results for other model portfolios as and when necessary

Notation

To aid easy reading, when a table is used to present data for comparison, the best result will be shown in blue-colored bold font. For example, the table below that compares portfolios Port 1 and Port 2 using three performance indicators (PI1, PI2 and PI3). For performance indicator PI1, Port 1 is the better than Port 2, hence the 3.40% is highlighted in a blue-colored bold font. Similarly, the highlighted fonts in the second column suggest that Port 2 performed better than Port 1 for PI2 and PI3.

	Port 1	Port 2
PI1	**3.40%**	2.30%
PI2	1.2	**1.5**
PI3	67%	**70%**

This is to help reader quickly identify the best performers when reading a table.

6.2 Performance of Global Model Portfolio

6.2.1 Assets Universe

The 23 countries that make up the MSCI Developed World index are:

- North America – Canada, United States of America
- Europe – Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom
- Asia Pacific – Australia, Hong Kong, Japan, New Zealand, Singapore

The following table shows the each country's constituent weights in the MSCI World, as at end July 2007:

Region	Country	Weight (%)
North America	USA	47.31
	Canada	3.87
Europe	United Kingdom	11.14
	France	4.96
	Germany	4.00
	Switzerland	3.22

Region	Country	Weight (%)
	Spain	2.03
	Italy	1.85
	Netherlands	1.73
	Sweden	1.28
	Finland	0.82
	Belgium	0.59
	Norway	0.51
	Denmark	0.43
	Ireland	0.38
	Greece	0.34
	Austria	0.30
	Portugal	0.19
Asia-Pacific	Japan	10.47
	Australia	3.02
	Hong Kong	0.92
	Singapore	0.54
	New Zealand	0.08

Source: FactSet

Table 6-a: Country Constituents Weights in MSCI World (as at end July 2007)

6.2.2 Summary of Portfolio Performance

The table below summarizes the performance of the global portfolio over the seven-year period from 2000 to 2006. The numbers (except for Hit Rate) were annualized:

Performance Indicators	Value
PI1 – Asset Allocation Value-added (%)	1.61
PI2 – Information Ratio	1.15
PI3 – Proportion of Out-performing Quarters (%)	75.00
PI4 – Average Turnover (%)	35.50
PI5 – Correlation with Market	0.12
PI7 – Trading Edge (bp)	40.26

Table 6-b: Performances of Model Global Country Allocation Portfolio

Annualized Value-added (VA) can be calculated using the formula: $\left[\prod_{t=1}^{T}\left(1+VA_t\right)\right]^{12/T}-1$,

where VA_t = value-added for t^{th} month

6.2.3 PI1 – Asset Allocation Value-added

The annualized MSCI World returns in US-dollar terms from 2000 to 2006 were -0.26% per annum. In contrast, our global model portfolio averaged 1.34% annually. This represents an out-performance or value-added of 1.61% annually.

The following chart shows the cumulative value-added of the portfolio over the 28 quarters. Apart from the dip in early 2002, the performance had been fairly stable:

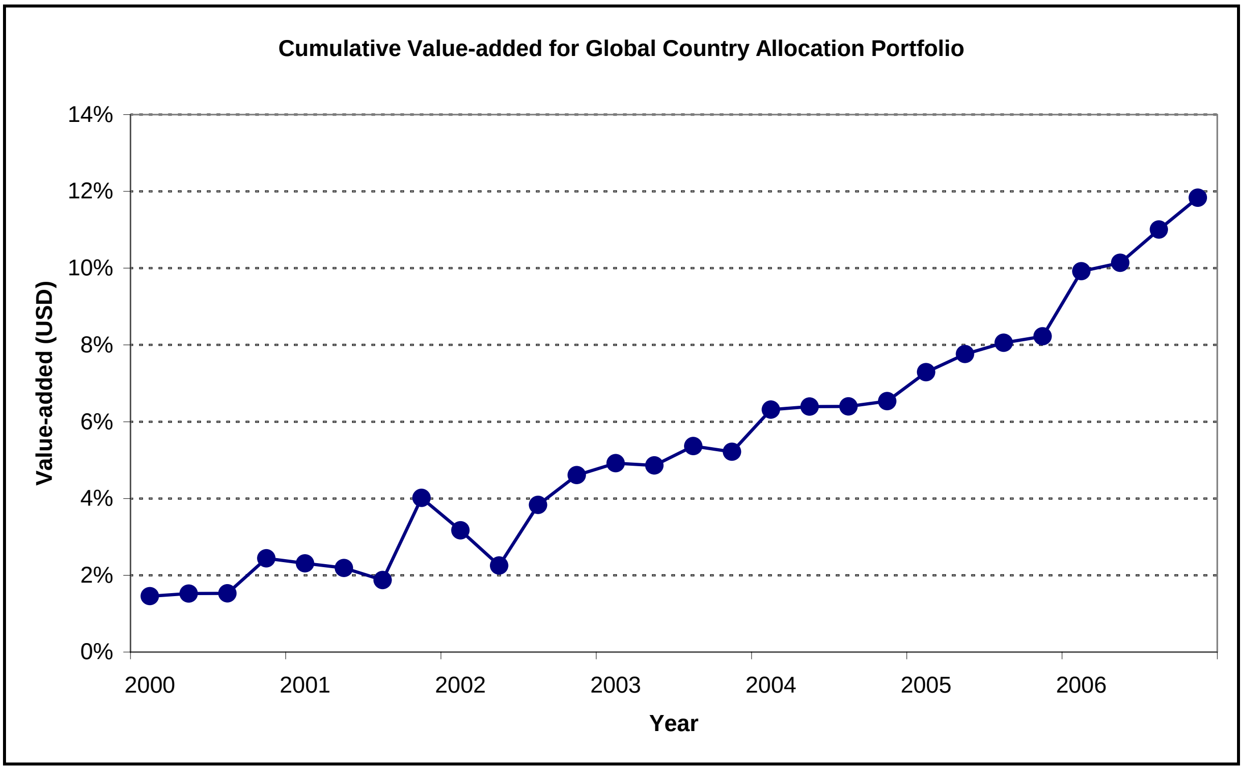

Figure 6-a: Cumulative Value-added for Model Global Portfolio

6.2.4 PI2 – Information Ratio

From the cumulative value-added chart above, we see that the value-added were accumulated steadily over the seven-year period and this is the reason why tracking error is low which resulting in a high Information

Ratio. The annual tracking error for our global asset allocation portfolio is 1.40%, with an annual value-added of 1.61%; the Information Ratio is an impressive 1.15.

According to Mercer GIMD database, we see that the top quartile managers who have a global mandate delivered an Information Ratio of 0.99 and 0.6 for the five-year and ten-year periods respectively. Comparing our global model portfolio against the managers in Mercer database, we see that its Information Ratio of 1.15 clearly ranks among the top managers.

The following table shows the median and top quartile information ratio of fund managers who were given global mandates over different periods as at end 2006:

Period	3-Year	5-Year	10-Year
Median IR	0.36	0.47	0.32
Top Quartile IR	1.05	0.99	0.6
Number of Managers	192	155	80

Source: Mercer Global Investment Manager Database

Table 6-c: Information Ratios for Median and Top Quartile Managers with Global Mandate

6.2.5 PI3 – Proportion of Out-performing Quarters

Out of the 28 quarters, our portfolio registered positive value-added in 21 quarters. This means that the proportion of out-performing quarters to the whole period is 75%. This statistics means that our portfolio is able to out-perform the benchmark regularly, theoretically, 75% of the times.

6.2.6 PI4 – Average Turnover

The portfolio also has a fairly low turnover of 35.5% compared to the turnover numbers we found in some research publications that constructed similar portfolios. In DeMiguel et al (2005), various asset allocation methodologies were used to construct a nine-country world portfolio. The quarterly turnover for the portfolios constructed using mean-variance and Bayes-Stein Shrinkage, both with constraints, are 63.6% (21.2% per month) and 53.7% (17.9% per month) respectively. Balvers and Wu (2005) constructed world portfolios using momentum and reversal strategies and noted that the average monthly turnover is 14.8%, which is about 59% per quarter.

Comparing these results with the turnover generated by our global portfolio, it showed that our portfolio's performance was generated without exceptionally high turnover.

6.2.7 PI5 – Correlation with Market

The period from 2000 to 2006 was very volatile as far as the equity market is concerned. The first three years saw MSCI world index dropped more than 50% and the next four years rebounded strongly. A sound portfolio construction process would not be affected too much from market volatility or fluctuations. To see if our portfolio performances were affected by the market ups and downs, we plotted the quarterly benchmark returns (an indication of the market's performance) against the value-added of our global portfolio. We see that the out-performance were quite evenly distributed out and looked independent of the benchmark returns. We then compute the correlation and found that it is only of 0.12.

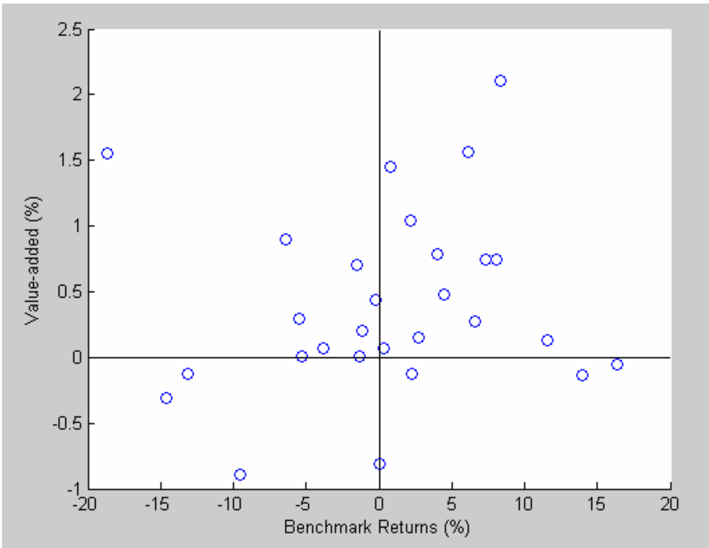

Figure 6-b: Scatter Plot of Model Global Portfolio Benchmark Returns Against Value-added

From the table below we see that for the seven years, regardless the market performance, the global portfolio is able to out-perform the market from 0.57% to 3.34%:

Year	Value-added (%)	Index Return (%)
2000	2.44	-10.77
2001	1.54	-15.25
2002	0.57	-25.20
2003	0.58	22.75
2004	1.25	9.49
2005	1.58	13.74
2006	3.34	13.52

Table 6-d: Yearly Value-added of Model Global Portfolio

6.2.8 PI7 – Trading Edge

The global country allocation portfolio has an average positive value-added of 65.37bp; this means that when the portfolio out-performed, its returns is 65.37bp above the benchmark returns. The average negative value-added is -35.06bp, which is about half of the amount of its average positive value-added.

Using the hit rate as the probability of out-performance, the trading edge of our global country allocation portfolios is 40.26bp. This means that we expect the portfolio to generate about 40bp more than the benchmark returns every quarter. This further illustrates the stability of out-performance of the pair-wise framework.

6.3 Model Portfolios in Other Markets

6.3.1 Markets Considered

We have shown empirically pair-wise portfolios constructed using Similarity Ratio as the filtering criteria works in the global country allocation portfolio. The other three model portfolios that we had constructed were:

- Developed Europe country allocation portfolio – Aim to out-perform MSCI Europe Index, which is a sub-set of the MSCI World Index with 16 countries in Europe region: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and United Kingdom.

- Emerging Asia country allocation portfolio – Aim to out-perform MSCI Emerging Asia Index. MSCI classified five countries in Asia as *developed countries*: Hong Kong, Australia, New Zealand, Singapore and Japan. The rest of the Asia is considered as *emerging* countries. Among them, ten were selected to be included in MSCI Emerging Asia Index: China, India, Indonesia, Korea, Malaysia, Pakistan, Philippines, Sri Lanka, Taiwan, and Thailand.

- Developed Europe ex-UK country allocation portfolio – Aim to out-perform the MSCI Europe ex-UK Index. United Kingdom is a large constituent country (around 30%) of MSCI Europe hence it is common that portfolios are measured against a European index that is excluding UK in order to give more beta exposure to continental Europe.

The set was chosen based on data availability. We needed country index values of at least 10-year of returns series history. We also required at least five-year of country constituents weights that made up the benchmarks.

We would like to test the pair-wise framework in other markets such emerging non-Asia (such as Latin America, Africa, Eastern Europe, etc) but there is no good source of index values and constituents' data for these countries. Using data with poor quality or insufficient history may not be meaningful and may run the risk of getting wrong conclusion.

Data Availability for Emerging Asia Country Index

As this is a relatively new index, we are only able to get the country weights data from 2002. Thus the analysis involved emerging Asia portfolios was done based on five-year data instead of the seven-year period we did for our global and European portfolios.

6.3.2 Summary of Portfolios Performance

Table 6-e summarizes the performances of the three model portfolios:

Performance Indicators	Europe	EM Asia[1]	Europe ex UK
PI1 – Value-added (%)	0.75	2.75	0.66
PI2 – Information Ratio	0.61	1.27	0.58
PI3 – Proportion of Out-performing Quarters (%)	67.90	72.73	71.43
PI4 – Average Turnover (%)	32.65	17.62	30.53
PI5 – Correlation with Market	-0.04	0.55	-0.03
PI7 – Trading Edge (bp)	18.85	68.64	16.51

Table 6-e: Performances of Model Portfolios in Europe, EM Asia and Europe ex-UK

From the table, we are convinced that the pair-wise framework also worked well in other markets. These model portfolios generated positive value-added with Information Ratio well above 0.5. The proportions of out-performing quarters are almost 70%.

The yearly value-added for the three model portfolios are plotted in the chart below. We see that the portfolios constructed using the pair-wise framework and Similarity Ratio out-performed their respective

[1] For emerging Asia, data is only available from 2002 hence only five years of portfolio performance

benchmarks for all the years except in the year 2002, our Europe ex-UK portfolio marginally under-performed the benchmark. This is a clear indication that the pair-wise framework and Similarity Ratio work in different markets and time periods.

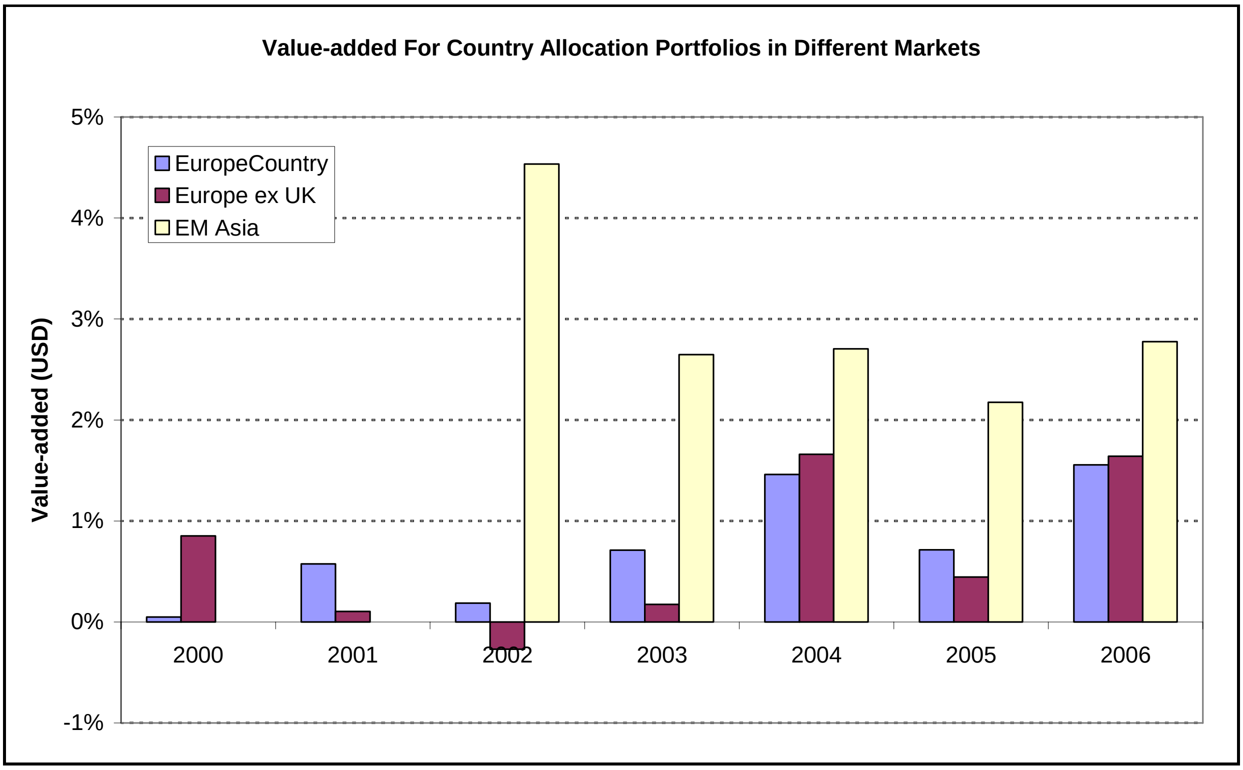

Figure 6-c: Yearly Value-added of Model Portfolios in Europe, EM Asia and Europe ex-UK

6.3.3 Comparison of Information Ratio with Other Managers

We compared the performance of the model portfolios with other managers using data obtained from Mercer GIMD. We are only able to get the performance data for European managers and global emerging markets managers hence the comparison were made using these data.

Europe

The following table shows the Information Ratios of the median and top quartile managers with the mandate of out-performing MSCI Europe:

Period	3-Year	5-Year	10-Year
Median IR	-0.01	0.16	0.17
Top Quartile IR	0.79	0.69	0.45
Number of Managers	102	83	43

Source: Mercer Global Investment Manager Database

Table 6-f: Information Ratios of Median and Top Quartile Managers with European Mandate

Our Europe and Europe ex-UK model portfolios delivered Information Ratios of 0.61 and 0.58 respectively. Although the model portfolios' Information Ratios were not as high as the top quartile managers over the five-year period, their performances were certainly ranked among the top. The model portfolios however did deliver higher Information Ratios than the top quartile managers over the ten-year period.

Emerging Asia

There is no data available for Emerging Asia thus we can only compare against the global emerging markets, which more than half of the market capitalization is from Emerging Asia:

Period	3-Year	5-Year	10-Year
Median IR	0.48	0.55	0.46
Top Quartile IR	0.95	0.95	0.6
Number of Managers	87	81	49

Source: Mercer Global Investment Manager Database

Table 6-g: Information Ratios of Median and Top Quartile Managers with Global Emerging Markets Mandate

Our model portfolio generated an information ratio of 1.27, which clearly out-performed even the top quartile managers.

6.4 Performances Comparison: Individual vs. Pair-wise Model

To test the hypothesis that modeling assets in pairs is better than treating them individually, we constructed a global country allocation portfolio using the individual model. The portfolio was also constructed using robust regression as the means to estimate the factor loadings, and used Similarity Ratio as the scoring measure. As a result, the difference in performance of this portfolio from our global model portfolio is solely due to whether the assets are being modeled individually or in pairs. This section compares the performances of this portfolio with our global model portfolio.

6.4.1 Comparison of Performance

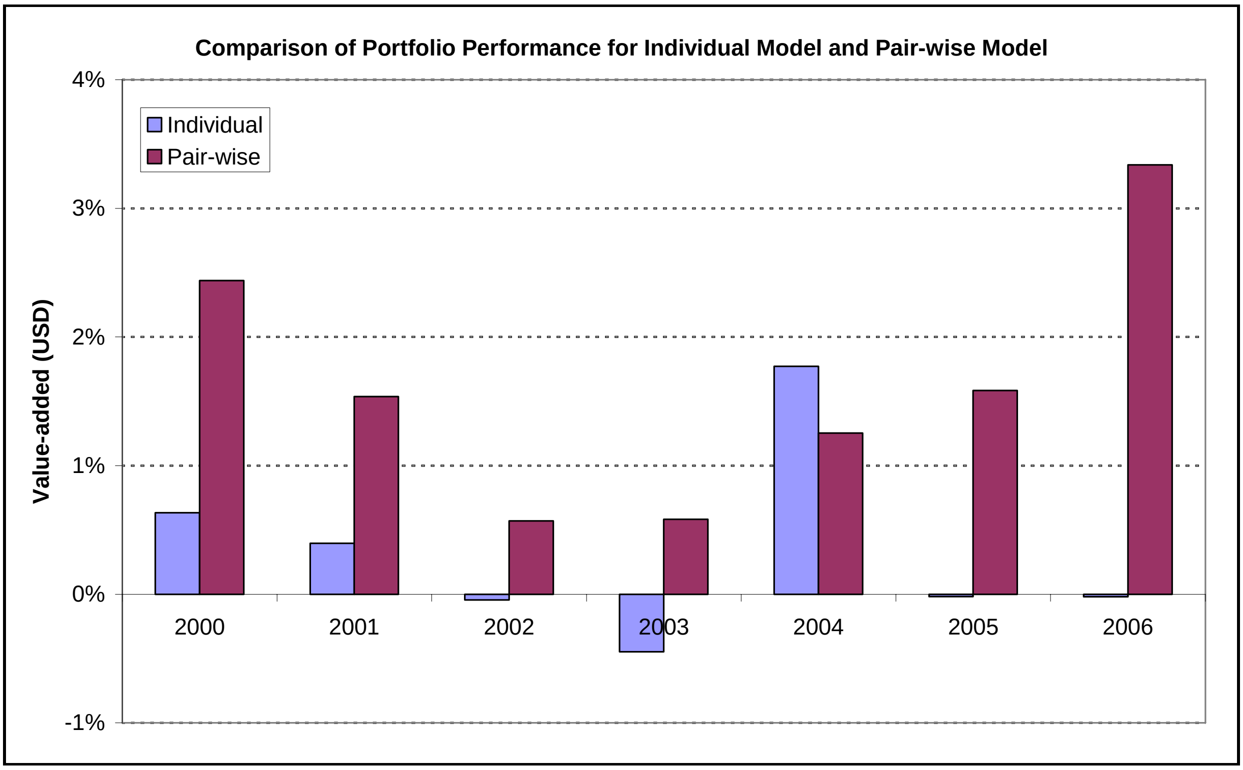

Figure 6-d: Comparison of Yearly Value-added for Individual and Pair-wise Models

Figure 6-d compares the yearly value-added of the two portfolios; the pair-wise model out-performed the individual model for six out of seven years. Even in year 2004 where the pair-wise model delivered a lower

value-added than the individual model, the pair-wise model still out-performed the benchmark return by 1.25%.

The following table compares the performance of the two global country allocation portfolios. The pair-wise model had delivered a higher value-added and Information Ratio. In addition, the proportion of out-performing quarters is also higher for the pair-wise model portfolio. All these performances were obtained with a lower turnover and less correlated with the market's performance as compared to those of individual model. It is clear that pair-wise model generates better results for all the performance indicators we used.

Performance Indicators	Individual Model	Pair-wise Model
PI1 – Asset Allocation Value-added (%)	0.33	**1.61**
PI2 – Information Ratio	0.66	**1.15**
PI3 – Proportion of Out-performing Quarters (%)	53.57	**75.00**
PI4 –Turnover (%)	13.47	**35.50**
PI5 – Correlation with Market	-0.18	**0.12**
PI7 – Trading Edge (bp)	8.45	**40.26**

Table 6-h: Comparison of Performances of Individual and Pair-wise Models

PI6 – Cumulative Contribution Curve

From the Cumulative Contribution Curves, we see that the Pair-wise model has a lower starting point; this indicates that the best trade in that portfolio is contributing less than that from the Individual model portfolio.

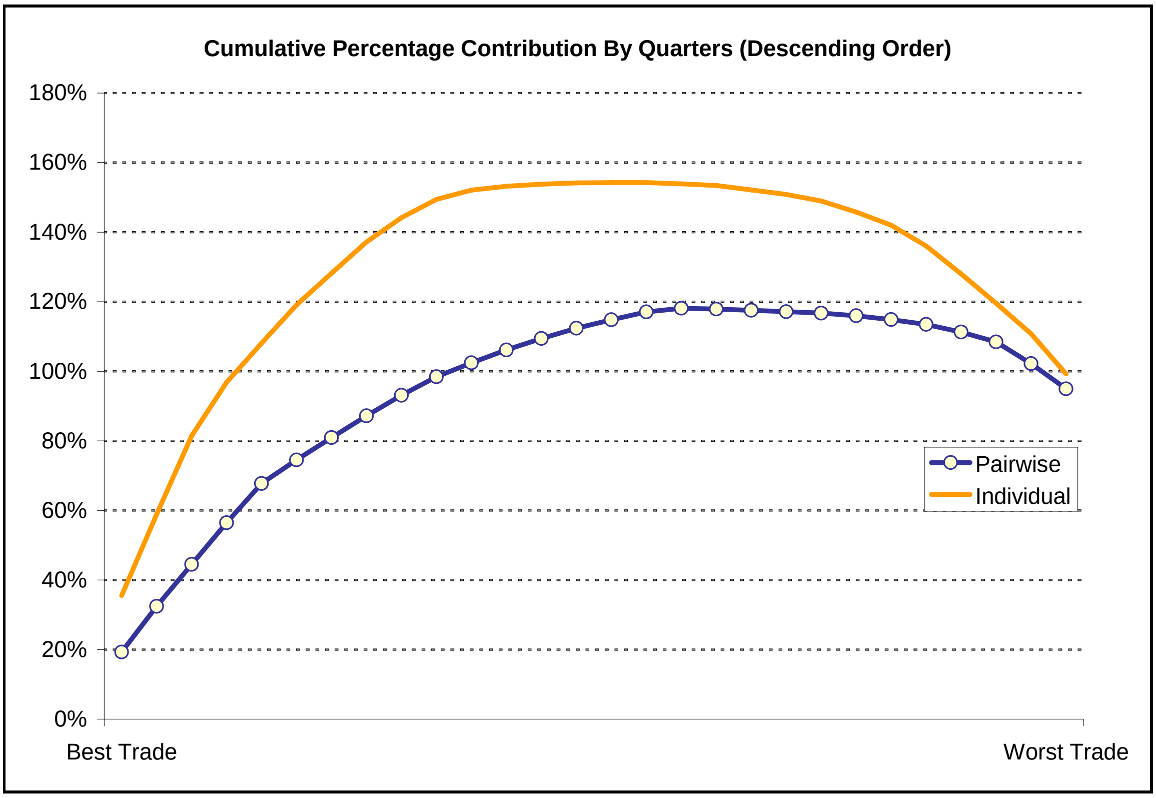

Figure 6-e: Cumulative Contribution Curves for Individual and Pair-wise Models

The curve for the Pair-wise model is flatter and is always below the curve of Individual model. This means that the profit and loss from each quarter was less volatile and had less impact on overall performance individually. This suggests that the portfolio construction process is able to generate portfolio with less fluctuating returns.

Lastly, we see that the Cumulative Contribution Curve of the individual model portfolio starts to taper off about half way point while the curve from the Pair-wise model continues to rise. This is a graphical representation of the proportion of out-performing quarters and again the Pair-wise model came up on top of the Individual model.

The analysis from the Cumulative Contribution Curves supports our claims that the using Pair-wise asset modeling in the yields better results that using the Individual asset modeling.

6.4.2 Verdict

The two global country portfolios were constructed in exactly the same manner except that in one portfolio, we model the relative returns of asset pairs (pair-wise model); and in the other, to forecast the asset return of each asset individually (individual model). Our empirical results showed pair-wise model yielded better results than individual model. With the choice of pair-wise or individual model being the only difference, it is clear the results support our claim that the pair-wise modeling works. It is clear to see that the results reveal two things:

6.5 Performances Comparison: Different Scoring Methods

To test if Similarity Ratio is indeed a better scoring measure than those available, we constructed global country allocation portfolios using different scoring measures to screen the pair-wise forecasts.

6.5.1 Scoring Method and Expected Portfolio Performances

Recalled that we have earlier suggested that we should look at three properties when assessing the scoring measures:

- P1. Measures directional accuracy of relative returns forecasts.
- P2. Uses magnitude of forecast only if directional accuracy of direction is not sufficient in distinguishing the quality of the two models
- P3. Has an aggregated score that it is not susceptible to outlier

We tabulated the candidates that we have short-listed in our study and assess their suitability based on the three properties and found that none of the six alternative scoring measures has all the three properties:

Scoring Measure	Range of Original Definition	Scaling	P1	P2	P3
R^2	0 to 1	Score = R^2	N	N	N
Hit Rate	0 to 1	Score = Hit Rate	Y	N	Y
IC	-1 to 1	Score = (IC + 1)/2	N	N	N
UIC	-1 to 1	Score = (UIC + 1)/2	Y	N	N
AC	-1 to 1	Score = (AC + 1)/2	Y	N	N
UI	1 to 0	Score = 1 – UI	N	N	N
Similarity Ratio	0 to 1	Score = Similarity Ratio	Y	Y	Y

Table 6-i: Summary of Characteristics of Alternative Scoring Measures

Based on the characteristics of the scoring measures, our initial assessment is that we expect Similarity Ratio to be the best measure since it meets all the three properties of a good scoring measure. Of the three properties, P1 (directional accuracy) is the most important, this suggests that Hit Rate, Un-centered Information Coefficient and Anomaly Correlation are expected to deliver reasonably good results. We should also expect to see portfolios constructed using R^2 and Theil's Accuracy Coefficient (UI) to perform poorly as they did not meet any of the three properties.

6.5.2 Comparison of Portfolios Performances

Figure 6-f shows the cumulative value-added of all the test portfolios constructed using different scoring measure. The portfolio that used Similarity Ratio to screen the pair-wise forecasts clearly out-performed the rest.

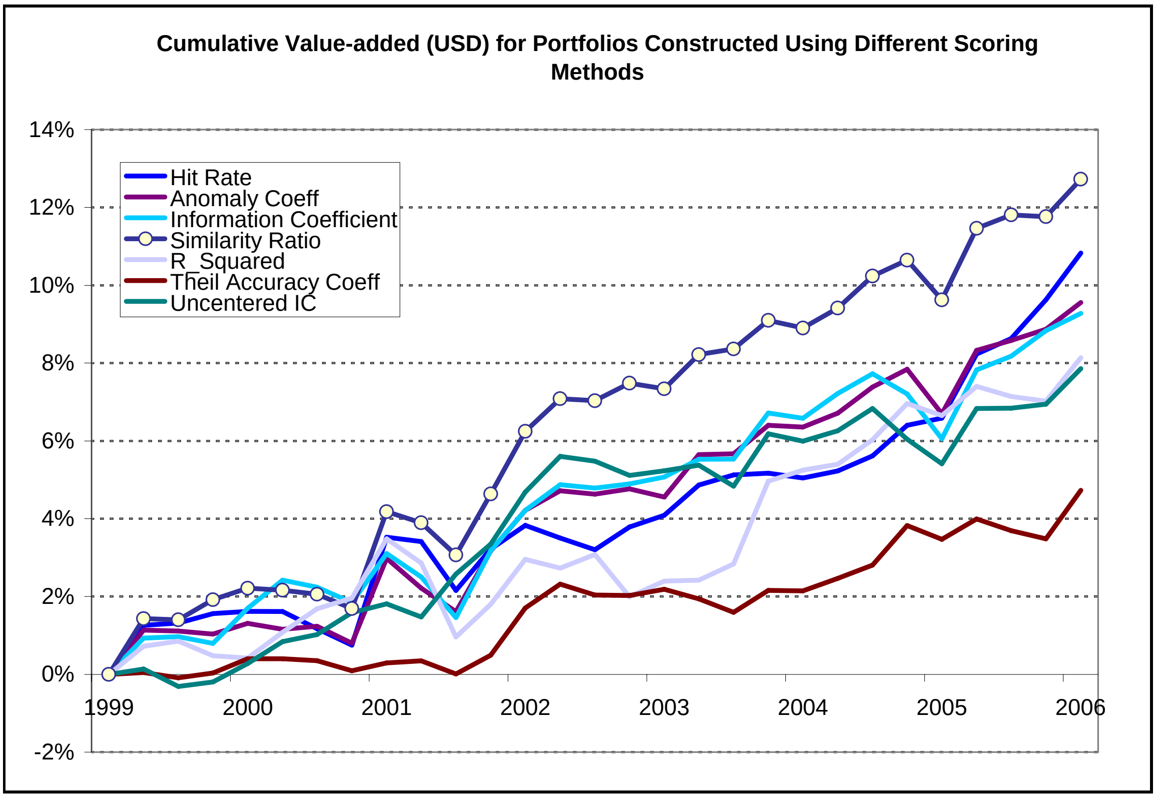

Figure 6-f: Cumulative Value-added for Global Portfolios Constructed Using Different Scoring Measures

The table below summarizes the performances of the global portfolios constructed using different scoring methods. With the winner in each category being highlighted in blue:

Performance Indicators	R^2	Hit Rate	Info Coeff (IC)	Un-centered IC	Anomaly Corr	Theil Forecast Accuracy	Similarity Ratio
PI1	1.12	1.48	1.28	1.09	1.31	0.66	**1.61**
PI2	0.73	0.98	0.90	0.93	0.91	0.74	**1.15**
PI3	67.86	71.43	67.86	71.42	64.29	57.14	**75.00**
PI4	37.45	35.46	33.81	36.41	35.59	**17.43**	35.49
PI5	0.20	0.29	**-0.08**	-0.27	0.09	0.14	0.12
PI7	28.26	37.05	31.98	32.44	32.90	16.61	**40.26**

Table 6-j: Performances of Global Portfolios Constructed Using Different Scoring Measures

From the above table, we see that the value-added delivered by the portfolios using Hit Rate and Anomaly Correlation as scoring measures were among the highest. This matches our initial assessment of the scoring measures using the three properties of ideal scoring measure.

Average Ranking

We ranked the portfolios of each scoring measure based on each performance indicator, with "1" being the best and "7" for the worst:

Performance Indicators	R^2	Hit Rate	Info Coeff (IC)	Un-centered IC	Anomaly Corr	Theil Forecast Accuracy	Similarity Ratio
PI1 (%)	5	2	4	6	3	7	**1**
PI2	7	2	5	3	4	6	**1**
PI3 (%)	4	2	4	3	6	7	**1**
PI4 (%)	7	3	2	6	5	**1**	4
PI5	5	7	**1**	6	2	4	3
PI7 (bp)	6	2	5	4	3	7	**1**

Table 6-k: Rankings of Global Portfolios Constructed Using Different Scoring Measures

With the rankings, we then calculate each portfolio's average ranking ("1" for the best and "7" for the worst) which are shown in the table below:

Scoring Method	Average Ranking
Similarity Ratio	**1.8**
Hit Rate	3.0
Information Coefficient	3.5
Anomaly Correlation	3.8
Un-centered IC	4.6
Theil Forecast Accuracy	5.3
R^2	5.7

Table 6-l: Average Ranking of Global Portfolios Constructed Using Different Scoring Measures

Again the results matched our initial assessment based on the three properties (P1 to P3) of an ideal scoring measure. We correctly predicted Similarity Ratio to be the best while R^2 and Theil's Accuracy Coefficient to be among the worst measures. This also suggests that the list of properties of an ideal scoring measure is well constructed and is a set of fair evaluation criterion to use in deciding the scoring measure.

Similarity Ratio portfolio performed the best as its overall ranking score is 1.8. The only in-sample based measure we have, R^2, had performed poorly as we had expected; its score of 5.7 is the worst among all the scoring measure. We did not have to weight the score based on the importance of different performance indicators because this will only skew the score towards Similarity Ratio since it ranked first in the most important indicators: value-added and information ratio. Using a weighted average ranking will not change the fact that Similarity Ratio ranked first, in fact, it would only increase its winning margin.

PI6 – Cumulative Contribution Curves

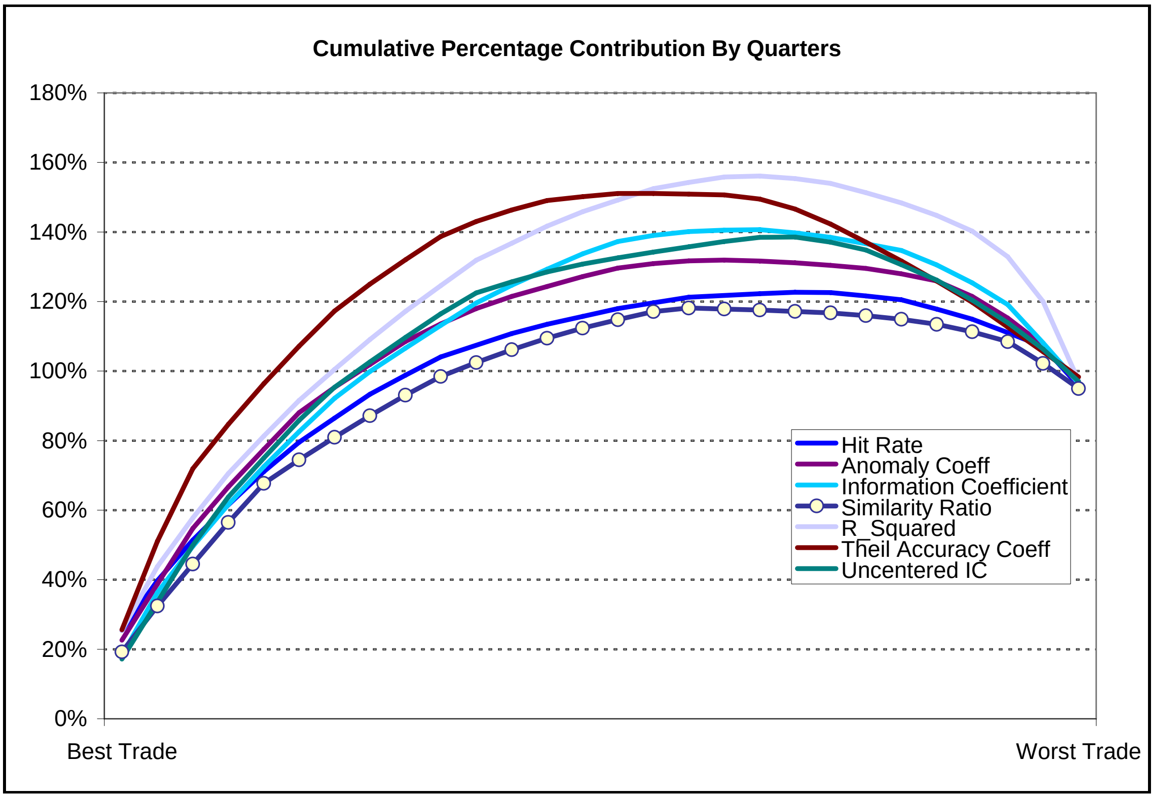

Figure 6-g: Cumulative Contribution Curves of Global Portfolios Constructed Using Different Scoring Measures

The chart plots the Cumulative Contribution Curves for the portfolios constructed using different scoring methods. We see that the Similarity Ratio's curve has the lowest starting point, and that it is the lowest (i.e. most flat). Thus from the Cumulative Contribution Curves, we see that the portfolio constructed using the

Similarity Ratio as the filtering criteria generated better performance than those portfolios using other scoring methods.

6.5.3 Verdict

Table 6-l shows the average ranking of each global portfolio constructed using different scoring measures. We see that the portfolio that used Similarity Ratio as the means to screen the pair-wise forecasts had the best ranking of 1.8. The next best scoring measure had a score of 3.0, which is relatively poor as compared to the ranking obtained by our model portfolio. From the Cumulative Contribution Curves, we also concluded that portfolio constructed using Similarity Ratio is more superior than those portfolios constructed using other scoring methods.

6.5.4 Pair-wise Model Out-performed Individual Model – even without Similarity Ratio

The results also added more evidences that the pair-wise model works better than modeling each asset individually. Almost all portfolios constructed using the pair-wise framework generated a value-added of more than 1%. Even the worst portfolio, the one that used Theil Forecast Accuracy as the scoring measure, delivered a value-added much higher than the 0.33% generated by the individual model. In fact, the individual model portfolio also has the lowest Information Ratio, proportion of out-performing quarters, and trading edge. Thus these results provide another indication that using pair-wise model generates better results, even if the scoring measure used was not Similarity Ratio.

6.6 Performances Comparison: Hit Rate vs. Similarity Ratio

Among the alternative scoring measures, Hit Rate is the closest competitor to the Similarity Ratio; Hit Rate has the next highest average ranking and its Cumulative Contribution Curve is closest to that of the

Similarity Ratio. To find out if Similarity Ratio is indeed a better scoring measure than Hit Rate, we extended the empirical tests for these two measures. We constructed test portfolios with other benchmarks: MSCI Europe, MSCI Emerging Asia, and MSCI Europe ex-UK.

6.6.1 Comparison of Performances

PI1 – Asset-allocation Value-added

Portfolios constructed using Similarity Ratio out-performed those constructed using Hit Rate for all the three other markets. As we can see from the chart below, the Similarity Ratio portfolios registered positive value-added in all three markets whereas while the Hit Rate portfolio even under-performed in the continental Europe portfolio:

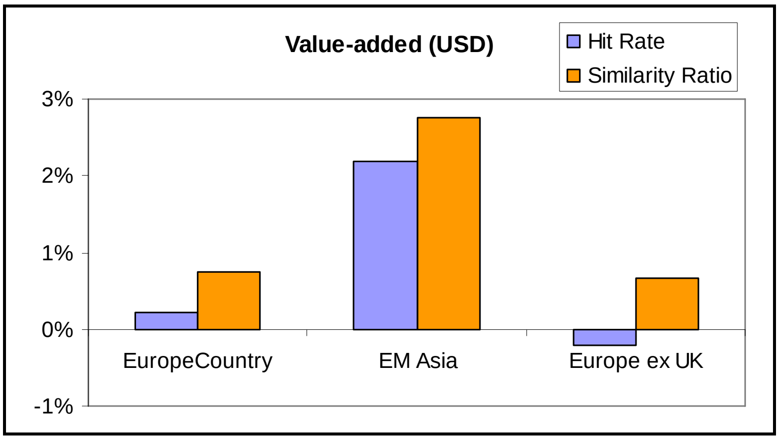

Figure 6-h: Value-added of Portfolios Constructed Using Hit Rate and Similarity Ratio

PI2 – Information Ratio

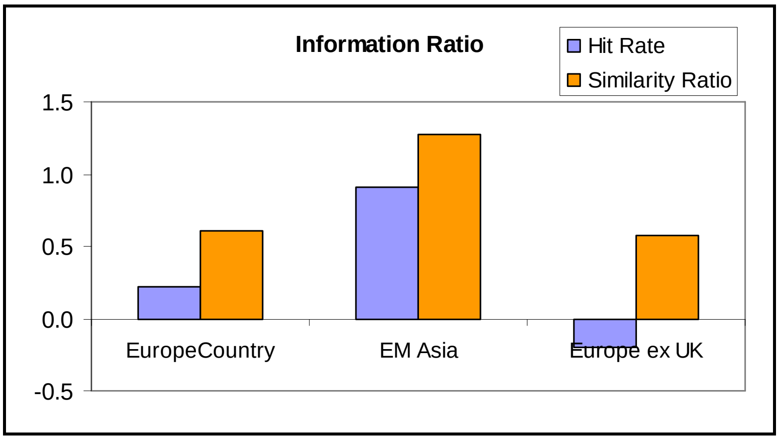

Figure 6-i: Information Ratios of Portfolios Constructed Using Hit Rate and Similarity Ratio

Not only did the Similarity Ratio portfolios generated higher value-added, they did it without taking excessive risk. This is evident from the higher Information Ratios that our Similarity Ratio portfolios generated. The Information Ratios of the portfolios constructed using Similarity Ratio were higher than that of portfolios constructed using Hit Rate.

PI3 – Proportion of Out-performing Quarters

From the Similarity Ratio portfolios' higher proportion of out-performing quarters than that of the Hit Rate portfolios, we see that the higher value-added of the Similarity Ratio portfolios were generated over more quarters than that of the Hit Rate portfolios.

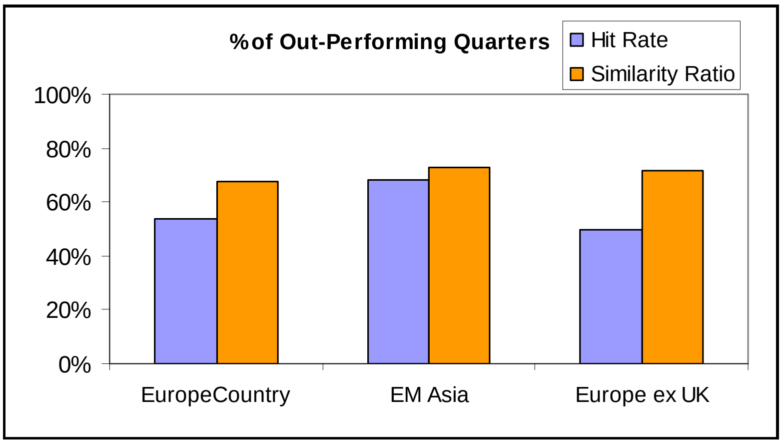

Figure 6-j: Percentage of Out-performing Quarters of Portfolios Constructed Using Hit Rate and Similarity Ratio

PI4 – Average Turnover

The average turnover for the European portfolios were very close for two sets of portfolios hence there is no clear indication that which set is considered better. However, for the Emerging Asia market, the Similarity Ratio portfolio's average turnover is clearly lower than that of the Hit Rate portfolio:

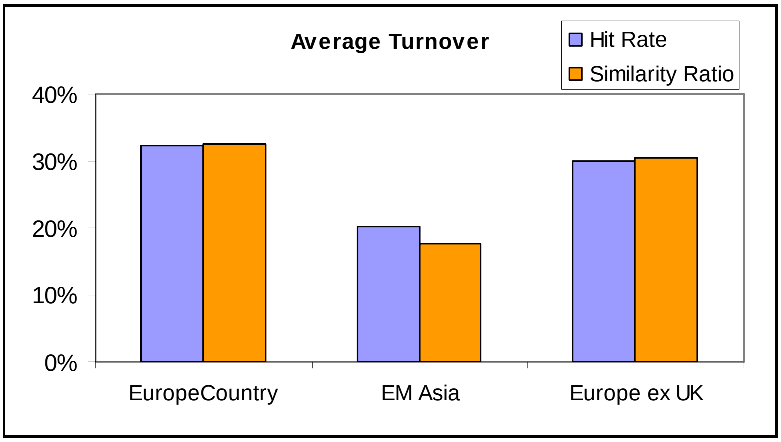

Figure 6-k: Average Turnover of Portfolios Constructed Using Hit Rate and Similarity Ratio

PI5 – Correlation with Market Performance

Both Hit Rate and Similarity Ratio portfolios have very high correlation with the market returns in the Emerging Asia market. This may be true due to the positive returns of the Asia markets in the past few years. However, over a longer period of seven years, we see that the Similarity Ratio portfolios have relatively lower correlation with the European market returns, as compared to the Hit Rate portfolios:

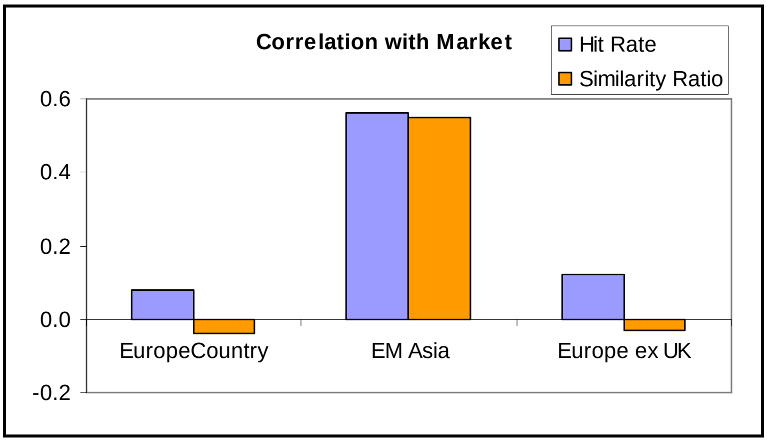

Figure 6-l: Correlation of Value-added with Market Returns of Portfolios Constructed Using Hit Rate and Similarity Ratio

PI7 – Trading Edge

The expected performances for the European portfolios constructed using Similarity Ratio were higher than those that used Hit Rate as the scoring measure. However, the hit rate based Emerging Asia portfolio has a marginally higher expected value-added that the Similarity Ratio based portfolio. However, with the trading edge for both at a high of almost 70bp, it should not matter much. On the whole, comparing the two sets of portfolios using trading edge only marginally favors those portfolios which were constructed using Similarity Ratio as the scoring measure.

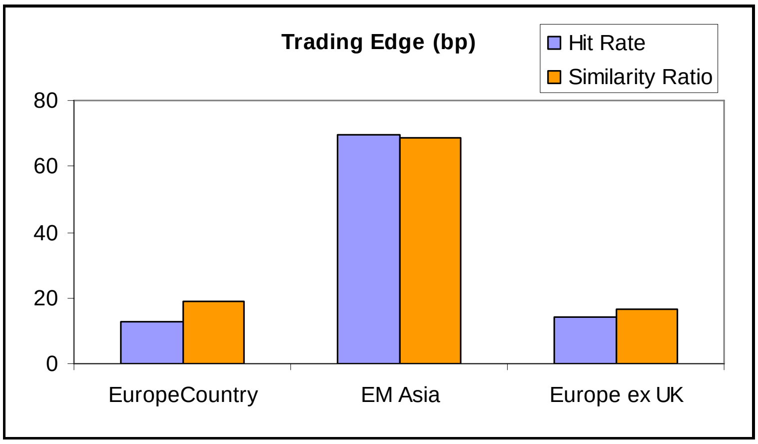

Figure 6-m: Trading Edge of Portfolios Constructed Using Hit Rate and Similarity Ratio

6.6.2 Verdict

Summarizing the findings in the following table, we see that the Similarity Ratio portfolios out-performed the Hit Rate portfolios in almost all the performance indicators except for the average turnover where both sets of portfolios are very close.

Performance Indicators	Europe Country		Emerging Asia Country		Europe Country ex-UK	
	Hit Rate	**Similarity Ratio**	**Hit Rate**	**Similarity Ratio**	**Hit Rate**	**Similarity Ratio**
PI1 (%)	0.22	**0.75**	2.18	**2.75**	-0.22	**0.66**
PI2	0.22	**0.61**	0.91	**1.27**	-0.20	**0.58**
PI3 (%)	53.57	**67.86**	68.18	**72.73**	50.00	**71.43**
PI4 (%)	**32.24**	32.65	20.33	**17.62**	**29.90**	30.53
PI5	0.08	**-0.04**	0.56	**0.55**	0.12	**-0.03**
PI7 (bp)	12.59	**18.85**	**69.54**	68.64	14.40	**16.51**

Table 6-m: Summary of Performances of Portfolios Constructed Using Hit Rate and Similarity Ratio

Thus we conclude that portfolios constructed using Similarity Ratio clearly out-performs those that were constructed using Hit Rate.

6.7 *Conclusion from Empirical Results*

Our empirical tests were designed to test the following hypotheses:

- Portfolios constructed using relative returns forecasts (pair-wise models) performed better than those based on individual asset returns forecasts (individual model)
- Portfolios constructed by using Similarity Ratio as the scoring measure in pair-selection performed better than those based on other scoring measures.

Comparing the results of the individual model portfolio with the pair-wise portfolios constructed using different scoring measures, we see from the portfolios performances that regardless of the scoring measure used in pair-selection, portfolios constructed using relative returns forecasts performed better than the portfolio that was constructed using individual asset returns forecasts. For example, the pair-wise portfolio with the worst Information Ratio (IR) is the one that used R^2 as the scoring measure. However, its IR of 0.73 is still higher than individual model's 0.66. The pair-wise portfolio with the worst value-added is the one that used Theil Forecast Accuracy as the scoring measure. Again, its value-added of 0.66% is double of the individual model value-added of 0.33%.

Among the pair-wise portfolios, the portfolio that used Similarity Ratio as the scoring measure performed that best. The portfolio's IR of 1.15 and value-added of 1.61% out-performed all other pair-wise portfolios. In addition, 75% of the times the portfolio out-performed the benchmark and this is the highest among the seven portfolios.

The results have provided strong empirical evidences to support our proposal of pair-wise framework and the use of Similarity Ratio as the scoring measure for pair-selection.

7 Generating Views with Classification Models

In our empirical studies, we had used momentum robust regression to estimate the factor loadings of the linear model for each asset pair. Regression is a common approach in building linear factor model and the empirical studies had showed positive results using robust regression in a pair-wise framework.

We extended our research to use classification techniques in generating sign-forecasts for the pair-wise relative returns so that we can test if pair-wise modeling works with different forecasting models. In addition, the results will also provide invaluable empirical evidences to show that forecasting models constructed with robust regression out-performed those constructed using classification techniques.

We review various representative classification techniques and how can they be applied in our pair-wise framework. The second part of the chapter shows the implementation details and presents the empirical results.

7.1 Classification Models

7.1.1 The Classification Problem in Returns Forecasting

We can formulate three possible out-comes for the relative performance of an asset pair:

1. First asset out-performed the second asset,
2. both assets performed the same thus no one asset performed better than the other, and
3. first asset under-performed the second asset.

With this formulation, we are able to introduce classification techniques into our pair-wise framework. Instead of using robust regression to estimate the factor loadings of the momentum indicators in each pair-wise model, we provide sample data to *train* the classifiers to explore the relationship between the indicators' values and the direction of relative returns. At decision time, the trained classifiers will then be used to forecast the direction of the relative returns with the latest set of indicator values.

7.1.2 Classification Techniques

The construction of a classification procedure from a set of data for which the true classes are sometimes known as *pattern recognition*, *discrimination*, or *supervised learning* (in order to distinguish it from unsupervised learning or clustering in which the classes are inferred from the data).

We consider here methods for supervised classification, meaning that the construction of a classification procedure based on a set of sample data with known classes. A classification method uses a set of features or parameters (predictors) to characterize each object. The supervised classification programs will map the training input data to the training output data to find the relationships that best classify the output data into different classes.

There are two phases to constructing a classifier. In the training phase, the training set is used to decide how the parameters ought to be weighted and combined in order to separate the various classes of objects. In the application phase, the weights determined in the training set are applied to a set of objects that do not have known classes in order to determine what their classes are likely to be.

Michie et al (1994) group the classification techniques into three categories:

- Statistical
- Machine Learning

- Neural Networks

Statistical Approaches

Two main periods of work on classification can be identified within the statistical community. The first, "classical" period, concentrates on derivatives of Fisher's early work on linear discrimination, e.g. quadratic discrimination and logistic discrimination. The second, "modern" period, exploits more flexible classes of models, e.g. *k*-nearest Neighborhood.

Machine Learning

Machine Learning attempts to mimic human reasoning behavior sufficiently to provide insight into the decision-making process. It is generally taken to encompass automatic computing procedures based on logical or binary operations that learn a task from a series of examples (supervised learning). The key technique in this category has been Decision Tree, in which the classification process is a result from a sequence of logical steps. Other techniques under this category are Genetic Algorithms and Inductive Logic Procedures (ILP). GA is a mathematical optimization algorithm with global search potential. The methodology is inspired by a biological metaphor and applies a pseudo-Darwinian process to evolve the solutions iteratively.

Neural Networks

Neural networks combine the complexity of some of the statistical techniques with the machine learning objective of imitating human intelligence. A broad class of techniques can come under the umbrella of neural networks. Generally, neural networks consist of layers of interconnected nodes, each node producing a non-linear function of its input. The inputs to a node may come from other nodes or directly from the input data. More complex networks may have output nodes connected with earlier nodes to provide feedback into the network. These recurrent networks can be useful when modeling time-series data, Elman back-propagation network is one such network configuration.

7.1.3 Research on Application and Comparison of Classification Techniques

Neural Network and Support Vector Machine are the Most Popular Candidates

There have been plenty of research works being done on the application of classification techniques in the area of financial forecasting. Refenes (1993) and Steurer (1993) found that neural network forecasting models have outperformed both statistical and other machine learning models of financial time series, achieving forecast accuracy of up to 58%. A survey on the recent works revealed that back-propagation neural network is a popular choice of network architecture. Just to name a few, we have Roman and Jameel (1996), Yu (1999), Chen et al (2003) and Kohara (2003).

The popularity of neural networks is also the reason why it is commonly treated as a *competitor* when other classification approaches are being proposed. For example, Tay and Cao (2001), Kim (2003), and Huang et al (2004) all recommended Support Vector Machine and all of them compared their results against the neural network models.

Support Vector Machine (SVM), first proposed by Cortes and Vapnik (1995), is another popular classification technique. Support Vector Machine is favored because researchers found that it is resistant to the over-fitting problem, and hence likely to lead to a high generalization performance. Another important property of the Support Vector Machine is that training the classifier is equivalent to solving a linearly constrained quadratic programming problem; this means that the solution is always unique and globally optimal. In contrast, neural networks' training process requires nonlinear optimization and often being accused of getting stuck at local minima. Some other works that recommended Support Vector Machine in financial modeling are Yang et al (2002), Yu et al (2005) and Chen et al (2006).

Statistical and Regression Techniques are Less Popular

We found that the less popular choices in the published research on the application of classification techniques in financial modeling are what Michie et al (1994) call *classical* statistical techniques. Regression is one such technique that is not commonly seen in research works. Perhaps these techniques have existed for a long time and are well established. Thus their roles in the research studies are often been that of alternative classifiers. They are used to compare the results when a more advanced technique, such as Support Vector Machine and Neural Network, is being proposed. For example, Huang et al (2004) compare Support Vector Machine against Discriminant Analysis (Linear and Quadratic). Poddig, T. (1995) compares the performance of Discriminant Analysis against Neural Network (Multi-layer perceptron and learning vector quantiser) in the area of bankruptcy prediction. Kim (2003) includes *k*-nearest neighbor algorithm as an alternative when he proposes applying Support Vector Machine in forecasting daily Korea Composite stock price index.

Baestaens and Bergh (1995) compare regression against neural networks in tracking the Amsterdam Stock Index while Refenes et al (1994) examine the use of neural networks as an alternative to regression. Zapranis et al (1996) also compare regression to neural network in the construction of a cash-equities portfolio. As many of the research works were to introduce more advanced classification techniques, it is not surprising that most empirical evidences supported the claims that regression technique is inferior when compared to classifier based on Neural Networks or Support Vector Machine.

7.1.4 Observations of Empirical Tests Setup of Research Publications

We reviewed a large collection of research articles on classification techniques; some with applications in financial modeling and many in other areas. We studied those works that applied classification techniques to stock indices forecasting and made some observations:

- **Short out-of-sample period –** Out-of-sample size can be as short as 20-day (Yang et al 2002) to three-year (Zapranis et al 1996). However, most conclusions were made based on empirical results using one-year out-of-sample data. For example, Huang et al (2004) use 36-week out-of-sample data for their empirical tests.

- **High frequency (daily) data** – This may limit the robustness and stability of the models over a longer history, say three years and above, which is common in active asset management. For example, Tay and Cao (2001), Kim (2003), Roman and Jameel (1996) conduct their tests using high frequency data.

- **Over-emphasis on forecasting accuracy** – Many of the works drew their conclusion based on forecasting accuracy, for example, Yang et al (2002), Tay and Cao (2001). By doing so, one has assumed that forecasting accuracy implies profitable portfolio, which is not necessary the case as the implementation of trades may result in high tracking error and transaction costs. There is no empirical result shown on how can one use classification techniques in actual portfolio construction.

- **Focus on absolute returns** – None of the publication work we found compares the performance of the classification technique with the benchmark returns. The high absolute returns generated by the trades may be due to the market component (beta) and there may be little alpha being added. There is no finding if an index funds would result in a higher profit.

- **Assumed the same set of factors** – There is no mention of a framework that explains the selection of best predictors in each decision time. A static set of predictors is used for the entire empirical study. Thus questions on how these predictors are to be selected at decision-time remain un-answered.

- **Modeling of single-asset with the aim of trading that asset** – Most works attempt to model the returns of an asset with the objective of making short-term trading gains. This does not suggest that the results are also applicable when the objective is to construct a portfolio that aims to out-perform a benchmark. Even when Roman and Jameel (1996) consider a basket of countries in their modeling, their aim is just to pick the best country to invest rather than constructing an optimal country allocation.

- **Empirical tests are mainly restricted to a single market or a small set of markets** – Most empirical results are confined to a single market, for example, Japan (Yu 1999, Kohara 2003), Taiwan (Chen et al 2003), Korea (Kim 2003) and Holland (Baestaens and Bergh 1995). Whether what they have proposed work for different markets remained unclear. Though Tay and Cao (2001) apply their model to different asset classes, such as US government bond, Germany government bond, US and France stock markets; they found that only the US stock market performance had a hit rate of more than 50%.

Comparing what we have observed so far, we see there are many differences in the way we had carried out the empirical studies of the pair-wise framework:

- We have a long out-of-sample period of seven years
- We use monthly data to forecast construct quarterly-rebalanced portfolio which means more stable and robust results
- We evaluate the portfolios using a full spectrum of performance indicators – from profitability to risk to correlation to market
- We attribute the performance appropriately to avoid drawing conclusion skewed by market returns.
- We adopt a contextual process that is repeatable as not test data specific
- We construct a basket of assets for long term performance instead of gaining from short-term trading
- We construct global and regional portfolios and not confined to a single market

We believe that our empirical framework is more comprehensive and extensive. Using this framework we will be able to give the various classification techniques a fair comparison with robust regression.

7.1.5 Implementation of Classification Techniques

Selecting the Classification Techniques

We selected a representative set of classification techniques to compare against robust regression. These techniques were selected either they had been tested in related research works (e.g. Support Vector Machine) or had exhibit characteristics that may potentially suit our needs (e.g. Decision Tree, Elman Network).

There are different variations of neural networks and it is not possible to test all networks configurations. We have chosen Probabilistic Neural Network (PNN) and Elman Network (ELM) because we find that they are the most suitable in our context. These network architectures work well in the area outside financial forecasting, for example, Comes (2003) uses PNN to classify micro-array data, Goh et al (2000) applies ELM to construct an intelligent system for time series prediction in product development, Tan and Silva (2003) make use of ELM in the study of human activity recognition by head movement. PNN is efficient to train and is less susceptible to outliers; it has been working well in classification problem. Elman Network takes into the consideration the sequences of events by having feedbacks to the contextual units; this suggests that it is suitable for time series modeling.

The final list of classification techniques selected for our empirical tests are:

- Linear Discriminant Analysis
- Quadratic Discriminant Analysis

- Logistic Regression
- *k*-nearest neighbors
- Decision Tree
- Support Vector Machine
- Probabilistic Neural Network (PNN)
- Elman Network (ELM)

Choosing the Model Variation and Setting the Parameters' Values

Each classification technique has a different set of parameters and each requires specific settings for the technique to be effective, for example, the choice of *k* in *k*-nearest neighbor and the number of neurons in Elman Network. In order to maintain objectivity of our empirical tests and to prevent data snooping, we first surveyed past research to see if there is any recommended parameter values or specific variation of the classification model that we should use. For example, in our implementation of SVM, we adopted the Gaussian RDF as the kernel function with bandwidth set to 10 as these were used in many research publications.

In situations that we do not find any recommended settings, or that the settings are too different from each other, we will apply cross validation with some qualitative assessments. For example, the number of neurons used in Elman Network can range from a handful to hundreds. This leads us to use cross validation and make a decision based on some qualitative analysis of the results.

Defining the Classes

Chen et al (2003)'s results showed that having a multiple threshold-triggering yields the best results. However, as to not to make the model too complex, we limit the classes to just three. Other than a class each for “out-performance” and “under-performance”, we include one where the relative performance is too small to be considered. In our implementation, we choose to ignore any forecast of relative returns within a ±1% range. Hence when designing classes for each asset pair, supposed asset *A* and asset *B*, we consider having a threshold of 1% in setting three possible outcomes in our empirical study, and coded them into +1, 0 and –1:

- +1: Asset *A* out-performs asset *B* by 1%
- 0: Asset *A* and asset *B* have a relative performance of between –1% to +1%.
- –1: Asset *A* under-performs asset *B* by 1% i.e. Asset *B* out-performs asset *A* by 1%

Forecasting the Expected Returns from the Classification Results

To use the results from the classification models in our pair-wise framework, we need more than just a forecast of direction of the relative returns. We need the magnitude of relative returns in order to form the view that can be used to generate the vector of expected returns, which is necessary in mean-variance optimization.

We considered two ways of doing so. First, we can use the relative direction obtained from the classification model to construct a ranking for each asset. For example, supposed we know asset A out-performs asset B, and Asset B out-performs asset C; we can infer that asset A is the best of the three, followed by asset B and then asset C. With this, we can then apply sort-based or rank-based optimization.

This approach, however, has a few problems:

- Asset with no view – We are unable to infer the performance of these assets and thus we will not be able to obtain a ranking for them

- Unrelated pair-wise views – For example, if asset A out-performs asset B, and asset C out-performs asset D, we will not know if asset A out-performs Asset C or vice versa. In this case, these two views are unrelated and thus we are not able to rank asset A and asset C.

In the second approach, we assumed past average performance would be a good estimate for future returns. Hence when we obtained the forecast of the sign of relative returns, we couple it with the average magnitude of the past performance of that sign:

Output from Classifier	**Expected Returns**
+	Average positive returns
0	0
-	Average negative returns

Table 7-a – Three Possible Outputs of Classifier

For example, if the output of the classification is a "+", we will use the average positive returns as the forecast of the expected relative returns.

7.2 Description and Implementation Consideration of the Classification Models Tested

7.2.1 Linear and Quadratic Discriminant Analysis (LDA and QDA)

Discriminant Analysis (DA) is a multivariate statistical technique commonly used to build a predictive/ descriptive model of group "discrimination". Linear Discriminant Analysis (LDA) can handle the case in which the frequencies within the classes are unequal by assuming the covariance of each class is constant. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set, thereby guaranteeing maximal distinction. Quadratic Discriminant Analysis (QDA) is similar to LDA, only dropping the assumption of equal covariance matrices. Therefore the boundary between two discrimination regions is allowed to be a quadratic surface (for example, ellipsoid, hyperboloid, etc) in the maximum likelihood argument with normal distributions.

Huang et al (2005) compared the performances of the two methods in forecasting the direction of weekly Nikkei 225 returns and found that QDA out-performs LDA in term of hit rate. They also compare the covariance matrices of the input variables and found that they are not equal as assumed by LDA, thus attributing the out-performance of the QDA to the relaxation of the equal covariance assumption.

7.2.2 Logistic Regression (Logit)

Logistic Regression is used to find the hyper-plane that best separates the classes. Logistic models belong to the class of generalized linear models (GLM) which generalizes the use of linear regression models to deal with non-normal random variables, and in particular binomial variables. In the context of classification, the binomial variable is an indicator that counts whether an example is in a class or not.

7.2.3 *K*-nearest Neighbor (KNN)

The earliest nonparametric classification method was the nearest neighbor rule of Fix and Hodges (1951), also known as the *k*-nearest neighbor (KNN) rule. The procedure is conceptually simple; we compute the distance from an observation x_i to all other points x_j using a distance function, typically the Euclidean Distance.

To classify x_i into one of the two groups (supposed G_1 and G_2), the k points nearest to x_i are examined, and if the majority of the k points belong to group G_1, then assign x_i to G_1; otherwise assign x_i to the other group G_2. If we denote the number of points from G_1 as k_1, with the remaining k_2 points from G_2, where $k = k_1 + k_2$, then the rule can be expressed as: Assign x_i to G_1 if

$$k_1 > k_2$$

and to G_2 otherwise. If the sample sizes n_1 and n_2 differ, the proportions should be used in place of counts: Assign x_i to G_1 if

$$\frac{k_1}{n_1} > \frac{k_2}{n_2}$$

Determining *k*

The best choice of *k* depends upon the data; generally, larger values of *k* reduce the effect of noise on the classification but make boundaries between classes less distinct.

Loftsgaarden and Quesenberry (1965) suggest using $\sqrt{n_i}$, where n_i is the size of a typical sub-sample, as the estimated value of *k*. The ambiguity arises is the definition of "typical", which can be subjective. Given that that higher *k* is preferred, we used the largest sample size to determine the value of *k* in our implementation:

$$k = \left\lfloor \sqrt{\max(n_1, n_2, ..., n_s)} \right\rfloor$$

where

s = number of sub-samples/ classes

Choosing the distance measure and tie-break rule

The popular choice is the Euclidean norm, for example, Kim (2003) uses Euclidean Distance when comparing against Support Vector Machine in forecasting daily KOSPI index values. However, when dealing with financial time series, the magnitude of the correlation coefficient between each input and the output in the training set is useful when determining the weight of each input when computing the distance measure for a new example. This idea was also implemented by Levitt and Vale (1995) in the trading of foreign exchange.

As such, we will use the following distance measure:

Distance = 1 – (*sample correlation between points*)

In the event of a tie, we assign the sample point to the class which the majority of the *k* nearest neighbors is from.

7.2.4 Decision Tree (Tree)

One method of representing a decision sequences is through a decision tree. The tree represents the decision-making process or a classification process as a series of nested choices or questions. At each step (end of a tree branch) in the process, a single binary or multinomial question is posed (each question is another branch out of the original branch), and the answer determines the next set of choices to be made. In a classification context, this method is sometimes referred to as recursive partitioning.

When the full tree is being generated, it can be pruned to reduce large size and over-fitting the data. This is done by turning some branch nodes into leaf nodes, and removing the leaf nodes under the original tree (Bolakova, 2002). The trees are pruned based on an optimal pruning scheme that the branches giving less improvement in error cost were the first to be pruned.

7.2.5 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a supervised learning technique from the field of machine learning applicable to classification. In Cortes and Vapnik (1995) own words:

> "*The support vector network is a new learning machine for two-group classification problems. The machine conceptually implements the following*

idea: input vectors are non-linearly mapped to a very high dimensional feature space. In this feature space a linear decision surface is constructed."

SVMs are based on the principle of *Structural Risk Minimization*, which searches to minimize an upper bound of generalization error. This induction principle is based on the fact that the generalization error is bounded by the sum of the training error and a confidence interval term that depends on the Vapnik-Chervonenkis (VC) dimension. This makes SVM less susceptible to over-fitting problem as compared to most neural networks. Neural networks implement the *Empirical Risk Minimization* principle, that is, minimizes the training error.

Choosing the kernel function and parameters

Xian (2005) uses a large number of samples to compare the accuracy rates of three kinds of kernel functions of SVM, the results indicate that Gaussian kernel is more accurate than polynomial kernel and multilayer perceptron (MLP) kernel. Gaussian kernel was commonly used when SVM is applied to financial modeling, for example, Kim (2003) and Tay and Cao (2001) both used the Gaussian kernel.

Chang (2005) notes that the results of using Gaussian kernel in support vector classification are sensitive to the size of the kernel width. Small kernel width may cause over-fitting, while a large width may result in under-fitting. The so-called optimal kernel width is selected based on the tradeoff between under-fitting loss and over-fitting loss. Kim (2003) varies the parameter values in various tests and observes hit rates varying between 50.4303% and 57.315% for out-of-sample period ending December 1998. Based on Kim's study, we set the kernel bandwidth to 10, a setting that was also coincides with the kernel width used in Tay and Cao (2001).

Handling more than two-group classification

SVM works for a two-group classification problem. In our case, we have to partition our output universe into three groups (G_1, G_2 and G_3). To handle the limitation in SVM, we first transform our sample set into two groups, the first belong to G_1 and the rest. We then applied SVM classification to identify all data points that belong to group G_1. We then partition the rest of the unclassified data into two groups: G_2 and G_3. SVM classification is then applied to the second set of sample to create the classification for G_2 and G_3.

7.2.6 Probabilistic Neural Network (PNN)

The PNN, proposed by Specht (1988), is a classifier that is built on the Bayesian method of classification, hence is also called Bayesian network. PNN is a type of neural network that uses kernel-based approximation to form estimates of the probability density functions of classes in a classification problem. It is able to deduce the class, with the maximum probability of success, of a given input vector after the training process is completed. As suggested by Chen et al (2003), PNN is appealing because:

- It is fast to train PNN thus suits the frequent update that is required in our contextual model
- The logic of PNN is less susceptible to outliers and questionable data points and thereby reduces extra effort on scrutinizing training data
- PNN provides the Bayesian probability of the class affiliation

7.2.7 Elman Network (ELM)

Elman (1990) proposes a partially recurrent neural network. The network topology contains mainly feed-forward nodes but also includes a set of carefully chosen feedback connections that let the network remembers cues from the recent past. The input layer is divided into the true input units and the context units. The context units hold a copy of the activations of the hidden units from the previous time step, thus allow the network activation produces by past inputs can be feedback to provide additional information for the processing of future inputs. This feedback feature makes ELM an ideal candidate for the financial time series forecasting, as remarked by Brabazon and O'Neill (2006).

Choosing the Number of Neurons and Number of Epochs

The number of hidden neurons varies from implementation to implementation. For example, Wang et al (2006) use 10 neurons while Yu (1999) uses only 6. In more complex settings, the number of neurons can exceed 100 (e.g. Goh et al 2000, Tan and Silva 2003). With the large differences in the number of neurons used, we used in-sample data to decide on the parameter values to be used in our empirical tests. We recorded the hit rates with different combinations of number of neurons and number of epochs. The number of neurons starts from 5 to 50, with a step-size of 5. For the number of epochs, we used a step-size of 100 for the range 100 to 800. The following tables show the hit rates from the two pairs, the USA-Japan pair and Austria-Belgium pair. Our cross-validation analysis suggests that Elman Network containing 25 hidden neurons with a maximum training 500 epochs is the most stable settings. Hence we used these settings in our implementation of the Elman Network.

		Number of Epochs							
		100	200	300	400	500	600	700	800
Number of Neurons	5	66.7	75.0	81.7	76.7	80.0	76.7	76.7	76.7
	10	75.0	76.7	75.0	76.7	76.7	76.7	76.7	81.7
	15	76.7	78.3	76.7	76.7	76.7	83.3	75.0	76.7
	20	73.3	75.0	76.7	75.0	81.7	75.0	73.3	75.0
	25	71.7	75.0	80.0	76.7	76.7	80.0	75.0	75.0
	30	73.3	73.3	76.7	76.7	75.0	75.0	76.7	76.7
	35	76.7	75.0	76.7	75.0	76.7	76.7	80.0	78.3
	40	73.3	73.3	76.7	73.3	80.0	75.0	75.0	76.7
	45	75.0	78.3	75.0	76.7	73.3	75.0	78.3	76.7
	50	73.3	75.0	76.7	75.0	75.0	76.7	73.3	75.0

Table 7-b: Elman Network Hit Rates for USA-Japan Pair with Different Number of Neurons and Epochs

		Number of Epochs							
		100	200	300	400	500	600	700	800
Numb	5	71.7	71.7	71.7	70.0	70.0	75.0	80.0	76.7

	10	71.7	73.3	70.0	75.0	75.0	75.0	73.3	76.7
	15	70.0	71.7	70.0	71.7	78.3	71.7	73.3	73.3
	20	66.7	71.7	71.7	71.7	76.7	73.3	71.7	73.3
	25	73.3	73.3	76.7	70.0	70.0	76.7	78.3	73.3
	30	75.0	71.7	81.7	71.7	71.7	71.7	75.0	73.3
	35	70.0	73.3	71.7	71.7	73.3	80.0	71.7	76.7
	40	71.7	75.0	70.0	71.7	73.3	76.7	71.7	81.7
	45	71.7	73.3	73.3	71.7	73.3	76.7	71.7	78.3
	50	73.3	71.7	70.0	75.0	78.3	76.7	73.3	76.7

Table 7-c: Elman Network Hit Rates for Austria-Belgium Pair with Different Number of Neurons and Epochs

7.3 Empirical Results

To compare the performances of various classification models, including that of regression model, we did two empirical studies. The first, with data from 1995 to 1999, is to apply the classification models to study the hit rates of various classification techniques in forecasting the direction of the three-month ahead relative returns for the six MSCI regional pairs. This is the same preliminary empirical study that we used to review the performance of various momentum and reversal indicators, which we documented in the appendix.

In the second set of empirical tests, we replaced robust regression in our pair-wise framework and apply the classification techniques to construct the forecasting models. All other aspects on how the empirical studies were carried out remained the same to isolate the impact of the different model construction approaches. That is, Similarity Ratio was used to evaluate the pair-wise forecasts, used the same optimizer with the same set of constraints, etc. We then compare the performances of the portfolios constructed using different classification methods with our model portfolios.

7.3.1 Preliminary Hit Rate Analysis

Table 7-d shows the hit rates of different classifiers (including robust regression) from 1995 to 1999 in forecasting the direction of the relative returns of the six pairs of regional assets:

Regional Pair	QDA	LDA	Logit	KNN	Tree	SVM	PNN	ELM	Regression
EU-AP	57.4	54.1	**67.2**	41.0	57.4	41.0	52.5	65.6	59.0
EU-JP	65.6	**67.2**	**67.2**	52.5	52.5	59.0	59.0	60.7	**67.2**
JP-AP	49.2	50.8	44.3	47.5	52.5	44.3	44.3	49.2	**55.7**
NA-AP	49.2	57.4	62.3	45.9	60.7	42.6	59.0	**63.9**	52.5
NA-EU	55.7	57.4	50.8	52.5	52.5	47.5	**59.0**	50.8	57.4
NA-JP	67.2	63.9	**70.5**	54.1	57.4	60.7	68.9	67.2	62.3

Table 7-d: Hit Rates of Different Classifiers in Predicting the Directions of Regional Pair-wise Relative Returns

From the table, we see that only three models generate at least 50% hit rate for all the asset pairs: Linear Discriminant Analysis, Decision Tree and Regression. Among the three, the difference between the maximum and minimum Hit Rate is the smallest for the Decision Tree, followed by the Regression model. This seems to suggest that these models do offer some stability in their forecasts.

	QDA	LDA	Logit	KNN	Tree	SVM	PNN	ELM	Regression
Minimum	49.2	50.8	44.3	41	52.5	41	44.3	49.2	**52.5**
Maximum	67.2	67.2	**70.5**	54.1	60.7	60.7	68.9	67.2	67.2
Range	18	16.4	26.2	13.1	**8.2**	19.7	24.6	18	14.7
Average	57.4	58.5	**60.4**	48.9	55.5	49.2	57.1	59.6	59.0

Table 7-e: Summary of Hit Rates of Different Classifiers in Predicting the Directions of Regional Pair-wise Relative Returns

We summarize the statistics of the hit rates in Table 7-e, we see that Decision Tree and Regression are the classifiers with good performances. They have a minimum hit rate that is above 50%, with smaller differences between the maximum and minimum hit rates then the rest. In addition, the average hit rates are above 50% with maximum hit rates above 60%. Other notable classifiers are LDA and logistic regression.

7.3.2 Performances of Global Country Allocation Portfolios

We shall now look at the performances of the portfolios constructed using these classification techniques. The following table shows the performances of the global country allocation portfolios:

	LDA	QDA	Logit	KNN	Tree	SVM	PNN	ELM	Regress
PI1 (%)	0.51	0.58	0.74	0.75	**1.64**	0.95	0.98	0.51	1.61
PI2	0.29	0.42	0.43	0.39	1.03	0.66	0.75	0.34	**1.15**

	LDA	QDA	Logit	KNN	Tree	SVM	PNN	ELM	Regress
PI3 (%)	57.14	50.00	53.57	57.14	**75.00**	57.14	64.29	60.71	**75.00**
PI4 (%)	**30.01**	35.45	39.74	37.16	45.02	42.57	41.47	42.90	35.49
PI5	**0.07**	0.09	0.1	0.35	0.07	0.44	0.21	0.25	0.12
PI7 (bp)	13.04	14.69	18.70	19.15	**41.06**	14.01	24.55	12.90	40.26

Table 7-f: Performances of Global Portfolios Constructed Using Different Classification Techniques

The empirical results show that Decision Tree has the highest annual value-added, 3bp higher than that of the robust regression model. However, due to a higher tracking error, its Information Ratio is relatively smaller than the regression model. Its average turnover is 4% higher than that of the regression model. Both have a higher hit rate of 75%. Comparing the performance on the pre-sample hit rate and actual empirical performance, robust regression presents the most stable model construction techniques.

Ranking each model with 1 being the best and 5 for the worst, we have the following:

	LDA	QDA	Logit	KNN	Tree	SVM	PNN	ELM	Regress
PI1 (%)	8	7	6	5	**1**	4	3	8	2
PI2	9	6	5	7	2	4	3	8	**1**
PI3 (%)	5	9	8	5	**1**	5	3	4	**1**
PI4 (%)	**1**	2	5	4	9	7	6	8	3
PI5	**1**	3	4	8	2	9	6	7	5
PI7 (bp)	8	6	5	4	**1**	7	3	9	2

Table 7-g: Rankings of Global Portfolios Constructed Using Different Classification Techniques

And the average ranking clearly suggests that the regression model is the best approach in constructing the forecasting models in our pair-wise framework:

Classification Method	Average Ranking
Robust Regression	**2.3**
Decision Tree	2.7
Probabilistic Neural Network (PNN)	4.0
Linear Discriminant Analysis (LDA)	5.3
Quadratic Discriminant Analysis (QDA)	5.5
Logistic Regression	5.5
K-nearest Neighbors (KNN)	5.5

Classification Method	Average Ranking
Support Vector Machine (SVM)	6.0
Elman Neural Network (ELM)	7.3

Table 7-h: Average Ranking of Global Portfolios Constructed Using Different Classification Techniques

Out of the three classification techniques that performed well during our Hit Rate studies (LDA, Logistic Regression, Decision Tree and Robust Regression), only Robust Regression and Decision Tree performed well in terms of portfolio performances.

Cumulative Contribution Curves

From the Cumulative Contribution Curves, we see that indeed the portfolio constructed based on Decision Tree offered the closest competition to the portfolio constructed using Robust Regression:

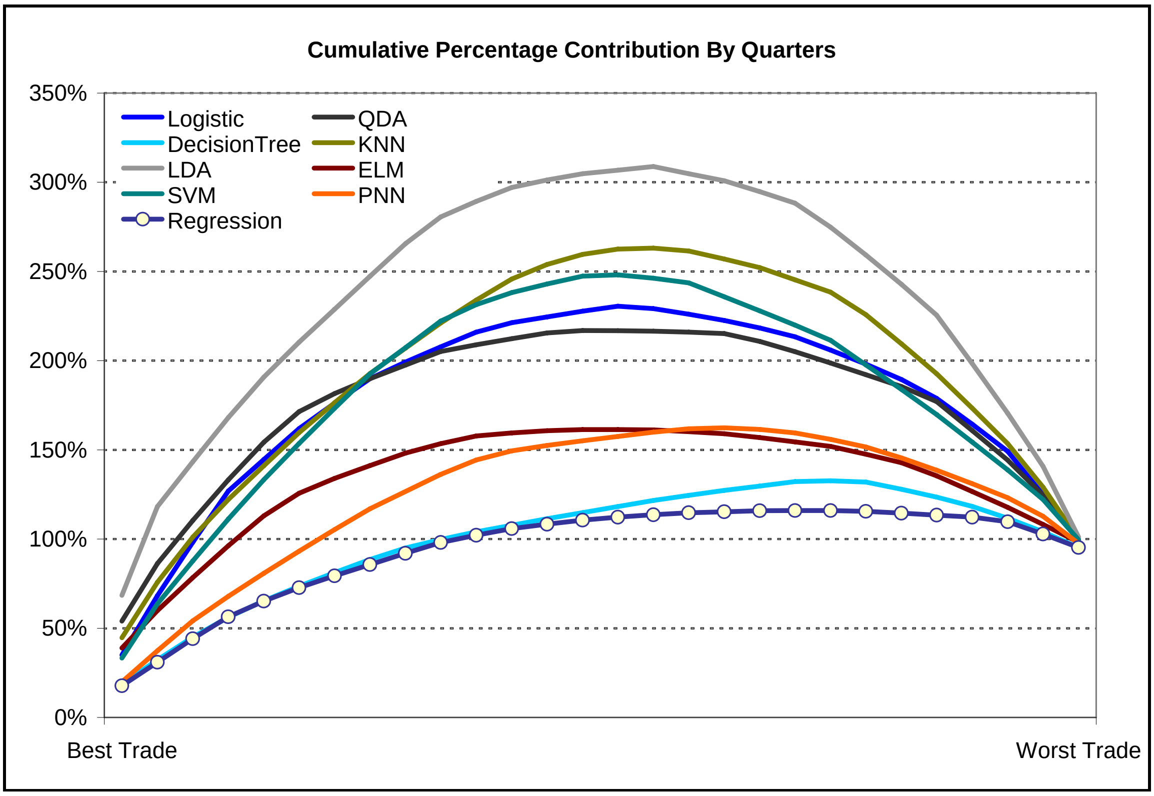

Figure 7-a: Cumulative Contribution Curves of Global Portfolios Constructed Using Different Classification Techniques

7.4 Comparing Decision Tree and Robust Regression in Other Markets

We have seen that the global country allocation portfolio constructed using the Decision Tree is comparable to that of Robust Regression. To confirm if Robust Regression indeed the better model construction technique in our framework, we shall compare these two models in different markets.

7.4.1 Performance Indicators

PI1 – Asset-allocation Value-added

Regression model generated positive value-added in all three markets. It out-performed the Decision Tree model in the two European portfolios but not the emerging Asia portfolio. However, both registered very high value-added in Asia thus the slight under-performance of Regression model should not be seen as a poor result.

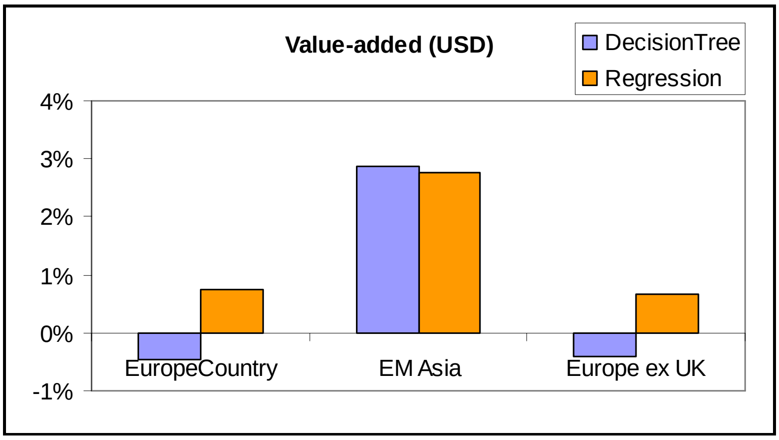

Figure 7-b: Value-added of Portfolios Constructed Using Decision Tree and Robust Regression

PI2 – Information Ratio

The higher value-added in Asia generated by Decision Tree model was achieved with a higher tracking error thus resulting it had a lower information ratio that the Regression model. In terms of Information Ratio, the Regression model clearly done better than the Decision Tree model.

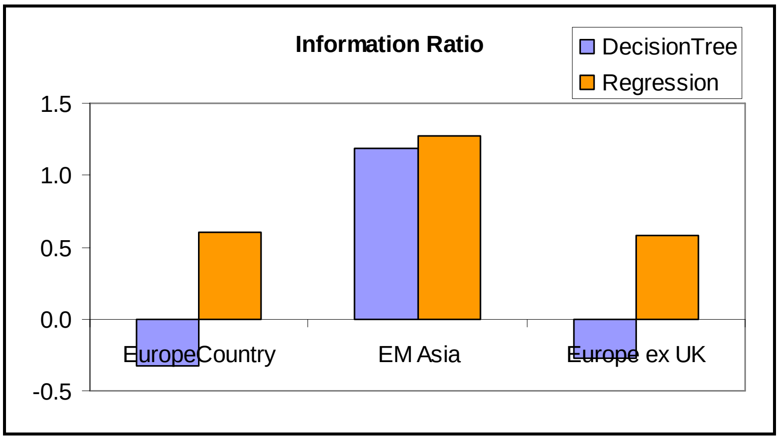

Figure 7-c: Information Ratio of Portfolios Constructed Using Decision Tree and Robust Regression

PI3 – Proportion of Out-performing Quarters

Again the Decision Tree models marginally out-performed in Asia but the Regression model Asia portfolio also managed a high of 77%. In contrast, the Regression model portfolios performed very well in Europe. Generally, the Regression model deliver stable performance overall.

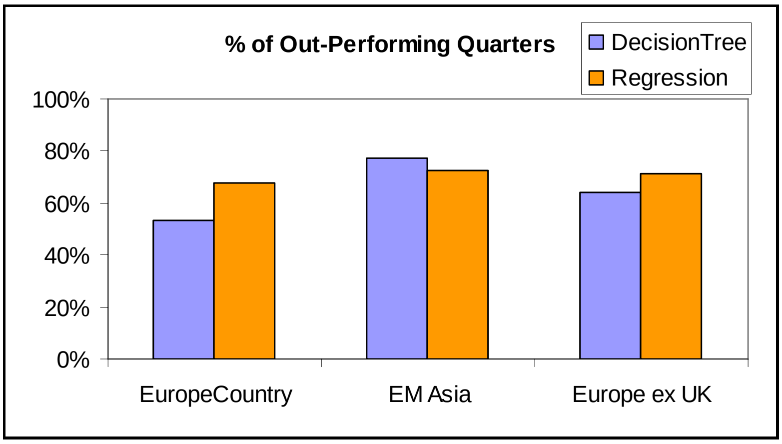

Figure 7-d: Percentage of Out-performing Quarters of Portfolios Constructed Using Decision Tree and Robust Regression

PI4 – Average Turnover

Portfolios constructed with the Regression models generated lower turnover than those portfolios constructed from the Decision Tree models. The differences range from 5% to 10%.

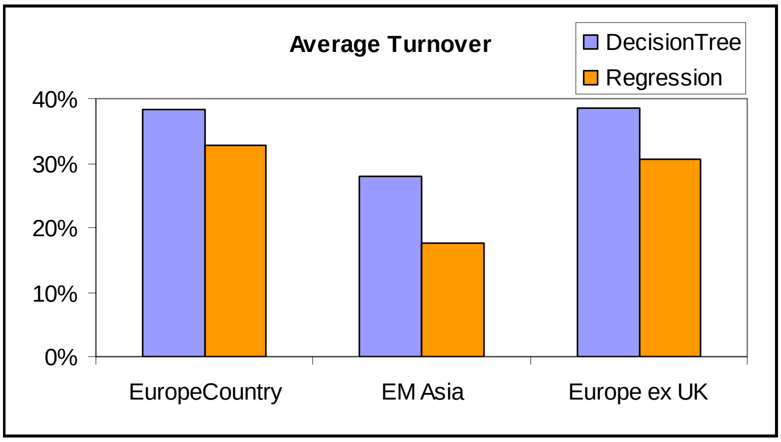

Figure 7-e: Average Turnover of Portfolios Constructed Using Decision Tree and Robust Regression

PI5 – Correlation with Market Performance

We see that the portfolios constructed using the Robust Regression technique have lower correlation with the market performance:

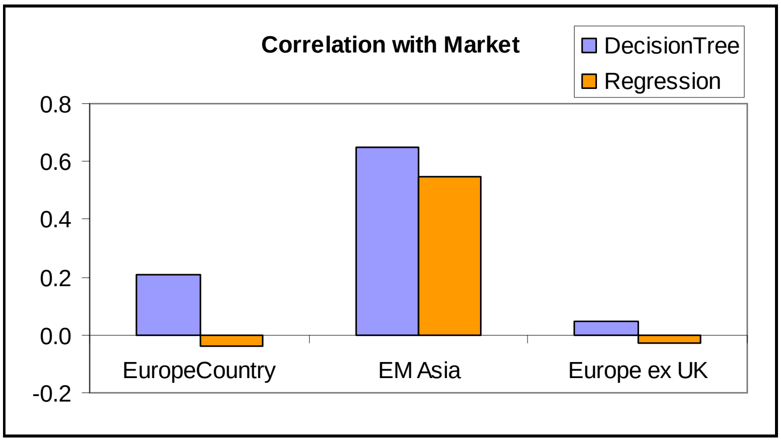

Figure 7-f: Correlation of Value-added with Market Returns of Portfolios Constructed Using Decision Tree and Robust Regression

PI7 – Trading Edge

We see that the portfolios constructed using the Robust Regression technique have positive expected value-added for all three markets. Hence even if the edge is slightly lower than the portfolio constructed using Decision Tree, we conclude that based on Trading Edge, Robust Regression out-performed Decision Tree.

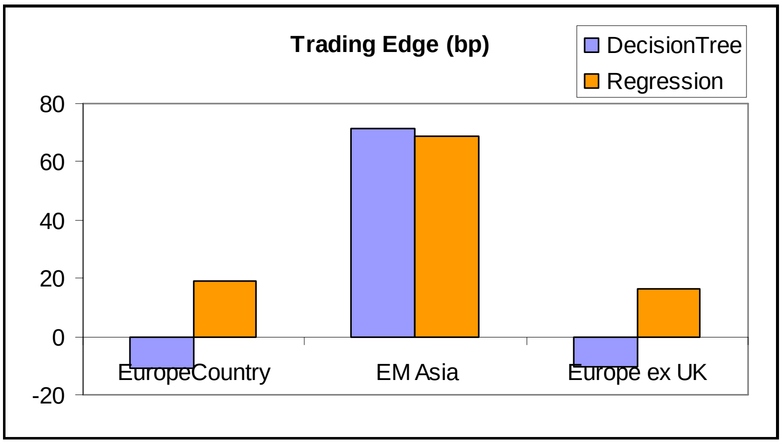

Figure 7-g: Trading Edge of Portfolios Constructed Using Decision Tree and Robust Regression

7.4.2 Verdict

Despite the slight *under*-performance in Asia, Regression models appear to be more stable and able to obtain overall better results in different markets. The outcome of this empirical study has shown that Regression model is the preferred model in our pair-wise framework. Table 7-i summarizes the performances of the test portfolios for different benchmarks. It is clear that portfolios constructed with Robust Regression clearly out-performed those constructed using Decision Tree:

Performance Indicators	Europe Country		Emerging Asia Country		Europe Country ex-UK	
	Decision Tree	Robust Regression	Decision Tree	Robust Regression	Decision Tree	Robust Regression
PI1 (%)	-0.45	**0.75**	2.87	**2.75**	-0.42	**0.66**
PI2	-0.33	**0.61**	1.19	**1.27**	-0.28	**0.58**
PI3 (%)	53.57	**67.86**	**77.27**	72.73	64.29	**71.43**
PI4 (%)	38.27	**32.65**	28.02	**17.62**	38.46	**30.53**
PI5	0.21	**-0.04**	0.65	**0.55**	0.05	**-0.03**
PI7 (bp)	-11.12	**18.85**	**71.7**	68.64	-10.21	**16.51**

Table 7-i: Summary of Performances of Portfolios Constructed Using Decision Tree and Robust Regression

7.5 Ensemble Method or Panel of Experts

7.5.1 Combining Opinions of Different Experts

An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way (typically by weighted or un-weighted voting) to classify new examples. The goal of ensemble methods is to generate an overall system that performs better than any individual classifier.

A great deal of research, including Dietterich (2000), has shown that ensembles are in practice often more accurate than the best of the classifiers they are composed of. Huang et al (2005) tested QDA, LDA, ELM and SVM on forecasting the weekly returns of Nikkei 225 index and compared the results against a combined model comprising of the four models. The weighted average is computed with weights are computed from the in-sample fitting. Using 36-week data in 1992, they found that this weighted results out-performs all individual models.

Qian and Rasheed (2007) tested different combinations with a set of building block models that consist of neural network, *k*-nearest neighbor and decision tree. They concluded that combining three models yields the lowest error rate.

We have seen that our regression model performed well against individual classification models. We will now compare the results of our regression model against the combined opinion of all the classification models.

Hansen and Salamon (1990) found that in order for an ensemble of classifiers to be more accurate than any of its classifiers, the classifiers must be accurate and diverse. Here accurate means better than random guessing and diverse means the classifiers make uncorrelated errors. As such, when we poll the models for their views on the asset pairs relative performance, we only include those “experts” that are at least 50%

confidence in their views. That is, those "experts" that have a track record that surpass the 50% of random guessing.

We consider two ways of combining the outputs of the classifiers:

- **Equally weighted i.e. one vote each –** In this scheme, we will give each model an equal vote. The final result will be the outcome with the highest vote.

- **Weighted-voting using Similarity Ratio –** In this scheme, each model's vote is weighted based on the confidence score, that is, if the score is higher, the model will receive higher weights for its vote.

7.5.2 Empirical Results and Concluding Remarks

The performance of the portfolios constructed using the two ensemble schemes are tabulated as follows. For comparison, we have also included the results we obtained from the regression model:

Performance Indicators	Equal-Weighted	Score-Weighted	Robust Regression
PI1 – Value-added (%)	-0.43	1.18	**1.61**
PI2 – Information Ratio	-0.33	**1.55**	1.15
PI3 – Proportion of Out-performing Quarters (%)	46.43	67.86	**75.00**
PI4 – Average Turnover (%)	40.04	36.86	**35.50**
PI5 – Correlation with Market	**-0.02**	0.18	0.12
PI7 – Trading Edge (bp)	-10.66	29.77	**40.26**

Table 7-j: Performances of Global Portfolios Constructed Using Different Ensemble Schemes

A few conclusions can be drawn from the above results:

(a) Robust regression model performed the best as compared to both ensemble schemes

(b) Among the two schemes, using Similarity Ratio as the weights generates better results than an equally-weighted ensemble scheme.

(c) The ensemble approach that uses Similarity Ratio as the means to weigh the votes out-performed most of the classification techniques except for the Decision Tree approach. The fact that the performance is much closer to the Decision Tree than the rest suggests that Similarity Ratio for the Decision Tree model is likely to be higher than the rest thus getting more weights for its votes.

Based on the empirical results, there is no evidence to suggest that there is a better approach in estimating factor loadings than Robust Regression. This may or may not be true in all contexts but certainly in our proposed pair-wise framework, Robust Regression has performed outstandingly and will be the estimation method that we would recommend in our framework.

8 Conclusion

8.1 Contributions of this Research

Against the backdrop of a lack of application of pair-wise strategies in active portfolio management, our works provide:

- a generic framework to implement pair-wise strategies and extensive empirical evidences to support the framework
- an innovative scoring measure that emphasizes on directional accuracy of relative returns forecasts
- a comparison of regression-based model to other classification techniques in predicting direction of relative returns.

We emphasize that the pair-wise framework does not depend on the type of forecasting models used to predict relative returns or the way in which the forecasts are transformed into a portfolio. What is important is that forecasting model should be built to predict relative returns, and Similarity Ratio is used as a scoring measure. By doing so, we allow investment firms to retain their respective competitive advantage in generating forecasts and portfolio construction while incorporating the pair-wise framework that we propose.

8.1.1 A Framework to Implement Pair-wise Strategies

The framework consists of two stages. In the contextual modeling stage, a contextual model for each asset pair is used to forecast the relative returns of the pair, together with a confidence score. The full set of forecasts then goes through the selection stage in which the set of forecasts to be used in the portfolio construction process is sieved out. As the confidence score is used in the selection process, it is critical that the score reflects as accurately as possible the forecasting model's ability in relative returns prediction. We propose Similarity Ratio as the ideal scoring measure to be used.

8.1.2 Innovative Scoring Measure that Emphasizes on Directional Accuracy

We designed Similarity Ratio to quantify the forecasting quality of a model. The scoring mechanism embedded in Similarity Ratio uses directional accuracy as the main consideration when assigning a score, and supplements with the magnitude of forecast.

For an actual-forecast pair (a, f), Similarity Ratio is defined as:

$$\text{Similarity Ratio }(a, f) = \begin{cases} 0 & \text{if } a^2 + af \le 0 \\ \dfrac{|f+a|}{|f+a| + |f-a|} & \text{otherwise} \end{cases}$$

Similarity Ratio for a model will be the average Similarity Ratio for every out-of-sample actual-forecast pair generated from the model.

The intuition of Similarity Ratio can be seen geometrically; the measure is derived from the distances of orthogonal projections of each actual-forecast pair to **BOTH** the *Good* ($y = x$) and *Bad* ($y = -x$) lines.

Not only did the portfolios constructed using Similarity Ratio out-performed other scoring measures empirically, Similarity Ratio also exhibits important characteristics of an ideal scoring measure:

- Model with a higher accuracy in forecasting the direction of relative returns (hit rate) has a higher score.

- If there are two models having the same hit rate, higher score will be given to the model with more accurate forecasts, that is the deviations from actual values observed is smaller.

- The model's score is not susceptible to the presence of outlier in the sample.

8.1.3 Comparison of Regression Model with Classification Techniques

We conducted extensive empirical study to find evidences to support that regression approach still works in the pair-wise framework. The empirical results provide invaluable evidences to show that the portfolios constructed using regression-based forecasting models out-performed those constructed using classification techniques such as Neural Network, Discriminant Analysis, etc. In addition, this study also shows that it is possible to incorporate different forecasting models into the framework and this supports our claim that the framework is model-independent.

8.2 Empirical Evidences for the Pair-wise Framework

The global country allocation portfolio benchmarked against MSCI World was constructed using the pair-wise framework that we had proposed. The portfolio delivered an annual value-added of 1.61% over its benchmark returns from 2000 to 2006. This was achieved with a low tracking error and led to a high Information Ratio of 1.15.

Based on the statistics that we have obtained from Mercer, we see that the top quartile managers with a MSCI World mandate have an Information Ratio of 0.99 and 0.6 for the five-year and ten-year period respectively. Comparing the Information Ratio alone, our country allocation portfolio clearly ranked among the top quartile performers, whom their Information Ratios are listed in the following table:

Period	**3-Year**	**5-Year**	**10-Year**
Median IR	0.36	0.47	0.32
Top Quartile IR	1.05	0.99	0.6

Number of Managers	192	155	80

Source: Mercer Global Investment Manager Database

Table 8-a: Information Ratios of Median and Top Quartile Managers with Global Mandates

8.2.1 Pair-wise Model Yields Better Results Than Individual Model

Our empirical results shows that modeling the relative returns of asset pairs (pair-wise model) yields better results than modeling each asset returns separately (individual model). The following table compares the performance of the two global country allocation portfolios constructed using pair-wise model and the individual model, and the pair-wise model clearly out-performed the individual model in all aspects:

Performance Indicators	Individual Model	Pair-wise Model
PI1 – Asset Allocation Value-added (%)	0.33	**1.61**
PI2 – Information Ratio	0.66	**1.15**
PI3 – Proportion of Out-performing Quarters (%)	53.57	**75.00**
PI4 –Turnover (%)	13.47	**35.50**
PI5 – Correlation with Market	-0.18	**0.12**
PI7 – Trading Edge (bp)	8.45	**40.26**

Table 8-b: Performances of Global Portfolios Constructed Using Individual and Pair-wise Models

Pair-wise Model Out-performed Regardless of Scoring Measure

When pair-wise model is used, the portfolios constructed out-performed the individual model, regardless of which scoring measure we used. This is supported by the empirical results obtained from comparing different pair-wise portfolios constructed using different scoring measures. Table 8-d compares the performances of all the pair-wise global country allocation portfolios constructed using different scoring measures and the portfolio constructed using the individual model with Similarity Ratio.

Almost all portfolios constructed under the pair-wise framework generated a value-added of more than 1%, with the exception of the portfolio that used Theil Forecast Accuracy as the scoring measure. These results were much higher than the 0.33% value-added generated by modeling each asset's returns individually, even the portfolio that used Theil Forecast Accuracy delivers a value-added that is twice that of the individual model portfolio. In fact, the individual model portfolio also has the lowest Information Ratio, proportion of out-performing quarters, and trading edge. Thus these results provide another indication that using pair-wise model generates better results, even if the scoring measure used was not Similarity Ratio.

Pair-wise Model Out-performed Regardless Model Construction Approaches

When we replaced robust regression with different classification techniques in the pair-wise framework, the performances of the global country allocation portfolios yield different results but all of them out-performed against the portfolio constructed using individual model:

These results are presented in Table 8-e: and they provide more empirical evidences to support our proposition that modeling assets in pairs yield better results, regardless the choice of model construction methods used to implement the framework.

Pair-wise Model Out-performed When Perfect Forecasts Were Used as Inputs

When actual relative returns were used as the views in the pair-wise framework, the portfolios constructed out-performed those that used actual assets' returns. This suggests that better quality relative forecasts yields better results than absolute forecasts of the same quality. The performances of the global portfolios constructed using the pair-wise and individual models are presented below:

Performance Indicators	Individual Model	Pair-wise Model
PI1 – Asset Allocation Value-added (%)	5.11	**7.75**
PI2 – Information Ratio	2.43	**4.83**
PI3 – Proportion of Out-performing Quarters (%)	89.29	**100.00**
PI4 –Turnover (%)	**33.15**	49.54
PI5 – Correlation with Market	-0.3362	**0.0058**
PI7 – Trading Edge (bp)	126	**189**

Table 8-c: Performances of Global Portfolios Constructed with Perfect Forecasts

Performance Indicators	Individual Model	Pair-wise Models						
	Similarity Ratio	R^2	Hit Rate	Info Coeff (IC)	Un-centered IC	Anomaly Corr	Theil Forecast Accuracy	Similarity Ratio
PI1 (%)	0.33	1.12	1.48	1.28	1.09	1.31	0.66	**1.61**
PI2	0.66	0.73	0.98	0.90	0.93	0.91	0.74	**1.15**
PI3 (%)	53.57	67.86	71.43	67.86	71.42	64.29	57.14	**75.00**
PI4 (%)	**13.47**	37.45	35.46	33.81	36.41	35.59	17.43	35.49
PI5	-0.18	0.20	0.29	**-0.08**	-0.27	0.09	0.14	0.12
PI7 (bp)	8.45	28.26	37.05	31.98	32.44	32.90	16.61	**40.26**

Table 8-d: Performances of Individual Model Against Pair-wise Model with Different Scoring Measures

Performance Indicators	Individual Model	Pair-wise Models								
	Regress	LDA	QDA	Logit	KNN	Tree	SVM	PNN	ELM	Regress
PI1 (%)	0.33	0.51	0.58	0.74	0.75	**1.64**	0.95	0.98	0.51	1.61
PI2	0.66	0.29	0.42	0.43	0.39	1.03	0.66	0.75	0.34	**1.15**
PI3 (%)	53.57	57.14	50.00	53.57	57.14	**75.00**	57.14	64.29	60.71	**75.00**
PI4 (%)	**13.47**	30.01	35.45	39.74	37.16	45.02	42.57	41.47	42.90	35.49
PI5	-0.18	**0.07**	0.09	0.1	0.35	**0.07**	0.44	0.21	0.25	0.12
PI7 (bp)	8.45	13.04	14.69	18.70	19.15	**41.06**	14.01	24.55	12.90	40.26

Table 8-e: Performances of Individual Model Against Pair-wise Model with Different Classification Techniques

8.2.2 Similarity Ratio as a Scoring Measure Picks Better Forecasts to Use

Other than Similarity Ratio, we also tested six different scoring measures in constructing global country allocation portfolios. For each performance indicator, we assigned "1" for the model that performed best and "7" for the worst. Table 8-f shows the average ranking for all the seven candidates and clearly Similarity Ratio came in top:

Scoring Method	Average Ranking
Similarity Ratio	**1.8**
Hit Rate	3.0
Information Coefficient	3.5
Anomaly Correlation	3.8
Un-centered IC	4.6
Theil Forecast Accuracy	5.3
R^2	5.7

Table 8-f: Average Ranking of Global Portfolios Constructed Using Different Scoring Measures

To confirm that Similarity Ratio is indeed a better scoring model to use, we also constructed country portfolios in Europe and Emerging Asia using Similarity Ratio and its closest "rival", the Hit Rate. Once again, the various performance indicators have indicated that the portfolios constructed using Similarity Ratio out-performed those that were constructed based on Hit Rate:

Performance Indicators	Europe Country		Emerging Asia Country		Europe Country ex-UK	
	Hit Rate	Similarity Ratio	Hit Rate	Similarity Ratio	Hit Rate	Similarity Ratio
PI1 (%)	0.22	**0.75**	2.18	**2.75**	-0.22	**0.66**
PI2	0.22	**0.61**	0.91	**1.27**	-0.20	**0.58**
PI3 (%)	53.57	**67.86**	68.18	**72.73**	50.00	**71.43**
PI4 (%)	**32.24**	32.65	20.33	**17.62**	**29.90**	30.53
PI5	0.08	**-0.04**	0.56	**0.55**	0.12	**-0.03**
PI7 (bp)	12.59	**18.85**	**69.54**	68.64	14.40	**16.51**

Table 8-g: Summary of Performances of Portfolios Constructed Using Hit Rate and Similarity Ratio

8.2.3 Empirical Results Support Our Propositions

Based on the comprehensive set of evaluation criteria to compare the performances of test portfolios, our empirical results support our claims that:

- Portfolios that are constructed using relative returns of asset pairs (pair-wise model) yield better results than those based on returns forecasts of individual asset (individual model).

- Portfolios constructed using Similarity Ratio as the scoring measure to rank relative returns forecasts out-performed those that used other scoring measures.

As these are the pillars of the pair-wise framework, hence the empirical results also show that the pair-wise framework works in constructing country allocation portfolios.

8.3 Comparison with Classification Techniques

In the implementation of our pair-wise framework, we proposed to use robust regression to estimate the factor loadings of our forecasting model. As the framework is generic, different model construction approaches can be used in paced of robust regression. We explored the possibilities of using classification techniques to model the direction of each asset pair's relative returns. We considered a total of eight different classification models, from statistical-based (e.g. Discriminant Analysis) to neural networks (e.g. Elman Network).

We constructed eight global country portfolios using the pair-wise framework by replacing robust regression with these classification models. The empirical results showed that the portfolio constructed using robust regression as the model estimation method performs best for the set of performance indicators we have selected.

For each performance indicator, we ranked the various models with "1" being the best and "9" for the worst. The average rankings for each classification model are as follows:

Classification Method	Average Ranking
Robust Regression	**2.3**
Decision Tree	2.7
Probabilistic Neural Network (PNN)	4.0
Linear Discriminant Analysis (LDA)	5.3
Quadratic Discriminant Analysis (QDA)	5.5
Logistic Regression	5.5
K-nearest Neighbors (KNN)	5.5
Support Vector Machine (SVM)	6.0
Elman Neural Network (ELM)	7.3

Table 8-h: Average Ranking of Global Portfolios Constructed Using Different Classification Techniques

We see that robust regression model came in first with an average ranking of 2.3 out of 9, with Decision Tree model came in a distant second of 2.7 out of 9. Further comparison of the two models in different markets (Europe and Emerging Asia) showed that robust regression model portfolios out-performed those based on Decision Tree.

Performance Indicators	Europe Country		Emerging Asia Country		Europe Country ex-UK	
	Decision Tree	Robust Regression	Decision Tree	Robust Regression	Decision Tree	Robust Regression
PI1 (%)	-0.45	**0.75**	2.87	**2.75**	-0.42	**0.66**
PI2	-0.33	**0.61**	1.19	**1.27**	-0.28	**0.58**
PI3 (%)	53.57	**67.86**	**77.27**	72.73	64.29	**71.43**
PI4 (%)	38.27	**32.65**	28.02	**17.62**	38.46	**30.53**
PI5	0.21	**-0.04**	0.65	**0.55**	0.05	**-0.03**
PI7 (bp)	-11.12	**18.85**	**71.7**	68.64	-10.21	**16.51**

Table 8-i: Summary of Performances of Portfolios Constructed Using Decision Tree and Robust Regression

It may not be fair to say that robust regression is the most powerful in all context, and the result need not necessary suggest that various classification techniques will never outperform the regression model. Perhaps classification techniques require more data for training, more inputs to learn the results, and those different types of models work for different markets. However, at this stage, we have seen enough to continue to use robust regression in our portfolio construction framework.

8.4 Conclusion

The pair-wise framework is based on the finding that the performance of asset allocation portfolio is mainly driven by the directional accuracy of pair-wise relative returns forecasts. The framework recommends that forecasting models are to be "customized" (i.e. contextual model) to predict the relative returns of each possible pair-wise combination of assets.

MSCI World consists of 23 countries; this means we need only a maximum of 22 pair-wise forecasts out of the possible 253 forecasts generated. This leaves us with less than 9% chance of success if the selection of pairs is done randomly. Similarity Ratio is a scoring measure that emphasizes on the directional accuracy of the forecasting models and we found it to be more superior to those alternative scoring measures that we have considered, including the popular Information Coefficient. Given the fact that the portfolios constructed based on Similarity Ratio have deliver good empirical performances, we believe that it is a scoring measure that is an effective measure in the pair-wise framework.

Results from our empirical study provide evidences to support our propositions of using contextual pair-wise modeling of relative returns and the use of Similarity Ratio as the scoring measure. Hence we believe that the pair-wise framework is ideal in constructing country allocation portfolio, that is:

- Portfolios that are constructed using relative returns of asset pairs (pair-wise model) yield better results than those based on returns forecasts of individual asset (individual model).

- Portfolios constructed using Similarity Ratio as the scoring measure to rank relative returns forecasts out-performed those that used other scoring measures.

Bibliography

Aggarwal, R. and Demaskey. A. (1997). “Using Derivatives in Major Currencies for Cross-hedging Currency Risks in Asian Emerging Markets”. *Journal of Futures Markets*, 17.

Alexander, C. (2001). *Market Models*. John Wiley & Sons.

Alexander, C. and Dimitriu, A. (2004). “Sources of Outperformance in Equity Markets: Mean Reversion, Common Trends and Herding”. *Journal of Portfolio Management*. Special European Issue (2004).

Alexander and Weddington (2001). “Cointegration and Asset Allocation: A New Active Hedge Fund Strategy”. *http://inquire.org.uk.loopiadns.com/inquirefiles/Attachments/Brighton2001/Paper-AlexanderAlln.PDF*

Alford, A., Jones, R. and Lim, T. (2003). “Equity Portfolio Management”. *Modern Investment Management*. Wiley Finance.

Anderson, C. and Schumacker, R. E. (2003). “A Comparison of Five Robust Regression Methods with Ordinary Least Squares Regression: Relative Efficiency, Bias, and Test of the Null Hypothesis”. *Understanding Statistics*, 2(2).

Androutsost, D., Plataniotist, K. N. and Venetsanopoulost, A. N. (1998). “Distance Measures for Color Image Retrieval”. *IEEE Xplore*, Vol. 2.

Baestaens, D. E. and Bergh, W. M. (1995). “Tracking the Amsterdam Stock Index Using Neural Networks”. *Neural Networks in the Capital Markets* p149-161. John Wiley & Sons.

Balvers, R. J. and Wu, Y. (2005). “Momentum and Mean Reversion Across National Equity Markets”. *Journal of Empirical Finance*, Vol. 13.

Best, M. J. and Grauer, R. R. (1991). “On the Sensitivity of Mean-Variance-Efficient Portfolios to Changes in Assets Means: Some Analytical and Computational Results”. *Review of Financial Studies* 4.

Bliemel, F. and Mackay, D. B. (1973). “Theil’s Forecast Accuracy Coefficient: A Clarification”. *Journal of Marketing Research*, Vol. X, Nov 1973.

Bolakova, I. (2002). “Pruning Decision Trees To Reduce Tree Size”.

Brabaxon, A. and O’Neill, M. (2006). *Biologically Inspired Algorithms for Financial Modelling*. Springer 2006.

Breiman, L. (1996). “Bagging Predictors”. *Machine Learning*, 24(2).

Brinson, G. P., Randolph, H. and Gilbert L. Beebower. (1986) , “Determinants of Portfolio Performance,” Financial Analysts Journal, July-August 1986, pp. 39-44.

Brinson, Gary P., Singer, Brian D., and Gilbert L. Beebower, “Determinants of Portfolio Performance II: An Update,” Financial Analysts Journal, May-June 1991, pp. 40-48.

Bulsing, M. and Sefton, J. (2004). “The Hurdle of Credulity”. *UBS Investment Research*.

Chan, K. C. L., Jegadeesh, N., Lakonishok, J. (1996). “Momentum Strategies”. *Journal of Finance,* Vol. 51 No. 5

Chang, Q., Chen, Q. and Wang, X. (2005). “Scaling Gaussian RBF kernel width to improve SVM classification”. *Neural Networks and Brain*. Oct 2005

Chen, A., Leung, M. T. and Daouk, H. (2003). “Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index”. *Computers & Operations Research*, Vol. 30 Issue 6.

Chen, H., and DeBondt, W. (2004). “Style momentum within the S&P-500 index“. *Journal of Empirical Finance*. Volume 11, Issue 4.

Chen, W., Shih, J. and Wu, S. (2006). Comparison of support-vector machines and back propagation neural networks in forecasting the six major Asian stock markets”. *International Journal of Electronic Finance*, Vol. 1, No. 1.

Chincarini, L. B. and Kim, D. (2006). *Quantitative Equity Portfolio Management*. McGraw-Hill Library of Investment and Finance.

Chopra, N., Lakonishok, J. Ritter, J. R. (1992). “Measuring Abnormal Performance: Do Stocks Over-react?”. *Journal of Finance* 57.

Chou, P., Wei, J. and Chung, H. (2007). “Sources of contrarian profits in the Japanese stock market”. *Journal of Empirical Finance*. Volume 14, Issue 3.

Comes, B. and Kelemen, A. (2003). “Probabilistic Neural Network Classification for Microarraydata”. *IEEE* 2003.

Cortes, C. and Vapnik, V. (1995). ”Support vector networks”. *Machine Learning*, 20.

Covel, M. W. (2007). *The Complete TurtleTrader: The Legend, the Lessons, the Results.*

DeBondt, W. and Thaler, R. (1985). “Does the Stock Market Overreact?”. *Journal of Finance* 40.

DeBondt, W. and Thaler, R. (1987). “Further Evidence on investor overreaction and Stock Market Seasonality”. *Journal of Finance* 42.

DeMiguel, V., Garlappi, L. and Uppal, R. (2005). “How Inefficient are Simple Asset-allocation Strategies”.

Dietterich, T. G. (2000). “Ensemble Methods in Machine Learning”. *Lecture Notes in Computer Science*

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. Weily-Interscience

Elman, J. L. (1990). “Finding Structure in Time”. *Cognitive Science*, 14.

Emran, S. M. and Ye, N. (2001). “Robustness of Canberra Metric in Computer Intrusion Detection”. *Proceedings of the 2001 IEEE Workshop on Information Assurance and Security*.

Fabozzi, F. J., Focardi, S. M. and Kolm, P. N. (2006). *Financial Modeling of the Equity Market: From CAPM to Cointegration*. Wiley Finance.

Fanelli, P. R. and Urias, M. S. (2002). "Designing the Investment Process for an Active Equity Fund". *Inside the Investment Process*. Morgan Stanley Global Equity and Derivative Markets, Sep 2002.

Forsythe, A. B. (1972), "Robust Estimation of Straight Line Regression Coefficients by Minimizing p^{th} Power Deviations". *Technometrics*, 14.

Frost, P. A. and Savarino, J. E. (1988). "For Better Performance: Constraints Portfolio Weights". *Journal of Portfolio Management*, 15.

Goh, W. Y., Lim, C. P. and Peh, K. K. et al (2000). "A Neural-Network-Based Intelligent System for Time-Series Prediction Problems in Product Development". *IEEE* 2000.

Green, R. C. and Hollifield, B. (1992). "When Will Mean-Variance Efficient Portfolios Be Well Diversified". *Journal of Finance*, 47.

Green, H. and Pearson, M. (1994). "Neural Nets for Foreign Exchange Trading". *Trading on the Edge. Wiley*. p123-129.

Griffin, J. M., Ji, X. and Martin, J. S. (2003). "Momentum Investing and Business Cycle Risk: Evidence from Pole to Pole". *The Journal of Finance*, Vol. 58, No. 6.

Griffin, J. M., Ji, X. and Martin, J. S. (2005). "Global Momentum Strategies". *Journal of Portfolio Management*. Winter 2005.

Grinblatt, M. and Moskowitz, T. J. (2004). "Predicting Stock Price Movements From Past Returns: The Role of Consistency and Tax-loss Selling". *Journal of Financial Economics* 71.

Guerard, J. B. (2006). "Quantitative Stock Selection in Japan and the United States: Some Past and Current Issues". *Journal of Investing*. Vol. 15, Issue 1.

Hansen, L.K.; Salamon, P. (1990). "Neural network ensembles". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 12, Issue 10.

Hartzmark, M. L. (1991). "Luck Versus Forecast Ability: Determinants of Trader Performance in Futures Markets". *The Journal of Business*, Vol. 64, No. 1.

Hong, H. and Stein, J. C. (1999). "A Unified Theory of Underreaction, Momentum Trading". Journal of Finance 54.

Huang, W., Nakamori, Y. and Wang, S. (2005). "Forecasting Stock Market Movement Direction with Support Vector Machine". *Computers & Operations Research* 32.

Idzorek, T. (2004). "A Step-By-Step Guide To The Black-Litterman Model – Incorporating user-specified confidence level". http://www.mccombs.utexas.edu/faculty/keith.brown/ChileMaterial/Idzorek%20WP.pdf

Jegadeesh, N. (1990). "Evidence of Predictable Behaviour of Security Returns". *Journal of Finance* 45.

Jegadeesh, N. and Titman, S. (1993). "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency". *Journal of Finance* 48.

Jobson, J. D. and Korkie, B. M. (1981). “Putting Markowitz Theory to Work”. *Journal of Portfolio Management*, 7.

Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis*, 4th Edition. Prentice Hall.

Jorion, P. (1991). “Bayesian and CAPM Estimators of the Means: Implications for Portfolio Selection”. *Journal of Banking and Finance*, 15.

Kim, K. (2003). “Financial Time Series Forecasting Using Support Vector Machines”. *Neurocomputing* 55 (2003)

Kohara, K. (2003). “Selective-Learning-Rate Approach for Stock Market Prediction by Simple Recurrent Neural Networks”. *Knowledge-based Intelligent Information and Engineering Systems 2003*. p141-147. Springer-Verlag Berlin Heidelberg 2003.

Krishnamurtia, T. N., Rajendrana, K., Vijaya Kumara, *et al* (2003). “Improved Skill for the Anomaly Correlation of Geopotential Heights at 500 hPa”. *American Meteorological Society Journal*. Volume 131, Issue 6.

Kung, E. and Pohlman, L. (2004). “Portable Alpha”. *Journal of Portfolio Management*. Vol. 30, Issue 3.

Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2005). *Applied Linear Statistical Models*. Fifth Edition. McGraw Hill International Edition.

Kwaitowski, D., Phillips, P., Schmidt, P., and Shin, Y. (1992). “Testing the Null Hypothesis of Stationarity against the Null Hypothesis of a Unit Root”. *Journal of Econometrics*.

Lachenbruch, P. and Mickey, R. (1968). “Estimation of Error Rates in Discriminant Analysis”. *Technometrics*, 10.

Laker, D. (2000). “What is this Thing Called Interaction?” *Journal of Performance Measurement*, Fall 2000.

Lemann, B. (1990). “Fads, Martingales and Market Efficiency”. *Quarterly Journal of Econom*. 105

Levich, R. M. and Thomas, L. R. (1981). “How to Compare Chance with Forecasting Expertise”. *Euromoney*, Aug 1981.

Levitt, M. E. and Vale, S. (1995). “Machine Learning for Foreign Exchange Trading”. *Neural Networks in the Capital Markets* p233-243. John Wiley & Sons.

Li, Z. (2005). “The Persistence of Runs – The Directional Movement of Index Returns”. *Quarterly Journal of Business and Economic*, 44.

Liu, C. and Lee, Y. (2001). “Does the Momentum Strategy Work Universally? Evidence from the Japanese Stock Market”. *Asia Pacific Financial Markets* 8.

Loftsgaarden, D. O. and Quesenberry, C. P. (1965). “A Non-parametric Estimate of a Multivariate Density Function”. *Annals of Mathematical Statistics*, 36.

Michie, D., Spiegelhalter, D. J. and Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood Series in Artificial Intelligence

Miyakoda, K., Sirutis, J. and Ploshay, J. (1986). “One-month Forecast Experiments – Without Anomaly Boundary Forcings”. *Monthly Weather Review*. Volume 114.

Palmer, T. N., Brancovic, C., Molteni, F. *et al*. (1990). “The European Center for Medium-Range Weather Forecasts (ECMWF) Program on Extended-range Prediction”. *Bulletin in American Meteorological Society*. Volume 71

Poddig, T. (1995). “Bankruptcy Prediction: A Comparison with Discriminant Analysis”. *Neural Networks in the Capital Markets* p311-323. John Wiley & Sons.

Pretorius, M., Rossi, G. D. and Jessop, D. (2007). “Alpha+: Pairs”. *UBS Investment Research*.

Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann.

Qian, B. and Rasheed, K. (2007). “Stock Market Prediction with Multiple Classifiers”. *Application Intelligence* 26.

Qian, E. (2003). “Tactical Asset Allocation with Pairwise Strategies”. *Journal of Portfolio Management.* Fall 2003.

Radok and Brown (1993). “Anomaly Correlation and an Alternative: Partial Correlation”. *Monthly Weather Review*. Volume 121.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*. Second Edition. Wiley Series in Probability and Statistics

Refenes, A. N. (1993). “Constructive Learning and its Application to Currency Exchange Rate Forecasting”. *Neural Networks in Finance and Investing*. p777-805. Irwin.

Refenes, A. N., Zapranis, A. and Francis, G. (1994). “Stock Performance Modeling Using Neural Networks: A Comparative Study With Regression Models”. *Neural Networks*, Vol 7, No. 2, pp. 375-388.

Refenes, A. P. (1995). “Testing Strategies and Metrics”. *Neural Networks in the Capital Markets* p67-76. John Wiley & Sons.

Rey, D. M. and Schmid, M. M. (2007). “Feasible Momentum Strategies: Evidence from Swiss Stock Market”. *Swiss Society for Financial Market Research 2007*.

Rieck, K., Laskov, P. and Muller, K. (2006). “Efficient Algorithms for Similarity Measures over Sequential Data: A Look Beyond Kernels”. *DAGM 2006*, LNCS 4174, pp. 374–383.

Roman, J. and Jameel, A. (1996). “Backpropagation and Recurrent Neural Networks in Financial Analysis of Multiple Stock Market Returns”. *Proceedings of the 29th Annual Hawaii International Conference on System Sciences*.

Rousseeuw, P.J. (1984). “Least median of squares regression”. *Journal of the American Statistical Association*, 79.

Rouwenhorst, K. (1998). “International Momentum Strategies”. *Journal of Finance,* Vol 53 No. 1.

Satchell, S. and Scowcroft, A. (2000). “A Demystification of the Black-Litterman Model: Manageing Quantitative and Traditional Portfolio Construction”. *Journal of Asset Management*, Vol. 1, 2.

Schapire, R. E., Freund, Y. and Bartlett, P et al. (1997). “Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods”. *Proceedings of the 14th International Conference on Machine Learning*

Scheidat, T. Vielhauer, C. and Dittmann, J. (2005). “Distance-Level Fusion Strategies For Online Signature Verification”. *IEEE International Conference*, Jul 2005.

Siegel, L. B. (2003). *Benchmarks and Investment Management*. The Research Foundation of AIMR.

Sorensen, E. H., Hua, R. And Qian, E. (2005). “Contextual Fundamentals, Models, and Active Management”. *Journal of Portfolio Management*, Vol. 31 No. 1

Sorensen, E. H., Hua, R. And Qian, E. (2007). “Aspects of Constrained Long-Short Equity Portfolios”. *Journal of Portfolio Management*, Vol. 33 No. 2

Steurer, E. (1993). “Nonlinear Modeling of the DEM/USD Exchange Rate”. Neural Networks in the Capital Markets. P199-211. John Wiley & Sons

Straumann, D. and Garidi, T. (2007). “Developing an Equity Factor Model for Risk”. *Risk Metrics Journal*, Volume 7, Number 1.

Tan, H. C. C. and Silva, L. C. (2003). “Human Activity Recognition by Head Movement using Elman Network and Neuro-Markovian Hybrids”. *Image and Vision Computing NZ*, 2003.

Tay, F. and Cao L. (2001). “Application of Support Vector Machines in Financial Time Series Forecasting”. *Omega – The International Journal of Management Science 29.*

Trafalis, T. B. and Ince, H. (2000). “Support Vector Machine for Regression and Applications to Financial Forecasting”. *IJCNN2000*. p348-353.

Tracton, M. S., Mo K. and Chen, W. (1989). “Dynamical Extended Range Forecasting (DERF) at the National Meteorological Center”. *Monthly Weather Review*. Volume 117.

Tsay, R. S. (2002). *Analysis of Financial Time Series*. Wiley Series in Probability and Statistics

Walczak, S. (2001). “An Empirical Analysis of Data Requirements for Financial Forecasting with Neural Networks”. *Journal of Management Information Systems*, Vol. 17 No. 4.

Wang, L., Liang, Y. and Shi, X. et al. (2006). “An Improved OIF Elman Neural Network and its Application to Stock Market”. *Knowledge-based Intelligent Information and Engineering Systems 2006*. p21-28. Springer-Verlag Berlin Heidelberg 2006.

Waring, M. B., and Siegel, L. B. (2003). “The Dimensions of Active Management.” *Journal of Portfolio Management*, vol. 29, no. 3 (Spring):35–51.

Weigend, A., Rumelhard, D. and Huberman, B. (1991). “Generalization by Weight Elimination with Application to Forecasting”. *Advances in Neural Information Processing Systems*. Vol 3.

Wolfram Mathematica ___. http://reference.wolfram.com/mathematica/ref/CanberraDistance.html

Wolberg, J. R. (2000). *Expert Trading Systems: Modeling Financial Markets with Kernel Regression*. John Wiley 2000.

Wolpert, D. H. (1992). “Stacked Generalisation”. *Neural Networks*, Vol. 5, Issue 2.

Wright, S.M. and Satchell, S.E. (2002). “Generalised Mean-Variance Analysis and Robust Portfolio Diversification”. *Advances in Portfolio Construction and Implementation*.

Xian G.; Zhang Y. and Liu L. et al (2005). “A novel intelligent system based on WPA-SVMs for fault signals recognition”. *Wireless Communications, Networking and Mobile Computing*. Proceedings. 2005 International Conference

Yang, H., Chan, L. and King, I. (2002). “Support Vector Machine Regression for Volatile Stock Market Prediction”. *Intelligent Data Engineering and Automated Learning* 2002, LNCS 2412, p391-396.

Yao, J. T. and Tan C. L. (2000). “Time Dependent Directional Profit Model for Financial Time Series Forecasting”. *International Joint Conference on Neural Networks 2000*. IEEE Computer Society.

Yoshihara, M. (1990). “Testing the Over-reaction Hypothesis in Japanese Stock Market”. *Journal of Japanese Financial Analyst* 28.

Yu, L., Wang, S. and Lai, K. K. (2005). “Mining Stock Market Tendency Using GA-Based Support Vector Machines”. *Workshop on Internet and Network Economics* 2005.

Yu, S. (1999). “Forecasting and Arbitrage of the Nikkei Stock Index Futures: An Application of Backpropagation Networks”. *Asia-Pacific Financial Markets*, 6.

Zapranis, A. D., Utans, J. and Refenes, A. N. (1996). “Specification Tests for Neural Networks: A Case Study in Tactical Asset Allocation”. *Decision Technologies for Financial Engineering*. p262-275.

Zirilli, J. S. (1997). *Financial Prediction Using Neural Networks*. Thomson Computer Press.

Appendix 1: Review on Momentum and Reversal Indicators

Straumann and Garidi (2007) categorize the equity factors commonly used in financial asset models into different types, some of these are useful in modeling country returns are as follows:

- **Macroeconomic** factors derived from macroeconomic variables, interest rates, FX rates and commodity prices. The problem with macroeconomic variables is that they are mostly reported with a certain time lag so that their use in real time has its limitations.

- **Technical** – constructed using historical price data, for example, momentum indicators

- **Statistical** –factors are extracted from the historical returns and fundamentals factors via statistical techniques such as principal component analysis (PCA).

As we can see there are many factors that one can choose as the inputs to the mathematical models for the quantitative portfolio construction. Given the almost infinitely large numbers of possible combinations of factors, one has to choose a set of factors that makes intuitive sense and support empirically with substantial back-testing.

To test the pair-wise model and Similarity Ratio empirically, we used the momentum and reversal indicators as the factors for our analysis. Before we embark on the search for empirical support to our pair-wise framework, we want to first establish the feasibilities of momentum and reversal indicators as the factors to use in our analysis. We reviewed the publications on these indicators, and followed by some hit rate analysis on the indicators in predicting the direction of relative returns. We concluded by constructing single-indicator models and test the directional accuracy of these models.

Research Works on Momentum and Reversal Indicators

The basic idea of momentum strategy is to buy securities that have performed well and to sell securities that performed poorly. The underlying belief is that the same trend will continue in near future. Typically it is believed that the performance in the past six to twelve months will continue for another three to twelve months. Sufficient evidences, statistically and economically, have been published to support the feasibility and effectiveness of using price momentum in constructing portfolios.

Among many research works on explaining the reason for the success of such indicators, Hong and Stein (1999) provide an intuitive analysis by creating a market model that consists of two groups of rational investors: "news-watchers" and "momentum traders". Each news-watcher observes some private information, but fails to extract other news-watchers' information from prices. If this privately held information diffuses gradually across the population, prices under-react in the short run. The under-reaction means that the momentum traders can profit by trend-chasing. However, if they can only implement simple (i.e., univariate) strategies, their attempts at arbitrage must inevitably lead to overreaction at long horizons and hence a reversal of price to the mean level.

Though the robustness and pervasiveness of the momentum strategies are generally accepted, practitioners and researchers were unable to come to an agreement as in whether this is a result of market inefficiency or can indeed be explained by rational asset pricing. Nonetheless, the abundance of evidences suggests that the success of momentum and reversal indicators is unlikely due to luck or just pure coincidence.

In Hong and Stein (1999), we see price momentum contributing in two ways – short term continuation and long term reversal. That is, assets (e.g. stocks) tend to continue their short-term trend but this trend continuation characteristic reversed for longer term. The short-term trend continuation is some times called *positive momentum* or simply *momentum*. The long-term trend reversal is also called *negative momentum* or *reversal*.

Publications on Momentum Indicators in Equity Markets

Jegadeesh and Titman (1993) were the first to document the momentum phenomenon in academic literature. They noted that over a medium-term (three- to twelve-month) horizon, firms with higher past returns continue to out-perform firms with lower past returns over the same period. In fact, Jegadeesh and Titman paper was so well received that many of the subsequent works followed the definitions and notations used by them: $\boldsymbol{J}$ was used to represent the look-back period and $\boldsymbol{K}$ was used to represent the investment, or holding, period. For example, a momentum portfolio with $\boldsymbol{J} = 3$ and $\boldsymbol{K} = 6$ means that the portfolio is constructed by comparing the latest price against the price 3-period (e.g. 3-month) ago. This portfolio is then held for the next 6 months.

One such paper that adopts the notations is by Griffin et al (2005); they study the global markets using $\boldsymbol{J} = 6$ and $\boldsymbol{K} = 6$ on winner-minus-loser portfolios. In their studies of more than 30 countries, data up to the year 2000 was used and they find that going *long* on those assets that exhibit positive price momentum and *short* those assets that exhibit negative price momentum yield positive returns.

Rouwenhorst (1998) studied twelve stock markets in Europe and found new evidence to support the momentum strategy. He finds that in the European stock markets, momentum strategies work well, either on each individual country or cross-country basis. This was partly confirmed by the study of Rey and Schmid (2007), whom their results on large Swiss stocks shows that price momentum is an effective way to construct single-asset portfolios.

As the leading financial center in the world and availability of large historical data, it is not surprising that US market is extensively studied and analyzed. For example, Chan et al (1996) included data from as early as 1977 to study price momentum strategies in selecting stocks from various US exchanges (NYSE, AMEX and NASDAQ) while Chen and DeBondt (2004) studied the style momentum stocks in S&P 500. All of which merely add on to the long list of empirical evidences that supports price momentum based indicators.

Publications on Reversal Indicators in Equities Markets

While there are abundance of empirical studies on the efficacy of positive momentum in equity market, research works on price reversal studies are less common. We believed this is largely due to the rapid global expansion that resulting in generally upward trending stock markets performances in recent decades. The most commonly sited market that exhibits reversal in price momentum is Japan; some of the empirical results can be found in Liu and Lee (2001), Yoshihara (1990) and Chou et al (2007).

Liu and Lee (2001) find that from the period 1975 to 1997, the price reversals were present in the stocks traded in Tokyo Stock Exchange. Yoshihara (1990) documented that a long term three- (five- to eight-year) return reversal in the Japanese stock markets. In a more general term, DeBondt and Thaler (1985, 1987) reported that long-term past losers (three- to five-year) out-performed long-term past winners over the subsequent three to five years. Chopra et al (1992) suggested that over periods from one to five years extreme stock market performance tends to reverse itself.

Publications on Integrating Both Indicators

Given the amount of evidence that appeared for either momentum or reversal strategies, it is natural to consider the effectiveness of combining the two apparently distinct strategies. Balvers and Wu (2005) explored the implications of an investment strategy that considers momentum and mean-reversion jointly. They applied these momentum-contrarians strategies to select among 18 developed equity markets on a monthly basis. They found that the combined strategies out-performed pure momentum or pure reversal strategies. Others who found some success when combining both mean-reversion and momentum among national equity markets are Cutler et al (1991) and Asness et al (1997).

To combine the two sets of indicators, there are some implementation issues to be answered. Supposed we wanted to construct a portfolio of assets with high returns in the previous 1-12 months period and low returns in the previous 3-5 years period. Should we hold the portfolio for one month or three years? How should we weigh the importance of momentum potential vis-à-vis mean reversion potential? These problems were also highlighted by Balvers and Wu (2005).

Holding Period

The answer to the question of holding period may not be solely dependent on the portfolio construction process but is determined by other factors such as futures contracts expiry dates, corporate rebalancing period, etc. As such, we shall leave this out for the moment.

Balancing the Importance of Momentum Potential vis-à-vis Mean-Reversion Potential

To strike the balance, Balvers and Wu (2005) adopted a parametric approach to build a linear model that comprises of both momentum and mean-reversion components. Regression was used to estimate the coefficients. Van der Hard et al (2002) combined using a scoring approach whereby stocks are ranked using multiple indicators. These indicators were normalized and standardized. The indicator values are then summed up to create a composite score for each stock.

Observations

In general, we found that the *ex-post* analysis of the publications confirmed that momentum and reversal strategies worked during the period of analysis, and in the markets where the analysis took place. Unfortunately many of the *ex-ante* aspects of the strategies were being left out. For example, how should one select the right indicators to include in the model? How the various parameters were set? And lastly, how frequently should the indicators be reviewed. Some other observations are as follows:

- Most publications work on stock selection and equally weighted portfolios. Most followed the standard process as described in Jegadeesh and Titman (1993)

- A fixed set of indicators that work for the selected markets during the testing periods and there are no mentioned on how was the set of indicators being selected.

- Typically short testing periods, especially out-of-sample testing period.

- Most used high frequency data (e.g. daily) and short holding period

- Allow shorting even to the small and illiquid stocks

- Momentum and reversal are used to model absolute stock returns

Empirical Hit Rate of Indicators

List of Tests

- To confirm that momentum and reversal indicators constructed using monthly price series also worked for low frequency data (e.g. monthly) with a long holding period (e.g. quarterly).

- To study if there is a single set of indicators that work for all countries and regional indices.

- To study if the set of indicators remains constant over time, that is, if a set of indicators work for a period, it continues to do so in subsequent periods.

- To test the hypothesis that applying the momentum and reversal indicators on pair of indices (i.e. pair-wise) is better than applying them individually.

Data Set

This study was carried out using monthly index values from the MSCI index series we obtained from FactSet[1]. We used the values of the four regional indices that made up the MSCI World index: North America (NA), Europe (EU), Japan (JP) and Asia Pacific (AP). We used data from 1986 to 1999 for this set of empirical tests.

In this empirical study, we considered the investment universe of the four regional indices that made up MSCI World. This gives us six asset pairs for our analysis. Though we used all the pairs for our study, for the purpose of documentation, we only present the detailed results we obtained from the North America (NA), Japan (JP) and the NA-JP pair in this report. For the results we obtained from the other regions and regional pairs, we only document the summary results.

Evidence of Momentum Signals for North America

As a means of visual investigation, we plotted the three-month holding returns against the positive momentum indicators in scatter plots. The following figure shows two of such plots (R_{t+3} against I_1 and R_{t+3} against I_2). As most of the points appear in the top left-hand quadrant, the charts suggest that relatively high hit rates for both indicators in predicting the direction of three-month holding returns.

---

[1] FactSet is a leading provider of global financial and economic information, including fundamental financial data on tens of thousands of companies worldwide. Combining hundreds of databases into its own dedicated online service, FactSet also provides the tools to download, combine, and manipulate financial data for investment analysis. For more information, see www.factset.com

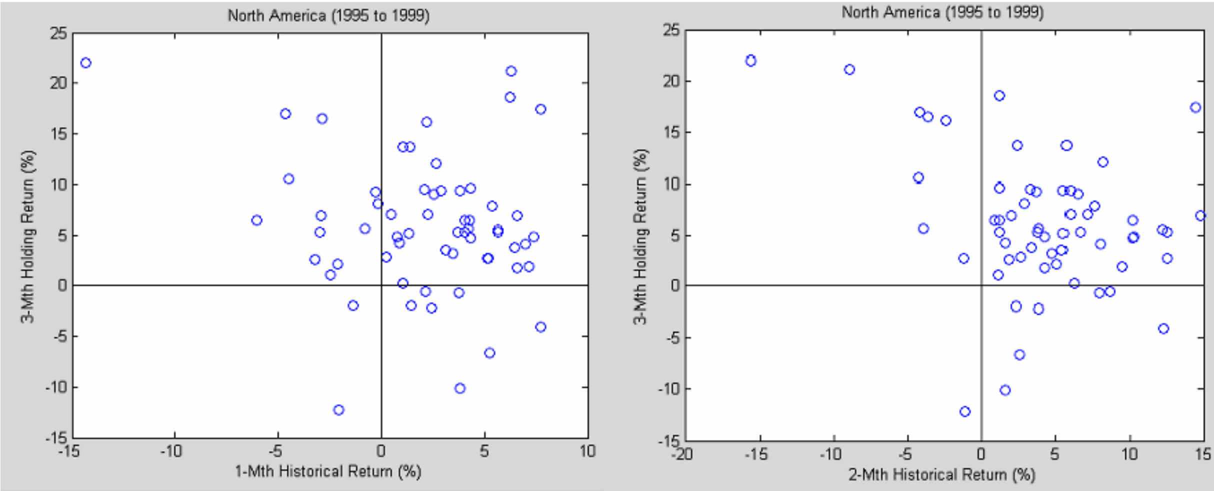

Figure A1-a: North America 1-mth and 2-mth Returns Against 3-mth Holding Returns (1995 – 1999)

The Hit Rate for each momentum indicator is tabulated as follows:

Indicator	Hit Rate (%)
I_1	66.67
I_2	73.33
I_3	70.00
I_6	78.33
I_9	85.00
I_{12}	83.33

Table A1-a: Hit Rate of Positive Momentum Indicators in Predicting Directions of North America 3-mth Holding Returns

We see that regardless of which momentum indicators we used, one would be able to predict the direction for the three-month returns of North America. In fact, during 1995 to 1999, if one were to make quarterly predictions on the direction of the 3-month holding returns based on past 9-month and 12-month returns, one would be correct 80% of the times.

Evidence of Reversal Signals for Japan

We used Japan to illustrate the predictive strength of the reversal indicators (I_{-24}, I_{-36}, I_{-60}). The scatter charts and the table show the hit rates of the reversal indicators:

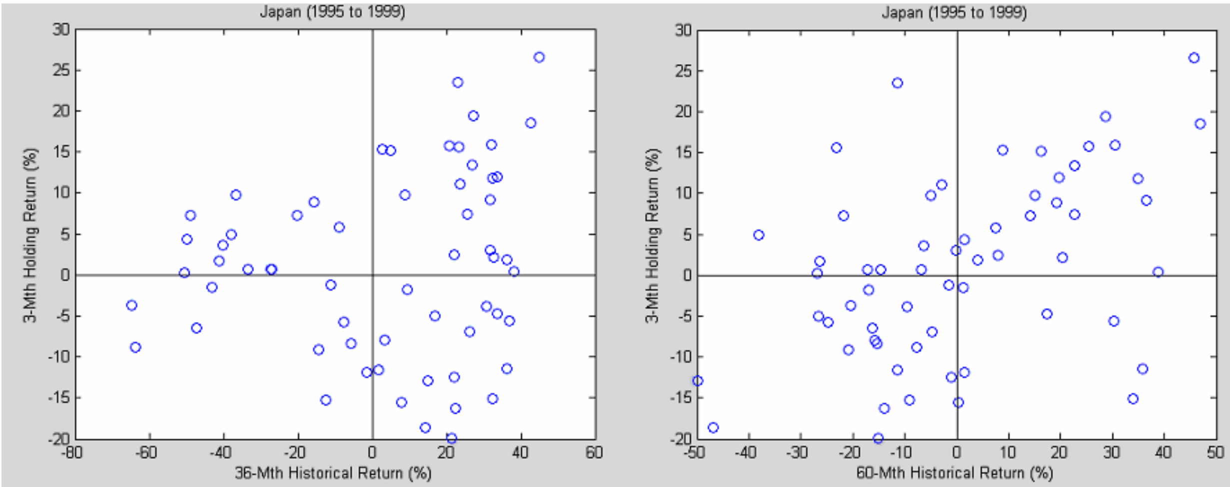

Figure A1-b: Japan 36-mth and 60-mth Returns Against 3-mth Holding Returns (1995 – 1999)

Indicator	Hit Rate (%)
I_{-24}	46.67
I_{-36}	51.67
I_{-60}	66.67

Table A1-b: Hit Rate of Reversal Indicators in Predicting Directions of Japan 3-mth Holding Returns

Clearly the results were less striking as compared to what we had presented on momentum indicators. However, it does show that the reversal indicators do have a certain degree of predictive powers. Many studies, including Yoshihara (1990), had confirmed that I_{-60} is among the best indicators.

Do we need both Momentum and Reversal Signals?

We have seen that North America is a region that one would do well by heeding the signals from momentum indicators. On the other hand, investment opportunities in Japan can be uncovered by reversal indicators. Does this mean that we expect the markets to continue to behave in the same manner month after month, quarter after quarter? Any investor would have to sit through various investment/ economic cycles. For example, Liu and Lee (2001) note that though Japan stock market in general exhibits reversal trend, when they split the their studies into two periods, 1975 – 1989 and 1990 – 1997, they found mixed results. Rey and Schmid (2007) also experience the different momentum strengths in Swiss Market Index from 1994 –1998 and 1998 – 2004. Further evidence was presented when Li (2005) tested the random walk hypothesis by examining the hazard ratio of runs. Li (2005) suggests that there may be two components of stocks returns. Longitudinally, stock return is cyclical and there is a reversal tendency; with the length of the run controlled, there is a positive autocorrelation in stock returns. This finding may explain why over different holding periods had impacts on the profitability of momentum trading and contrarians trading. Classifying a stock market by just a “momentum” or “reversal” may not be sufficient, at least not always the case, we need to answer the question of which component dominates the other.

For example, while North America has been rising significantly over the five years in our study, and hence long term reversal indicators do not work so well in North America, as shown in Table A1-c

Indicator	**Hit Rate (%)**
I_{-24}	15.00
I_{-36}	15.00
I_{-60}	15.00

Table A1-c: Hit Rate of Reversal Indicators in Predicting Directions of North America 3-mth Holding Returns

Conversely, Japans market had been less exciting for a long period in the 90s and this probably explain why even though the momentum indicators work decently well (see Table A1-d), the long-term indicators perform even better.

Indicator	**Hit Rate (%)**
I_1	50.00
I_2	53.33
I_3	60.00
I_6	58.33
I_9	61.67
I_{12}	56.67

Table A1-d: Hit Rate of Positive Momentum Indicators in Predicting Directions of Japan 3-mth Holding Returns

The best indicator for North America is I_9 while the best for Japan is the I_{-60}. Thus to build a framework that invest across different markets, we will need to include both momentum and reversal indicators in our models.

Indicator's Predictive Strength Varies over Time

We have seen from our empirical tests that different indicators work in different markets but does the best indicator always stay the best? Will the predictive power of an indicator change over time? Intuitively speaking, the strength of each indicator would change with changing market conditions.

To test the hypothesis that predictive strength does vary with time, we compute the hit rate for each indicator over a rolling five year period in a region. We then take note of the number of times each indicator came in top (i.e. being the indicator that has the highest hit rate in the region). The exercise was carried out for all the four regions and results tabulated as follows:

Indicator	North America	Europe	Japan	Asia Pacific ex-Japan
I_1	0	0	0	37.7
I_2	0	0	0	0
I_3	0	39.3	0	0
I_6	0	0	0	0
I_9	98.4	41.0	18.0	45.9
I_{12}	1.6	11.5	0	11.5
I_{-24}	0	0	1.6	3.3
I_{-36}	0	8.2	0	0
I_{-60}	0	0	80.4	0.6

Table A1-e: Percentage of Times Each Indicator Recorded Highest Hit Rate in Predicting Regional Returns (1995 to 1999)

While I_9 has been a consistent *winner* in North America, the same cannot be said about other indicators. In particular Europe and Asia Pacific ex-Japan, where there is no one clear *winner* in the regions. This is an indication that the predictive strength of an indicator does change over time and what works on in a period may not always work best in other times. This also suggests that single-indicator forecasting model may not be sufficient.

Now that we see evidence of indicators' predictive power do vary with time, do we know when would be a good time to review these indicators? Would arbitrarily select a review period be sufficient?

Evidence of Indicators' Predictive Power in Relative Returns

The proposed pair-wise model based on the forecasting model to predict the relative returns of each asset pairs. Thus we also studied the predictive powers of the indicators in forecasting the direction of relative returns. The results have not been very impressive as compared to what we have earlier presented on the single-market model. However, we still see that it does have some predictive powers; if the right indicators were used, achieving a hit rate that exceeds 60% is possible:

Indicator	NA-EU	NA-JP	NA-AP	EU-JP	EU-AP	JP-AP
I_1	48.33	58.33	53.33	56.67	56.67	56.67
I_2	51.67	61.67	50.00	66.67	61.67	48.33
I_3	56.67	76.67	56.67	70.00	61.67	46.67
I_6	60.00	78.33	70.00	73.33	65.00	51.67
I_9	53.33	78.33	73.33	71.67	61.67	60.00
I_{12}	56.67	75.00	76.67	68.33	46.67	61.67
$I\text{-}_{24}$	51.67	35.00	30.00	45.00	41.67	45.00
$I\text{-}_{36}$	53.33	41.67	30.00	50.00	36.67	58.33
$I\text{-}_{60}$	43.33	30.00	23.33	40.00	26.67	50.00

Table A1-f: Hit Rate of Indicators in Prediction Regional Pair-wise Relative Returns (1995 to 1999)

We repeat the above exercise, moving back one month at a time, to obtain the hit rate for each indicator in predicting the sign of the relative returns over a rolling five-year period. This exercise was repeated for a five-year period, generating 60 samples. For each of the sample (five-year period), we identify the indicator that has recorded the highest hit rate for each of the six regional pairs. Table A1-g shows the number of months, as a percentage of 60 months, that each indicator was identified as having the highest hit rate. Like the case in single-market model, we see that not one indicator came in top for the whole period:

Indicator	NA-EU	NA-JP	NA-AP	EU-JP	EU-AP	JP-AP
I_1	0.00	0.00	0.00	0.00	0.00	0.00
I_2	0.00	0.00	0.00	0.00	0.00	0.00
I_3	10.00	13.33	0.00	38.33	6.67	0.00
I_6	6.67	11.67	26.67	5.00	30.00	0.00
I_9	0.00	0.00	45.00	0.00	3.33	0.00
I_{12}	0.00	75.00	5.00	25.00	0.00	45.00
$I\text{-}_{24}$	0.00	0.00	21.67	31.67	60.00	0.00
$I\text{-}_{36}$	83.33	0.00	0.00	0.00	0.00	55.00
$I\text{-}_{60}$	0.00	0.00	1.67	0.00	0.00	0.00

Table A1-g: Percentage of Times Each Indicator Recorded Highest Hit Rate in Predicting Regional Pair-wise Relative Returns (1995 to 1999)

Take the EU-AP pair for example; we see from Table A1-f that $I\text{-}_{24}$ was among the worst indicators for the sample period from 1995 to 1999, having a hit rate of only 41.67%. However, as we start to roll the sample window backwards at a monthly interval, we see from Table A1-g that it was the best indicator 60% of the times in predicting the direction of Europe's returns relative to Asia Pacific ex-Japan. In contrast, we see from Table A1-f that I_6 was the best indicator from 1995 to 1999 with a hit rate of 65%, but on a rolling five-year basis, it only came in top 30% of the times. The predictive powers to these two indicators clearly have changed during the sample period.

The historical hit rates of each indicator in predicting the direction of the relative returns of the EU/ AP pair are plotted in Figure 2-a. We see that while the indicators generally stay relevant and effective for a certain period, the length of the periods vary from indicator to indicator. For example, we see that the predictive power of $I\text{-}_{24}$ diminishes over time while I_6 gets better in later part of the sample period. This suggests that having a fixed period to review the efficacy of the indicators may not be able to capture failing indicator, or include the more effective indicator in the forecasting model. This study supports the need to have a contextual model where the most appropriate indicators are used at the time when investment decision is to be made.

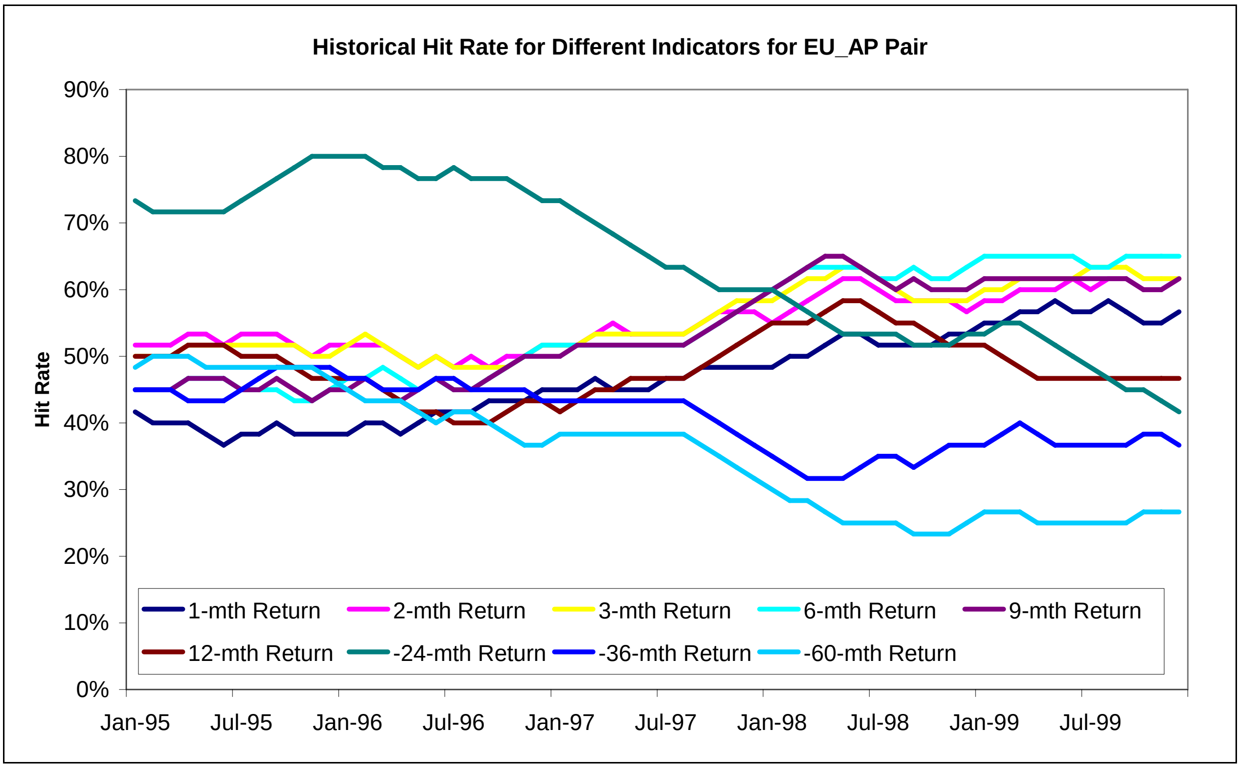

Figure A1-c: Historical Hit Rates for Various Indicators for EU_AP Pairs

Forecasting Model for Relative Returns

To forecast the relative performance, we can do it in two ways:

(1) Forecast the returns for both assets, $r_{asset1,t}$ and $r_{asset2,t}$,separately and use the difference as the forecast

(2) Forecast the relative returns, $r_{asset1-asset2,t}$, directly

To compare if modeling relative returns is indeed better than modeling the asset returns individually, we carried out some preliminary empirical tests. We compare the hit rates we obtained from using both approaches in predicting the relative performance of six regional pairs in MSCI World. We used one-variable regression forecasting models in both approaches.

Constructing a Forecasting Model

We defined pair-wise relative returns for an asset pair, say asset 1 and asset2, as the difference of their assets returns, which is consistent with Qian (2003):

$$r_{asset1-asset2,t} = r_{asset1,t} - r_{asset2,t}$$

By this definition, the indicator for an asset pair is the difference of the indicators of the two assets:

$$I_m(asset1, asset2) = I_m(asset1) - I_m(asset2)$$

To test approach one, we will choose the single best indicator for each market:

I_9: North America (NA), Europe (EU), Asia Pacific ex-Japan (AP)

I_{-60}: Japan (JP)

We next identified the best indicator for each of the asset pairs and use it to construct the linear model for each of the market. In our implementation, we have chosen hit rate as the evaluation criteria to select the indicators to use.

The indicators chosen for the models are as follows:

I_3: EU-JP

I_9: NA-JP, NA-AP, EU-AP

I_{-24}: JP-AP

I_{-36}: NA-EU

Validating the Regression Models

Among the models constructed for the six pairs, all have regression coefficients that are significantly different from zero except for the JP-AP pair. The residuals from the regression models were also tested for normality and independence with the following tests (all hypothesis tests were conducted at 95% confidence level):

- Jarque-Bera test –normality
- t-test – zero mean
- Durbin Watson test – independence

We see that the residuals from the JP-AP pair had violated our validation tests so it suggests that the results obtained from the model may not be stable. We still include its results for completeness sake.

	NA-EU	NA-JP	NA-AP	EU-JP	EU-AP	JP-AP
t-stats	2.7939	2.5512	3.3785	3.9469	3.9946	-0.8481
p-value	0.0070	0.0134	0.0013	0.0002	0.0002	0.3999
Residual Normal	Yes	Yes	Yes	Yes	Yes	No
Residual zero mean	Yes	Yes	Yes	Yes	Yes	No
Residual independence	Yes	Yes	Yes	Yes	Yes	No

Table A1-h: Test Statistics for Regional Pair-wise Regression Models

Out-of-sample Analysis

The forecasting models were then used to forecast a 3-month ahead relative returns and the hit rates were recorded:

Model	NA-EU	NA-JP	NA-AP	EU-JP	EU-AP	JP-AP
Individual	49.18	65.57	52.46	49.18	44.26	49.18
Pair-wise	57.37	62.30	52.46	67.21	59.02	55.74

Table A1-i: Comparison of Hit Rate in Predicting Directions of Regional Pair-wise Relative Returns for Individual Model and Pair-wise Model

We see that pair-wise models generally performed better as compared to individual forecasts. The individual models did not even deliver 50% hit rate in four out of the six pairs. In the case of NA-EU pair where the individual model out-performed, the pair-wise model still managed to score a hit rate exceeding 60% so pair-wise model did not fare too badly either.

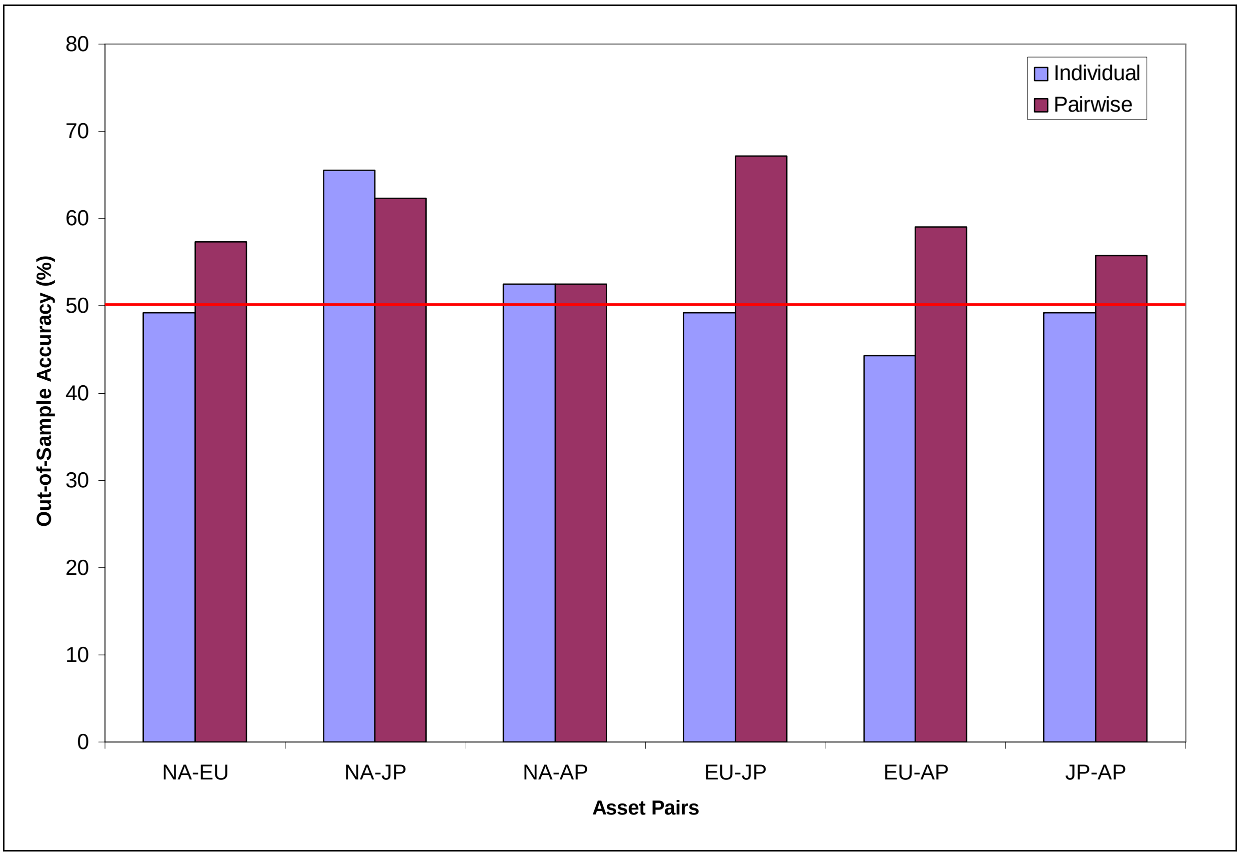

Figure A1-d: Comparison of Hit Rate in Predicting Directions of Regional Pair-wise Relative Returns for Individual Model and Pair-wise Model

Findings from Empirical Studies

From the series of empirical tests we had conducted, our initial conclusions are:

- Momentum and reversal indicators also worked for low frequency data (e.g. monthly) with a long holding period (e.g. quarterly).

- There is no single set of indicators that work for all countries and regional indices.
- The predictive powers of the indicators vary over time.
- Pair-wise forecasting models performed better in predicting direction of relative returns than those single-asset models.

The two most important results that our empirical studies have revealed are:

- Indicators' predictive power vary over time
- Pair-wise model translate to better signals in forecasting relative returns

These results provide the sufficient evidences to support a contextual pair-wise model. In addition, it also supports the use of momentum and reversal indicators in the implementation of the generic pair-wise framework.

Appendix 2: Asset Allocation Portfolio Management

Investment Goal and Three Parameters

In portfolio management, the first step is to identify the investment goal. It is this goal that determines how the portfolio should be managed. Alford et al (2003) point out that the investment goal is specified by three parameters:

1. Portfolio's benchmark
2. Risk-return target
3. Investment Constraints

Portfolio's Benchmark

Siegel (2003) defines *benchmarks* as "*paper portfolios constructed for comparison with real portfolios to see whether the latter are being managed effectively*". The benchmark represents the client's neutral position and the starting point for any active portfolio. It embodies the opportunity set of investments. For example, a portfolio that contains US stocks with large market capitalization can be benchmarked against S&P500 or MSCI US.

Risk-return Target

The fact that the existence of a benchmark is to see if an active portfolio is being managed effectively leads us to the second parameter of the investment goal – risk-return target. Indeed, not only does the benchmark portfolio serve as a reference portfolio to the active managers, it also serves as a performance target for the portfolios. When an active portfolio is benchmarked, then it must aim to deliver a higher return than the benchmark portfolio (or the *passive* portfolio). This excess return is often termed as *active return* or *value-added*.

To prevent portfolio managers from deviating too much from the benchmark portfolio in search for better investment returns, the portfolio's risk is usually measured relative to the benchmark, usually in the form of *active risk* or *tracking error*. (Ex-post) tracking error is calculated as the standard deviation of a portfolio's value-added. To compare how well different active portfolios are being managed, one can look at the value-added generated for each unit of active risk taken. This quantity is called the *Information Ratio* and it offers a convenient way to compare the performances of active portfolios with different tracking error. The risk and return targets, or a target Information Ratio helped to ensure the portfolio manager to stay within the boundaries implied by the given benchmark and avoid unfair attribution of performance (or blame) caused by the asset class policy returns.

Investment Constraints

Several studies have also shown that adding constraints to the optimizer leads to better out-of-sample performance, for example, Frost and Savarino (1988). Some considerations that went into the decision on setting the limits are market capitalization, mandates' restriction, regulatory restriction, liquidity, etc. This also helps to avoid large concentration in one particular holding and thus helps to maintain a well-diversified portfolio.

Holdings constraints can also come in the form of active weights constraints. That is, instead of having a bound for the holdings, we fixed the permissible range for active positions. This form is useful if a portfolio is to be measured against a benchmark and that all measures (e.g. risk, performance, positions) are considered relative to the benchmark.

Instruments Used to Implement Asset Allocation Portfolio

Given an optimal country allocation mix, there are a few ways an investment manager can implement the portfolio:

1. Stocks
2. Index derivatives e.g. futures and forward contracts
3. Country Index funds e.g. Exchange Traded Funds[1] (ETF)

Construction using stocks involves high transaction costs and requires active management of the stocks in the portfolio, which will have an impact on the value-added. To mitigate risks associated with active stock selection, one can choose to select the stocks in proportion to the stocks' benchmark weights in each country. However, this requires close monitoring and high operational setup to reduce slippages in tracking. The more cost effective and less operational effort ways are to use index derivatives or ETFs.

Traditional and Quantitative Approach to Investment

One major difference between the two groups of investment managers are the way the active positions are being determined. Quantitative managers often used an optimizer to find the optimal holdings that meet the risk-return target while staying within the investment constraints. This is rarely the case for traditional managers as the active positions are usually obtained via qualitative assessment and consciously stay within the boundaries caused by the investment constraints. Because the quantitative managers often used the optimizer, there is a need for them to generate a forecast of the expected returns for each asset that to be fed into the optimizer. This is often pointed out to be the weakness of a quantitative approach as rarely the case one can have a view of each and every asset. This is also the main reason why optimizer is seldom used by the traditional managers. Not only that they may not have view on every security, they have found it difficult to quantify their analysis of a company or stock into a single number called *expected return*.

[1] Exchange-traded funds are exchange-listed, equity securities backed by a basket of stocks from which they derive their value. Unlike closed-end funds, the basket of securities can be expanded as demand for the product increases. ETFs are designed to track country, sector, industry, style and fixed income indexes.

The pair-wise framework that we had proposed is essentially a quantitative approach. Chincarini and Kim (2006) list numerous advantages of quantitative equity portfolios over the traditional qualitative oriented portfolios. The two most important advantages offered by a quantitative approach are objectivity and better risk control. In a quantitative approach, the outputs are driven by the mathematical models and this significantly lessens the impact of the manager's biases on the portfolio. Risk control can also be implemented in the portfolio construction process and this helps the portfolio avoid large swings and keeps volatility low, giving rise to a more stable stream of portfolio performances over a long horizon

Appendix 3: Results of Portfolios Constructed Based on Perfect Forecast

To further illustrate the feasibility of pair-wise modeling, we construct portfolios, assuming perfect foresight. We use perfect forecasts as the views to obtain expected returns vector. In the pair-wise case, we set up the view matrices using the actual relative returns as the "perfect relative views". In the case of Individual model, we use the assets' absolute returns as the views (i.e. "perfect absolute views") into the Black-Litterman model. We then run the optimizer using the same objective function and same set of constraints to obtain the holdings weights. For the pair-wise model, the scores to the pairs were assigned randomly.

Three sets of perfect forecasts were used:

1. Perfect Forecasts – actual returns are used
2. Perfect Direction Forecasts – assumed the forecasts correctly predicted the direction, a magnitude of 0.5 was used
3. Perfect Magnitude Forecasts – assumed the forecast correctly predicted the magnitude, with the signs of the forecasts chosen randomly

Performances of Portfolios Constructed Using Perfect Forecast

The performances of the two global portfolios are tabulated below:

Performance Indicators	Individual Model	Pair-wise Model
PI1 – Asset Allocation Value-added (%)	5.11	**7.75**
PI2 – Information Ratio	2.43	**4.83**
PI3 – Proportion of Out-performing Quarters (%)	89.29	**100.00**
PI4 –Turnover (%)	**33.15**	49.54
PI5 – Correlation with Market	-0.34	**0.01**

Performance Indicators	Individual Model	Pair-wise Model
PI7 – Trading Edge (bp)	126	**189**

Table A3-a: Performances of Global Portfolios Constructed with Perfect Forecasts

The results clearly indicate that having perfect forecast for relative returns yield better results than having perfect forecast of assets' absolute returns. This supports our claim to model asset pairs' relative returns.

Comparison of Portfolios Constructed Based on Perfect Direction Forecast

We have also state that having the right direction is critical to the success of country allocation. To test the importance of direction in the pair-wise framework, we constructed test portfolios using the sign of the actual returns, and set a magnitude of the forecasts to the average of those historical relative returns that are of the same sign as our "prediction":

Performance Indicators	Individual Model	Pair-wise Model
PI1 – Asset Allocation Value-added (%)	4.53	**7.23**
PI2 – Information Ratio	2.25	**4.41**
PI3 – Proportion of Out-performing Quarters (%)	85.71	**100.00**
PI4 –Turnover (%)	**36.03**	49.16
PI5 – Correlation with Market	-0.36	**0.06**
PI7 – Trading Edge (bp)	112	**176**

Table A3-b: Performances of Global Portfolios Constructed with Perfect Direction Forecasts

The results again support our claim of pair-wise modeling, in addition, we find that the performances of the portfolios constructed based on perfect direction forecasts are not too different to those that were constructed using perfect forecasts. This confirms the importance of direction forecasts in the pair-wise framework.

Comparison of Portfolios Constructed Based on Perfect Magnitude Forecast

To complete the study, we also construct portfolios based on perfect magnitude forecasts, that is, we use the magnitude of the actual returns as inputs. The signs of the inputs are selected randomly. Not surprisingly, the pair-wise model once again out-performed the individual model:

Performance Indicators	Individual Model	Pair-wise Model
PI1 – Asset Allocation Value-added (%)	0.50	**0.95**
PI2 – Information Ratio	0.33	**0.58**
PI3 – Proportion of Out-performing Quarters (%)	57.14	**61.71**
PI4 –Turnover (%)	51.18	**50.59**
PI5 – Correlation with Market	0.27	**-0.16**
PI7 – Trading Edge (bp)	13	**24**

Table A3-c: Performances of Global Portfolios Constructed with Perfect Magnitude Forecasts

The performances of the two portfolios are generally very poor as compared to those constructed using perfect direction forecasts. This provides more evidence that direction forecast accuracy is the most important aspect in the forecasting models within the pair-wise framework.

Appendix 4: MATLAB Code Segments

Fitting the Regression Model

```
function [WinnerX, WinnerScore, WinnerModel] = FitRegressionModel(X, Y,
WeightFunction, ...
    InSampleIndex, SemiOutOfSampleSize, SampleSize, InitialInSampleSize, ...
    ScoreMethod, StepwiseOn)

    k = size(X, 2);
    if StepwiseOn ~= 1
        N = 1;
        C = ones(1, k);
    else
        N = 2^k-1;
        C = zeros(N, k);
        for m = 1:N
            C(m, :) = bitget(uint8(m), k:-1:1);
        end
    end

    StepScore = zeros(N, 1);
    for OneStep = 1:N

        StepX = X(:, strmatch(1, C(OneStep, :)'));

        % Semi-out-of-sample Fit Model
        % ----------------------------
        S = 0;
```

```
    a1 = zeros(SemiOutOfSampleSize, 1);
    a2 = zeros(SemiOutOfSampleSize, 1);
    NumberOfForecast = 0;
    for i = 1:SemiOutOfSampleSize
        InSampleX = StepX(InSampleIndex(SampleSize, i), :);
        InSampleY = Y(InSampleIndex(SampleSize, i), :);
        [b, s] = robustfit(InSampleX, InSampleY, WeightFunction);
        X_ForSemiOutSampleForecast = StepX(InitialInSampleSize + i, :);
        F = [1 X_ForSemiOutSampleForecast] * b;
        a1(i) = F;
        a2(i) = Y(InitialInSampleSize + i);
        if abs(F) > 0.01
            NumberOfForecast = NumberOfForecast + 1;
            if sign(F) == sign(Y(InitialInSampleSize + i))
                S = S + 1;
            end
        end
    end

    StepScore(OneStep) = ComputeConfidenceScore (a1, a2, ScoreMethod, 999);
end

% Find the winner
% ---------------
[Sorted, SortedIndex] = sort(StepScore, 'descend');
if numel(SortedIndex)>0
    WinnerCombination = C(SortedIndex(1), :);
    WinnerModel = strmatch(1, WinnerCombination');
    WinnerX = X(:, WinnerModel);
```

```
        WinnerScore = StepScore(SortedIndex(1));
    else
        WinnerModel = 1:k;
        WinnerX = X;
        WinnerScore = 0;
    end
end
```

Compute Scoring Measures

```
function [Score] = ComputeConfidenceScore (F, A, ScoreMethod, IgnoredValue)

      SampleToUse = strmatch(1, (F ~= 999) .* (F ~= IgnoredValue));
      a1 = F(SampleToUse);
      a2 = A(SampleToUse);
      s1 = sqrt(sum(a1 .^ 2)/ (size(a1, 1) - 1));
      s2 = sqrt(sum(a2 .^ 2)/ (size(a2, 1) - 1));

      switch (ScoreMethod)
         case {'HitRate'}
            Score = sum(max(0, sign(a1 .* a2))) ./ size(a1, 1);

         case {'AnomalyCoeff'}
            Score = ((((a1' * a2)./(s1 * s2))./(size(a1, 1) - 1)) + 1)./ 2;

         case {'UncenteredInfoCoeff'}
            s1 = sqrt(sum((a1 - mean(a1)) .^ 2)/ (size(a1, 1) - 1));
            s2 = sqrt(sum((a2 - mean(a2)) .^ 2)/ (size(a2, 1) - 1));
            Score = ((((a1' * a2)./(s1 * s2))./(size(a1, 1) - 1)) + 1)./ 2;
```

```
        case {'SimilarityRatio'}
            SR_Index = ((a2.^2 + a2.*a1) <= 0);
            SR_Index = strmatch(0, SR_Index);
            SR = zeros(size(a1, 1), 1);
            SR(SR_Index) = abs(a1(SR_Index) + a2(SR_Index)) ./ ...
                (abs(a1(SR_Index) + a2(SR_Index)) + abs(a1(SR_Index) - _
                a2(SR_Index)));
            Score = mean(SR);

        case {'TheilForecastAccuracy'}
            % Theil's Forecast Accuracy Coeff
            Score = 1 - (sqrt(sum((a1 - a2).^2)) ./ (sqrt(sum(a1 .^ 2)) _
                + sqrt(sum(a2 .^ 2))));

        case {'InfoCoeff'}
            Score = (corr(a1, a2) + 1) ./ 2;
    end

end
```

Views Selection

```
function [Views] = SelectViews (PairMatrix, PairScore, N, ...
    MinConfidenceLevel)

    Views = [];
    AssetSelectionMatrix = zeros(N, 1);
    NumberOfPairs = size(PairMatrix, 1);
    AllAssetSelected = (min(AssetSelectionMatrix) ~= 0);
    i = 1;
```

```
while (AllAssetSelected == 0) && (i <= NumberOfPairs)

    % Find the max score
    % ------------------
    [max_score, max_index] = max(PairScore);

    % If the max score is less than 50% then no need to proceed
    if max_score < MinConfidenceLevel
        break
    end

    Views(i) = max_index;
    Asset1 = PairMatrix(max_index(1), 1);
    Asset2 = PairMatrix(max_index(1), 2);
    AssetSelectionMatrix(Asset1) = 1;
    AssetSelectionMatrix(Asset2) = 1;



    % Clear all scores that involves two selected assets
    % --------------------------------------------------
    for j = 1:NumberOfPairs
        if (AssetSelectionMatrix(PairMatrix(j, 1)) && _
              AssetSelectionMatrix(PairMatrix(j, 2)))
            PairScore(j) = 0;
        else
            if (AssetSelectionMatrix(PairMatrix(j, 2)) && _
                AssetSelectionMatrix(PairMatrix(j, 1)))
                PairScore(j) = 0;
            end
```

```
        end
    end

    AllAssetSelected = (min(AssetSelectionMatrix) ~= 0);
    i = i + 1;
end

if isempty(Views) ~= 1
    Views = Views';
else
    Views = [];
End
```

Appendix 5: Robust Regression Techniques

Robust regression techniques available are *least trimmed squares* (LTS) regression, *least median squares* (LMS) regression, *least absolute deviations* (L1) regression, and *M-estimates* of regression.

Least Trimmed Sqaures (LTS) Regression

Rousseeuw (1984) introduced LTS regression for fitting linear regression models. The LTS is equivalent to ordering the residuals resulting from a least squares fit, trimming the observations that correspond to the largest residuals, and then computing a least squares regression model for the remaining observations. However, the use of an ordinary least squares estimator lacks robustness because a single observation can cause the parameter estimates to take on any value.

Least Median Squares (LMS) Regression

Unlike LTS, LMS minimizes the median of the squared residuals instead of the sum of the squared residuals. LMS regression has a high breakdown point of almost 50%. That is, almost half of the data can be corrupted in an arbitrary fashion, and the estimates obtained by LMS continue to model the majority of the data well. One problem of LMS is that it is statistically inefficient.

Least Absolute Deviation (L_1) Regression

This method gives less weight to larger errors than the least squares method. It finds the coefficients estimates $\hat{\beta}_{L1}$ that minimize the sum of the absolute values of the residuals:

$$\sum_{i=1}^{n} \left| r_i \beta \right|$$

There exist also L_p regression methods that minimize

$$\sum_{i=1}^{n} \left| r_i \beta \right|^p$$

Forsythe (1972) suggested, on the basis of a 400-replicates Monte Carlo study, that a value of p = 1.5 would be a good general choice. It led to estimates that were no worse than 95% as efficient[1] as least squares when the errors were actually normal.

M-Estimates of Regression

The M-estimator of regression was first introduced by Huber in 1973. For a given ρ function, an M-estimate of regression $\hat{\beta}_M$ minimizes the objective function:

$$\sum_{i=1}^{n} \rho\left(\frac{r_i \beta}{\sigma}\right)$$

Least squares regression corresponds to $\rho(x) = x^2$ and L_1 regression corresponds to $\rho(x) = |x|$. Generally, the value of $\hat{\beta}_M$ is dependent on the value of σ, which is usually unknown.

Although M-estimates are protected against wild values in the response variable, they are sensitive to high leverage points, which have very different x values compared to the other data points in a model. In particular, a typographical error in an explanatory variable can have a dramatic affect on an M-estimate, while least trimmed squares handles this situation easily. One advantage of M-estimates is that they can be computed in much less time than LTS or other high-breakdown point estimators.

The weights are fitted using an iterative procedure. In the fitting algorithm, an initial model is calculated using traditional weighted least squares by default. The algorithm computes a new set of weights based on

[1] Efficiency was defined as the ratio (mean square error for least squares)/ (mean square error for power p)

the results of the initial fit. The new weights are then used in another weighted least squares fit and so on until some convergence criteria are satisfied or a specified maximum number of iterations is reached.

Appendix 6: Country Weights of Global Model Portfolio