# Analysis of Pair Trading Strategy using Kalman Filter

*Abstract*— **In this paper we devise a pair trading strategy using Kalman filter and test it on Indian stocks. Pair trading is a market neutral strategy consisting of a long position in one security and a beta (hedge ratio) adjusted short position in another that bets on statistical relation between the two. This paper outlines how Kalman filter is applied to the spread for smoothening out any random noise and ensuring correct calculation of beta. We shortlist pairs for further analysis using statistical tests, define stop-loss methodology and backtest the strategy on shortlisted pairs. The performance of strategy is then analyzed at the portfolio level and for individual pairs based on a few parameters. The paper concludes with summarizing the strategy's performance over different backtesting horizons and suggests additional techniques to further optimize the strategy for superior performance.**

## I. History

The first instance of statistical pairs trading can be attributed to Nunzio Tartaglia, a Morgan Stanley quant in 1980. At that time, a group of statisticians, mathematicians, and computer scientists were assembled by him to develop state of the art statistical arbitrage trading strategies. One of the techniques used by them involved trading in pairs whose prices tend to move together. Whenever a deviation in this relation was observed, the pair would be traded based on the assumption that this anomaly would correct itself. This idea came to be known as pairs trading and has since increased in popularity for hedge fund and proprietary trading firms.

## II. Introduction

Pairs trading strategy bets on cointegration between the stock pairs, when this cointegration temporarily weakens both stocks of the pair move in opposite direction resulting in widening of the spread. This strategy would then go long on the underperforming stock and short the over performing stock betting that spread between the two would converge back to its original level (betting on mean reversion of the spread). The spread between the two stocks is calculated as X-β*Y where β is the beta coefficient (beta of one security with respect to another) X is the dependent variable and Y is the independent variable. We use beta adjusted spread so that the strategy has almost zero cost during initiation. Pair trading strategy is called an arbitrage strategy, however there is risk involved in it. The strategy bets that statistical relation between the two stocks in the pair will hold good and if there is any breakdown in this relation it will lead to underperformance of the strategy. To understand how we trade such a strategy, consider the figure 1. We plot the beta adjusted spread against the long-run mean of the spread and rolling positive/negative one standard deviation band of the beta adjusted spread. Whenever the spread crosses above the positive one standard deviation band we go short on the spread i.e. we short X and go long beta times Y till the spread converges to its long-run mean and when the spread crosses below negative one standard deviation of the spread we go long on spread i.e. we short beta times Y and go long X till the spread converges to its long-run mean.

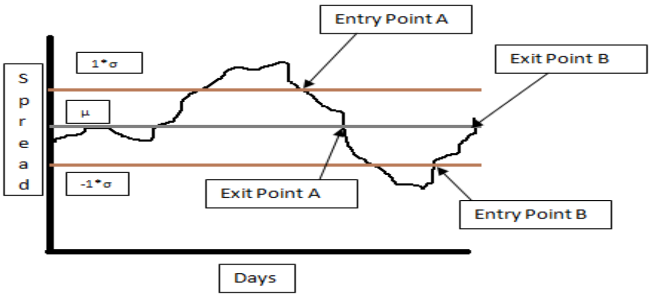


Fig.1 Example of a basic pair trading strategy

## III. Motivation

In any mean reverting strategy such as pairs trading we just analyze historical prices to find out hedge ratios (betas) for corresponding stocks and generate the spread. This has a disadvantage that if the look-back period is short or if it has spurious movements at the beginning or at the end of time series of prices it can have an artificial impact on the hedge ratios. Using a moving average or an exponential weighing scheme is not optimal as well. Hence, we decide to tackle this problem by using Kalman filter to update the hedge ratios which avoids the problem of choosing a weighing scheme arbitrarily.

Kalman filter is an optimal linear algorithm that updates the expected value of a hidden variable based on the latest value of an observable variable. It can be thought of as a technique to smooth out random noise from the observed spread in case of pairs trading.

Kalman filtering process is a 3 Step Process, the steps involved are:

*Prediction:* In this step we predict the next state of the system based on the knowledge of current state and estimate error in our prediction

*Observation:* In this state we take an actual reading from the predicted state after the prediction time has elapsed. Similar to prediction state we also estimate the error with our observation. The observation along with an estimate of the error constitutes the observation step.

*Correction:* we now have the estimates from prediction and observation states, the correction step would involve reconciliation of the above two states estimated taking into consideration the magnitude of the errors i.e. the predicted estimate is corrected based on the observation. This reconciled estimate from the correction state is the final estimate of the current state system. The corrected state is calculated based on the formula below:

Corrected State= Predicted State + K*(Actual observation – Predicted observation)
Where K -> Kalman Gain

The above process is repeated for the state at the next time instance, making Kalman filtering an iterative prediction-correction method. The flow chart below outlines the Kalman filter algorithm used in this pair trading strategy

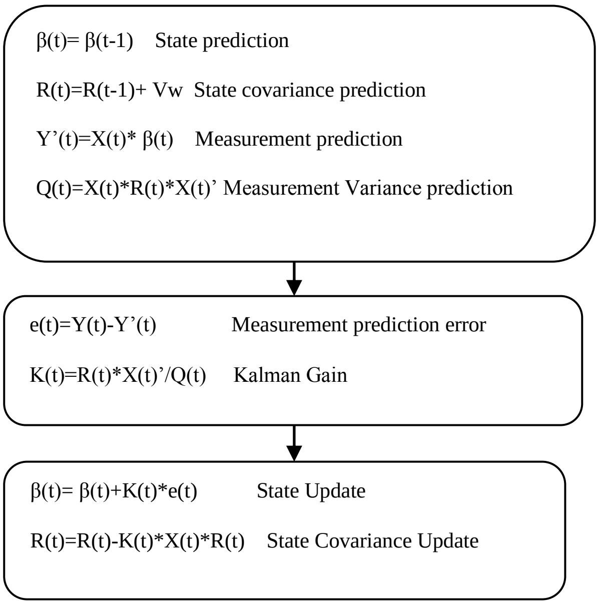

β(t)= β(t-1) State prediction
R(t)=R(t-1)+ Vw State covariance prediction
Y'(t)=X(t)* β(t) Measurement prediction
Q(t)=X(t)*R(t)*X(t)' Measurement Variance prediction
e(t)=Y(t)-Y'(t) Measurement prediction error
K(t)=R(t)*X(t)'/Q(t) Kalman Gain
β(t)= β(t)+K(t)*e(t) State Update
R(t)=R(t)-K(t)*X(t)*R(t) State Covariance Update

Fig.2 kalman Filter Flow Chart

## IV. Selection of Pairs and Backtesting Data

We now go ahead with applying the algorithm described above on universe of stocks traded on the Nifty Index, which is the benchmark stock market index for Indian equity markets. In our analysis, we choose five major sectors from Nifty Index for selection of pairs namely IT (information technology), private sectors banks, FMCG (fast moving consumer goods), energy and pharmaceutical. Within each sector, we only include stocks with large market capitalization.

Next step involves choosing stock pairs with each sector for carrying out further tests. We hence employ a short listing criteria based on the intra-sector correlation. This would involve calculating the correlation matrix for all the major stocks within each sector and then implementing a rule to reduce the universe of stock pairs based on the computed correlation matrix. The rule is to select a maximum of four pairs having a correlation of at least 75% within each sector. We use maximum pairs per sector constraint so that one particular sector doesn't dominate our portfolio as stocks within certain sectors tend to have a high correlation with each other.

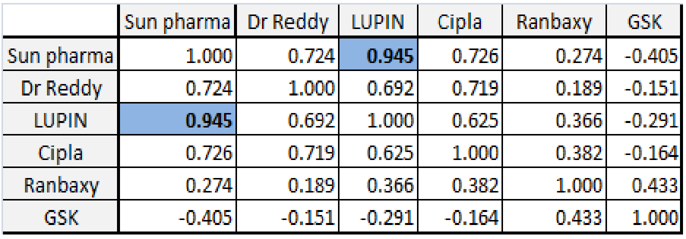
For the purpose of short listing pairs and backtesting the strategy we will be using historical data for all stocks from period 1-Jan-2011 till 18-Feb-2015. This data would be divided into three subsets as described below:

*Subset 1 (1-Jan-2011 to 1-Jan-2013):* Required for calculation of correlation matrix to shortlist stock pairs, testing for cointegration, calculation of initial beta and signal to noise ratio

*Subset 2 (2-Jan-2013 to 10-Mar-2013):* Required for calibration of parameters such updated betas, mean and standard deviation of beta-adjusted Spread in the pair trading algorithm

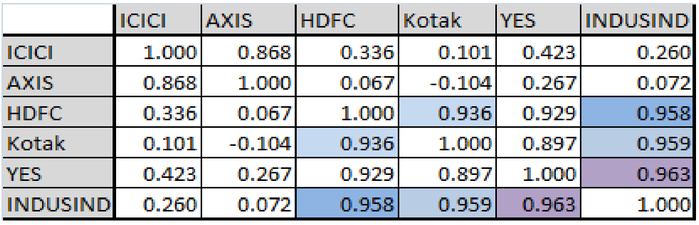
*Subset 3 (11-Mar-2013 to 18-Feb-2015):* Required for backtesting and performance analysis of the pair trading Strategy

We now calculate the correlation matrix for all stocks and shortlist the stock pairs for further tests. The figure below shows sector wise correlation matrix and shortlisted stock pairs



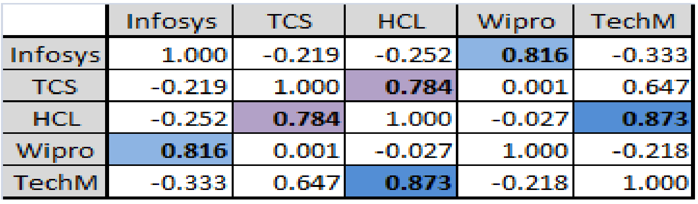
| | Sun pharma | Dr Reddy | LUPIN | Cipla | Ranbaxy | GSK |
|---|---|---|---|---|---|---|
| Sun pharma | 1.000 | 0.724 | **0.945** | 0.726 | 0.274 | -0.405 |
| Dr Reddy | 0.724 | 1.000 | 0.692 | 0.719 | 0.189 | -0.151 |
| LUPIN | **0.945** | 0.692 | 1.000 | 0.625 | 0.366 | -0.291 |
| Cipla | 0.726 | 0.719 | 0.625 | 1.000 | 0.382 | -0.164 |
| Ranbaxy | 0.274 | 0.189 | 0.366 | 0.382 | 1.000 | 0.433 |
| GSK | -0.405 | -0.151 | -0.291 | -0.164 | 0.433 | 1.000 |

Table 1 Correlation matrix pharmaceutical sector



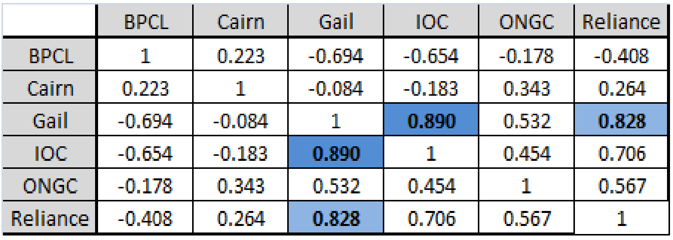
| | ICICI | AXIS | HDFC | Kotak | YES | INDUSIND |
|---|---|---|---|---|---|---|
| ICICI | 1.000 | 0.868 | 0.336 | 0.101 | 0.423 | 0.260 |
| AXIS | 0.868 | 1.000 | 0.067 | -0.104 | 0.267 | 0.072 |
| HDFC | 0.336 | 0.067 | 1.000 | 0.936 | 0.929 | 0.958 |
| Kotak | 0.101 | -0.104 | 0.936 | 1.000 | 0.897 | 0.959 |
| YES | 0.423 | 0.267 | 0.929 | 0.897 | 1.000 | 0.963 |
| INDUSIND | 0.260 | 0.072 | 0.958 | 0.959 | 0.963 | 1.000 |

Table 2 Correlation matrix private sector Banks



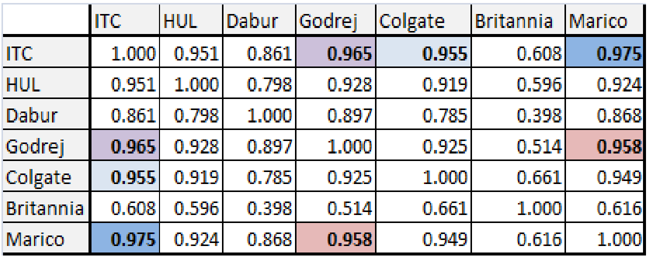
| | Infosys | TCS | HCL | Wipro | TechM |
|---|---|---|---|---|---|
| Infosys | 1.000 | -0.219 | -0.252 | **0.816** | -0.333 |
| TCS | -0.219 | 1.000 | **0.784** | 0.001 | 0.647 |
| HCL | -0.252 | **0.784** | 1.000 | -0.027 | **0.873** |
| Wipro | **0.816** | 0.001 | -0.027 | 1.000 | -0.218 |
| TechM | -0.333 | 0.647 | **0.873** | -0.218 | 1.000 |

Table 3 Correlation matrix IT sector



| | BPCL | Cairn | Gail | IOC | ONGC | Reliance |
|---|---|---|---|---|---|---|
| BPCL | 1 | 0.223 | -0.694 | -0.654 | -0.178 | -0.408 |
| Cairn | 0.223 | 1 | -0.084 | -0.183 | 0.343 | 0.264 |
| Gail | -0.694 | -0.084 | 1 | **0.890** | 0.532 | **0.828** |
| IOC | -0.654 | -0.183 | **0.890** | 1 | 0.454 | 0.706 |
| ONGC | -0.178 | 0.343 | 0.532 | 0.454 | 1 | 0.567 |
| Reliance | -0.408 | 0.264 | **0.828** | 0.706 | 0.567 | 1 |

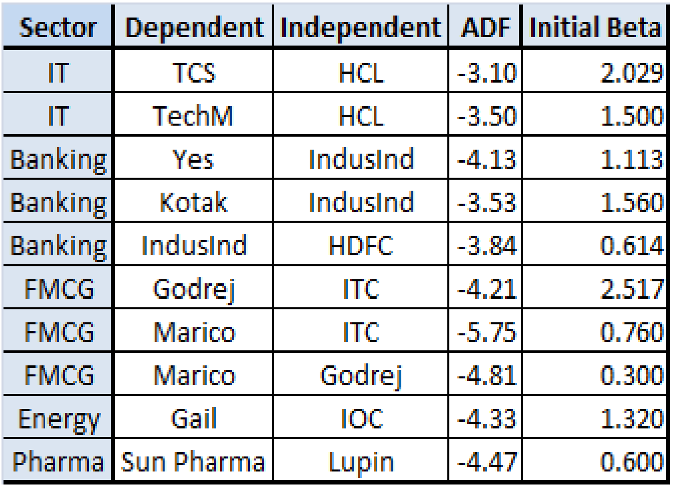
Table 4 Correlation matrix energy sector



| | ITC | HUL | Dabur | Godrej | Colgate | Britannia | Marico |
|---|---|---|---|---|---|---|---|
| ITC | 1.000 | 0.951 | 0.861 | **0.965** | **0.955** | 0.608 | **0.975** |
| HUL | 0.951 | 1.000 | 0.798 | 0.928 | 0.919 | 0.596 | 0.924 |
| Dabur | 0.861 | 0.798 | 1.000 | 0.897 | 0.785 | 0.398 | 0.868 |
| Godrej | **0.965** | 0.928 | 0.897 | 1.000 | 0.925 | 0.514 | **0.958** |
| Colgate | **0.955** | 0.919 | 0.785 | 0.925 | 1.000 | 0.661 | 0.949 |
| Britannia | 0.608 | 0.596 | 0.398 | 0.514 | 0.661 | 1.000 | 0.616 |
| Marico | **0.975** | 0.924 | 0.868 | **0.958** | 0.949 | 0.616 | 1.000 |

Table 5 Correlation matrix FMCG sector

We now conduct a cointegrated ADF test at the 90th percentile on beta-adjusted spread i.e. $X-\beta*Y$ to check if the shortlisted pairs are cointegrated. The table below shows the summary of cointegrated pairs along with their t-statistic values and initial beta estimates.



| Sector | Dependent | Independent | ADF | Initial Beta |
|---|---|---|---|---|
| IT | TCS | HCL | -3.10 | 2.029 |
| IT | TechM | HCL | -3.50 | 1.500 |
| Banking | Yes | IndusInd | -4.13 | 1.113 |
| Banking | Kotak | IndusInd | -3.53 | 1.560 |
| Banking | IndusInd | HDFC | -3.84 | 0.614 |
| FMCG | Godrej | ITC | -4.21 | 2.517 |
| FMCG | Marico | ITC | -5.75 | 0.760 |
| FMCG | Marico | Godrej | -4.81 | 0.300 |
| Energy | Gail | IOC | -4.33 | 1.320 |
| Pharma | Sun Pharma | Lupin | -4.47 | 0.600 |

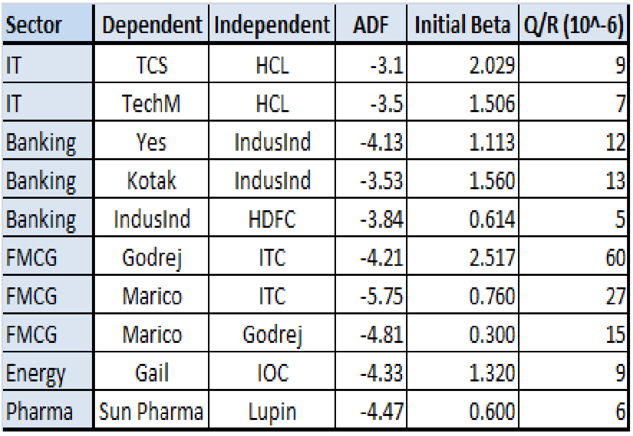
Table 6 Cointegrated Pairs across sectors with initial Betas

## V. SIGNAL-TO-NOISE RATIO(Q/R) TEST

One of the important aspects to look at when using Kalman filter algorithm for pair trading is to analyze the signal-to-noise (Q/R) for the pairs. Signal-to-Noise ratio (Q/R) is defined as the ratio of variance of beta process (Q) over the variance of the spread(R) i.e. $R = Y - \beta*X$. The idea behind using Q/R is that relation between two stocks must be more stable(Less volatile) than the stock process itself.

If the variance of beta process is low relative to price process we can determine beta quite accurately over time and obtain accurate estimates of true price Y based on X. Price process is given as $P=Y-Y*$ i.e. $Y-\beta*X$. We now compute Q/R for all the stocks and further shortlist stocks that have Q/R less 10^-3, as any Q/R with higher values will give noisy estimates of price process P.

From the table, it is evident that all shortlisted pairs have relatively lower signal-to-noise ratio, and hence the estimated beta can accurately predict the price process over time. We now calculate the betas of the stocks using data in the subset 1 and proceed with calibrating the algorithm using data in subset 2 for backtesting the strategy using data in subset 3.



| Sector | Dependent | Independent | ADF | Initial Beta | Q/R (10^-6) |
|---|---|---|---|---|---|
| IT | TCS | HCL | -3.1 | 2.029 | 9 |
| IT | TechM | HCL | -3.5 | 1.506 | 7 |
| Banking | Yes | IndusInd | -4.13 | 1.113 | 12 |
| Banking | Kotak | IndusInd | -3.53 | 1.560 | 13 |
| Banking | IndusInd | HDFC | -3.84 | 0.614 | 5 |
| FMCG | Godrej | ITC | -4.21 | 2.517 | 60 |
| FMCG | Marico | ITC | -5.75 | 0.760 | 27 |
| FMCG | Marico | Godrej | -4.81 | 0.300 | 15 |
| Energy | Gail | IOC | -4.33 | 1.320 | 9 |
| Pharma | Sun Pharma | Lupin | -4.47 | 0.600 | 6 |

Table 7 Cointegrated Pairs along with Signal-to-Noise ratio

## VI. STRATEGY DESCRIPTION

This strategy would involve a periodic rebalancing of the portfolio through long/short signals for the beta adjusted pair spread using the end of day prices. The threshold for entering into long or short position is designed based on a multiple of a standard deviation of the beta adjusted spread. For the purpose of analysis, we test this strategy using |1|*standard deviation and |1.5|*standard deviation bands.

Based on the current level of beta-adjusted spread long /short signals are generated as:-

Long spreads entry when spread < standard deviation of spread

Long spreads exit when spread > mean of spread (after crossing above the positive standard deviation band)

Short spreads entry when spread > standard deviation of spread

Long spreads exit when spread < mean of spread (after crossing the below negative standard deviation band)

| Generation of Long/Short Signals | |
|---|---|
| Long Entry | Spread < Std. dev of spread |
| Long Exit | Spread > Mean of Spread |
| Short Entry | Spread > Std. dev of spread |
| Short Exit | Spread < Mean of Spread |

Table 8 Long/Short Signal Generation

## VII. STOP-LOSS METHODOLOGY

As a measure of risk management and to prevent large losses a stop loss mechanism is included in the algorithm for individual pairs. The stop loss level is calculated using historical returns deviation of the pair. We run the pair trading algorithm for all the shortlisted pairs using data in the first subset and calculate the daily standard deviation of returns for individual pairs.

The intraday stop-loss threshold is then set as 2.5*(daily standard deviation) for the respective pairs. This means whenever the threshold is breached for any pair the positions are squared off. The multiple is chosen as 2.5 to ensure we do not incur any tail losses. Such a stop-loss measure provides a significant improvement in our backtested results.

## VIII. Performance analysis:

As we have completed all necessary prerequisite steps, we now run the algorithm on subset 3 to test the out of sample performance. The performance of this algorithm is analyzed at aggregate portfolio level as well as for individual pairs. We analyze the performance of the strategy at the aggregate level by allocating equal capital to each pair and aggregating gains and losses generated by them over time. The strategy is backtested using both |1|*σ and |1.5|*σ bands. We evaluate the aggregate performance of the strategy on basis of 4 parameters namely

*APR*-The annual percentage return generated by the strategy

*Sharpe ratio*- Defined as the ratio of excess return above standard deviation of returns

*Maximum Drawdown*- Peak-to-trough decline during a specific record period of a strategy. Usually quoted as the percentage between the peak and the trough.

*Calmar's Ratio*- Another risk-adjusted parameter defined as the ratio of APR over maximum drawdown.

The table below provides the summary of the performance of the strategy with and without use of stop-loss

| Stoploss | Band | APR | Sharpe | Max DD | Calmar |
|---|---|---|---|---|---|
| Yes | \|1σ\| | 17.70% | 1.93 | -3.27% | 5.41 |
| | \|1.5σ\| | 16.32% | 1.88 | -2.08% | 7.85 |
| No | \|1σ\| | 13.58% | 1.18 | -3.38% | 4.41 |
| | \|1.5σ\| | 12.57% | 1.02 | -2.74% | 4.58 |

Table 9 Portfolio strategy Performance with and without stop-loss

From the table, we observe that strategy provides better APR with |1|*σ bands, on the other hand, risk-adjusted parameters are fairly better for strategy using |1.5|* σ bands which is not surprising as the variability and frequency of long/short signal reduces as the we widen our standard deviation bands.

Strategy performs well using both bands, providing good Sharpe ratios, impressive drawdown, and Calmar's ratio. This is due to the diversification effect of strategy that is realized on aggregating the gains/losses generated by the individual pairs.

The table below shows the yearly return profiles of this strategy with stop-loss on both |1|*σ and |1.5|*σ bands

| Band | Year | APR | Sharpe | Max DD | Calmar |
|---|---|---|---|---|---|
| \|1σ\| | Year1 | 15.94% | 1.6 | -3.27% | 4.875 |
| \|1σ\| | Year2 | 19.13% | 2.54 | -1.50% | 12.75 |
| \|1.5σ\| | Year1 | 16.30% | 1.87 | -2.08% | 7.84 |
| \|1.5σ\| | Year2 | 16.34% | 2.12 | -1.72% | 9.22 |

Table 10 Yearly performance of the strategy using stop-loss

From the table, we observe that strategy using |1.5|*σ band performs marginally better in year 2 than in year 1, whereas strategy using |1|*σ band performs significantly better in year 2 than in year 1, and gives superior Sharpe, maximum drawdown, and Calmar values. The chart below shows the yearly return profiles for both the bands.

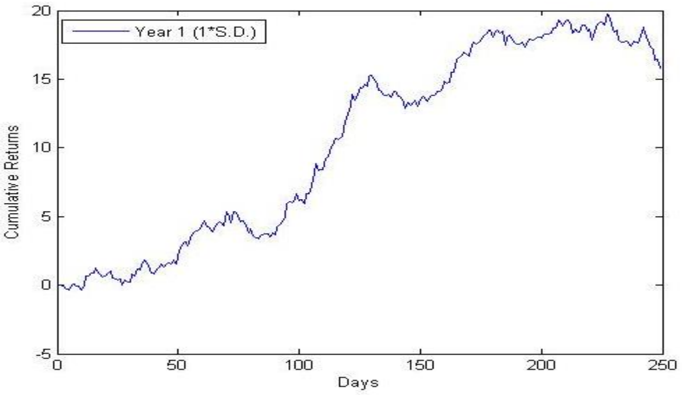


Fig.3 Cumulative returns of the strategy for 1st Year using |1|*SD bands

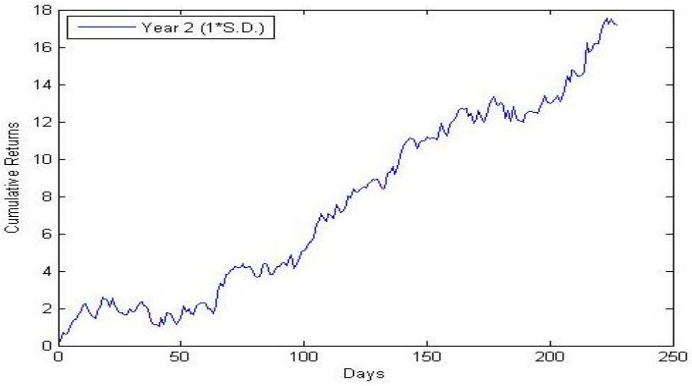


Fig.4 Cumulative returns of the strategy for 2nd Year using |1|*SD bands

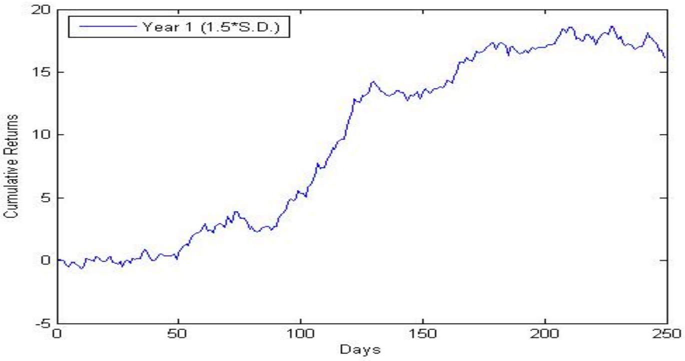


Fig.5 Cumulative returns of the strategy for 1st Year using |1.5|*SD bands

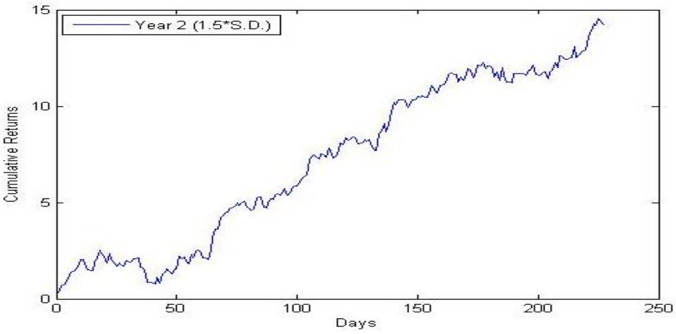
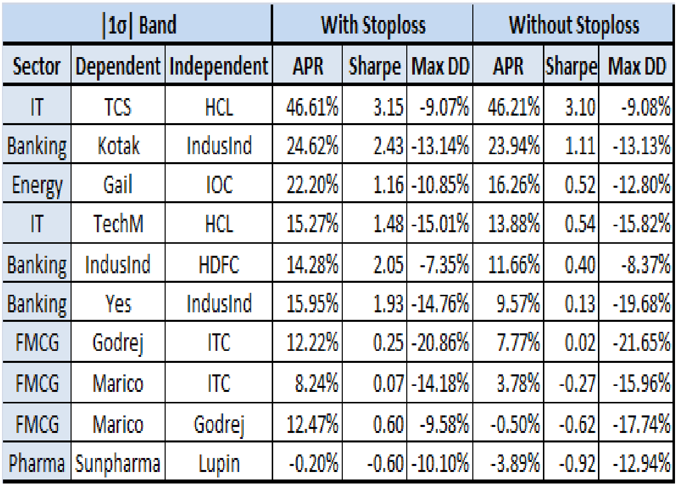


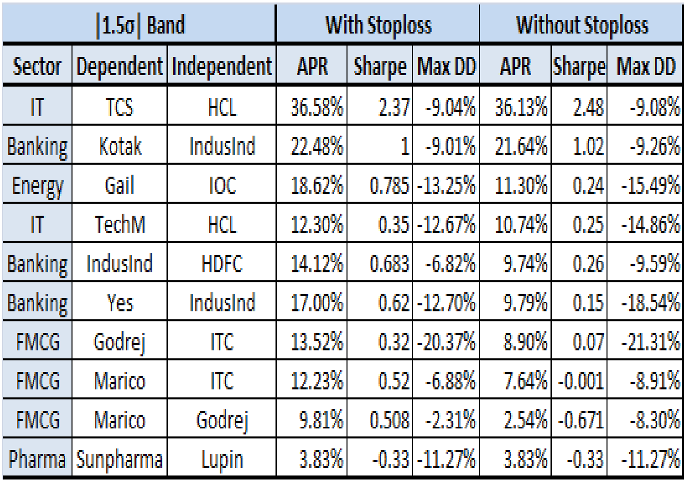
Fig.6 Cumulative returns of the strategy for 2nd Year using |1.5|*SD bands

The performance breakdown of individual pairs is shown in below table.



| \|1σ\| Band | | | With Stoploss | | | Without Stoploss | | |
|---|---|---|---|---|---|---|---|---|
| Sector | Dependent | Independent | APR | Sharpe | Max DD | APR | Sharpe | Max DD |
| IT | TCS | HCL | 46.61% | 3.15 | -9.07% | 46.21% | 3.10 | -9.08% |
| Banking | Kotak | IndusInd | 24.62% | 2.43 | -13.14% | 23.94% | 1.11 | -13.13% |
| Energy | Gail | IOC | 22.20% | 1.16 | -10.85% | 16.26% | 0.52 | -12.80% |
| IT | TechM | HCL | 15.27% | 1.48 | -15.01% | 13.88% | 0.54 | -15.82% |
| Banking | IndusInd | HDFC | 14.28% | 2.05 | -7.35% | 11.66% | 0.40 | -8.37% |
| Banking | Yes | IndusInd | 15.95% | 1.93 | -14.76% | 9.57% | 0.13 | -19.68% |
| FMCG | Godrej | ITC | 12.22% | 0.25 | -20.86% | 7.77% | 0.02 | -21.65% |
| FMCG | Marico | ITC | 8.24% | 0.07 | -14.18% | 3.78% | -0.27 | -15.96% |
| FMCG | Marico | Godrej | 12.47% | 0.60 | -9.58% | -0.50% | -0.62 | -17.74% |
| Pharma | Sunpharma | Lupin | -0.20% | -0.60 | -10.10% | -3.89% | -0.92 | -12.94% |

Table 11 Performance of individual pairs using |1|*SD bands



| \|1.5σ\| Band | | | With Stoploss | | | Without Stoploss | | |
|---|---|---|---|---|---|---|---|---|
| Sector | Dependent | Independent | APR | Sharpe | Max DD | APR | Sharpe | Max DD |
| IT | TCS | HCL | 36.58% | 2.37 | -9.04% | 36.13% | 2.48 | -9.08% |
| Banking | Kotak | IndusInd | 22.48% | 1 | -9.01% | 21.64% | 1.02 | -9.26% |
| Energy | Gail | IOC | 18.62% | 0.785 | -13.25% | 11.30% | 0.24 | -15.49% |
| IT | TechM | HCL | 12.30% | 0.35 | -12.67% | 10.74% | 0.25 | -14.86% |
| Banking | IndusInd | HDFC | 14.12% | 0.683 | -6.82% | 9.74% | 0.26 | -9.59% |
| Banking | Yes | IndusInd | 17.00% | 0.62 | -12.70% | 9.79% | 0.15 | -18.54% |
| FMCG | Godrej | ITC | 13.52% | 0.32 | -20.37% | 8.90% | 0.07 | -21.31% |
| FMCG | Marico | ITC | 12.23% | 0.52 | -6.88% | 7.64% | -0.001 | -8.91% |
| FMCG | Marico | Godrej | 9.81% | 0.508 | -2.31% | 2.54% | -0.671 | -8.30% |
| Pharma | Sunpharma | Lupin | 3.83% | -0.33 | -11.27% | 3.83% | -0.33 | -11.27% |

Table 12 Performance of individual pairs using |1.5|*SD bands

TCS-HCL was the best performing pair for both bands and on analyzing the performance of strategy at sector level, we see that Information Technology and Private Sector banks performed well compared to other sectors whereas Pharmaceutical sector was the worst performer. This could be attributed to the fact that companies in Information Technology and Banking sector as affected by similar set of fundamental, regulatory and economic factors and hence tend to be more correlated with other stocks within the sector, whereas companies in pharmaceutical sector tend to be relatively uncorrelated with each other, as each stock is affected by parameters like FDA rulings, R&D, and new drug development that are company specific.

## IX. Conclusion and Future Scope of the Strategy:

Based on the backtested performance we found that the strategy performs relatively well giving good annual returns and provides very good risk-adjusted returns on a portfolio level. The stop-loss methodology used apart from being a good measure of risk management also helps in improving the performance of strategy. For optimal performance of the strategy monthly/bi-monthly recalibration must be undertaken to check for cointegration of pairs and recalculation of returns deviation for stop-loss calculation, Return profiles, Maximum Drawdown and such. The returns we observed do not take leverage into consideration, the annual returns provided by the strategy can be significantly magnified through leverage. Hence to further improve the strategy we suggest using a metric like Kelly's formula defined as ratio of mean excess return by variance of excess return to incorporate leverage into our strategy and to further optimize the strategy we can add signal-to-noise ratio filter which will ignore trading signals when this ratio corresponding to a particular pair under consideration exceeds some specified level to avoid spurious estimates of the price process. Another way to optimize the strategy at portfolio level would be to periodically allocate varying weights to individual pairs depending on its performance, we can define a fixed minimum and maximum weight constraints for all pairs and calculate the weight for all pairs in the portfolio using an optimizer that maximizes a fitness function like Sharpe or Calmar's ratio under the defined constraints.

## References

[1] Robert J. Elliotty, John Van Der Hoek*z and William P. Malcolm ”Pairs Trading”, 11 April 2005

[2] Yuxing Chen (Joseph), Weiluo Ren (David), Xiaoxiong Lu” Machine Learning in Pairs Trading Strategies”

[3] Daniel Herlemont “Pairs Trading, Convergence Trading, Cointegration ”

[4] Ganapathy Vidyamurthi “Pairs Trading: Quantitative Method and Analysis ” Wiley Finance; pp 52-64, 73-75

[5] Ernest P. Chan “Algorithmic Trading: Winning Strategies and their Rationale” Wiley; pp 75-83

[6] Damodar Gujarati “Basic Econometrics:4th Edition” ; pp 792-834