Statistical Analysis of Strike Independent Volatility

MS&E 444 – Investment Practice

Babajide Kolade

Rajiv Kurian

Bilal Bidawi

Rana Mansoor Ali Khan

Project Advisors: Professor Kay Giesecke, Jeremy Evnine (EvA), Lisa Borland(EvA)

Abstract:

In this paper we study the strike- independent implied volatility changes. The strike independent volatility offers reduced dimensionality and thus can be used to identify the patterns that might exist in the volatility space and the stock space. In this paper we perform the Principal Component Analysis on the implied volatility, historical volatility and stock returns to study whether some clustering can be observed along these measures and whether there exists any correlation between these clusters. The second part of the paper is to look at the residual returns in the implied volatility space and in the normal stock return space. We study whether the residuals can be modeled properly and whether we can form a trading strategy using this model.

1) Introduction:

The q-alpha sigma model can be employed to reduce the dimensionality of the volatility space and it was created to address the deficiency of the famous Black-Scholes model. The Black-Scholes model assumes that the underlying stock follows a geometric Brownian motion and has constant volatility. In practice however the volatility surface fluctuates over a period of time as it varies along both strike and time to maturity. Due to this reason the Black-Scholes model fails to accurately predict the option prices. To address this deficiency in the Black-Scholes models, many models have been developed that study the variation of the volatility across strike and time to maturity. Q-alpha sigma model is one such model.

The q-alpha model replaces the underlying Brownian motion of the stock with a stochastic process that allows for statistical feedback as a model for the underlying stock. Such a model allows to accurately modeling of the returns' distribution matching those that are found empirically. The q-alpha sigma model derives the stochastic model from a class of processes that have been developed in statistical physics under the field of “Tsallis Nonextensive Thermo statistics”, (Borland, 2002). The q-alpha sigma model interprets the stochastic process as if the process is governed by a fat-tailed Tsallis distribution. One can use the model to invert the q-alpha-sigma implied volatilities from observed option prices and can reduce the volatility surface to become nearly flat across the strikes. This way the two-dimensional volatility surface can be reduced to a single dimension.

We perform the statistical analysis of this strike independent volatility by performing the Principal Component Analysis (PCA) and applying the Generalized Autoregressive Heteroskedasticity (GARCH) model to the volatility fluctuation of the 65 stocks in the

S&P 100. We take the implied volatility data for these stocks from January 1, 2006 till August 10, 2008 and break the data into quarters and maturities to study the short term effect of the variation in volatility and to identify the quarter-on-quarter patterns that can exist in this data. We also take the financial data like Revenue, Market Cap, Assets and Revenue to Market Cap ratio[1] of these stocks to study the impact of change in volatility on the financial indicators. To study the impact on long term basis, we also analyzed how the changes in volatility impact the data annually.

2) Data Collection:

We had primarily two sources of data:

1- The Q-Alpha implied volatility varying by ticker, date, and maturity of the option. It was provided by Eva. The data covered implied volatility quotes between the start of 2006 till the end of 2008. However, the main issue with this part of the data was that it wasn't homogeneous across tickers: The dates when implied volatility data was available were not the same for all tickers.
2- The price data of all the tickers. The data is available widely and we relied primarily on Stanford Business School library.

In order to conduct a "meaningful" PCA, We should find a set of dates which will have implied volatility data available for all tickers. Unfortunately, we couldn't simply take the common dates for all tickers because this set will become very small or empty (depending on the maturity). We thought about two different ways of solving this problem:

1- Interpolate the implied volatilities on the missing dates
2- Choose a subset of tickers having a substantial number of dates

We quickly discarded the first approach given the high variance of the volatility from date to date. Using smoothing to complete the missing data would have an impact on the validity of the results. The second solution needs to be more precise: to what extent we're willing to leave out tickers for the benefit of the number of dates? We decided that the number of tickers should never get less than 50 (out of 72), while maximizing the number of dates.

[1] Compustat Database: http://wrds.wharton.upenn.edu.gsbproxy.stanford.edu/ds/comp/secm/

This problem is algorithmically very difficult: We need to find the subset (of cardinal greater than 50) having the maximal number of common dates. The size of this problem is around 2^{71}. The only way to solve this problem efficiently is to find an approximate solution (not the optimal one, but a solution that is satisfying) by taking advantage of the particular form of the problem we have: there is a lot of clustering between tickers on certain dates.

We define a good "approximate" solution by a set that would have 100 dates less than the maximum number of dates available (This will leave us with a subset of tickers having at least 200 common dates per maturity which is deemed enough to run the PCA)

We use a greedy algorithm to find the approximate solution: we first put together all the tickers that have the same dates. We group the subsets by pairs per iteration. The pairs are formed between the closest subsets (the distance between two subsets is the number of dates lost by the merging of these two subsets) and we merge the pair if the subsets are close enough.

The algorithm is as following:

Let $T_1, T_2, \dots, T_N$ be the tickers. Let D_i be the set of dates of the ticker LetT_i. Define the distance between two sets A, B as $d(A,B) = card(A \Delta B)$

-Initialization:

Let$T = \{T_1, T_2, \dots, T_N\}; i = 1$;

While $(T \text{ is not empty})$

Take T_j one element of$T$, delete T_j from T

Define $S_i = \{T_j\}$

For all T_k in T

If ($d(T_j, T_k) = 0$) add T_k to S_i and delete T_k from T

End

Increment I;

End

-Start Greedy Algorithm

Let Tolerance = 50 (we are using here $\sum \frac{50}{2^i} \leq 100$)

j=0

While (a change in the sets happens)

For each Set S_i

Find the closest set S_k to S_i as

If $(d(S_k, S_i) \leq \frac{Tolerance}{2^i})$

Merge S_k to S_i

update the common dates for the new set

Signal that a change happened

End

End

End

By the end of the algorithm, we end up with big sets of tickers who have many dates in common. Choosing the set having the biggest number of dates was always enough in our case to find the set of tickers that would be the same for all maturities and that would have enough number of dates. We were planning, in case we don't get a set of size at least 50 tickers, to reduce the size of tolerance to reduce as much as possible the size of the problem, without missing on a lot of dates. Then we explore all the case possible by exploring the tree of all possibilities using the branch and bound technique to find the optimal solution as soon as possible.

2.1) Financial Data

The financial indicators i.e. Revenue, Market Capitalization, Assets, Revenue to Market Capitalization ratio are extracted from the Compustat Database. While the revenues for the stocks are taken to be quarterly, we take the average number in the case of Market Cap, Assets and Revenue to Market Capitalization ratio like for example, to compute the market capitalization we took the average outstanding shares of a particular company

for a quarter and multiplied it with the average stock price for that company for that quarter.

2.2) Categorization of the Stocks

Based on the financial data, we categorized the companies into two segments

1. Octile Categorization
2. Mean and Standard Deviation Categorization

Like the name implies, the octile categorization divides the stocks into 8 categories along each financial indicator; with 1 indicating the set of companies that has the smallest contribution to that financial indicator and 8 being the set of companies that has the largest contribution to that indicator.

The mean and standard deviation categorization divides the data according to the spread of the company's financial indicator to the overall mean and standard deviation of the financial indicator. Like for example, if the revenue of a particular company is less than the mean of the revenues of the 65 stocks minus 1.5 times the standard deviation for the stocks, we categorize this company as 1. Similarly we have 7 categories based on the mean and standard deviation spread of the companies' data. The detailed list can be found in appendix.

3) PCA Analysis Methodology

Principal component analysis (PCA) is employed in determining the systematic component of the financial data is standard and is along the lines of such utilized by Avellaneda and Lee, 2008. We briefly describe the methodology here.

First, the multivariate time series data is standardized: the mean of each variable is subtracted from the observation and the result is divided by the standard deviation of the variable. Then, a correlation matrix is calculated from the standardized input. The eigenvectors and eigenvalues of the correlation matrix are determined. The correlation matrix is a real symmetric matrix and therefore Hermitian. So its eigenvalues are real, finite, and orderable. The eigenvectors of the correlation matrix are sorted in order of decreasing magnitude of the corresponding eigenvalues. These eigenvectors comprise a set of orthonormal basis vectors which spans the space of the input data. One may now recast the input data as projections onto this space.

Next, the number of eigenvalues/ eigenvectors used in the reduced dimensionality is determined. Three options are provided in our implementation: 1) a fixed fraction of

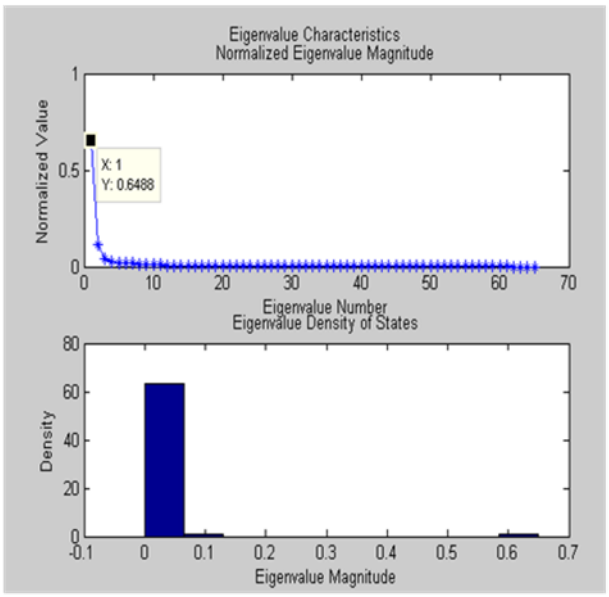
the total number of eigenvalues (fixed fraction of the rank of the correlation matrix), 2) a specified percentage of the explained variance: (That is, include enough eigenvalues that sum up to a specified percentage of variance starting from the largest eigenvalue.), and 3) include all the eigenvalues that are within a specified standard deviation away from the largest eigenvalue. The number of eigenvalues/ eigenvectors represents the cardinality of the reduced space and corresponds to the number of risk factors driving the systematic dynamics of the market. The time-varying risk factors are calculated from a dot product of original input data and the selected eigenvectors. The correlations of individual time series to the risk factors, beta, are determined by least squares. The remainder of the time series after accounting for the systematic portion is the residual/ idiosyncratic portion.

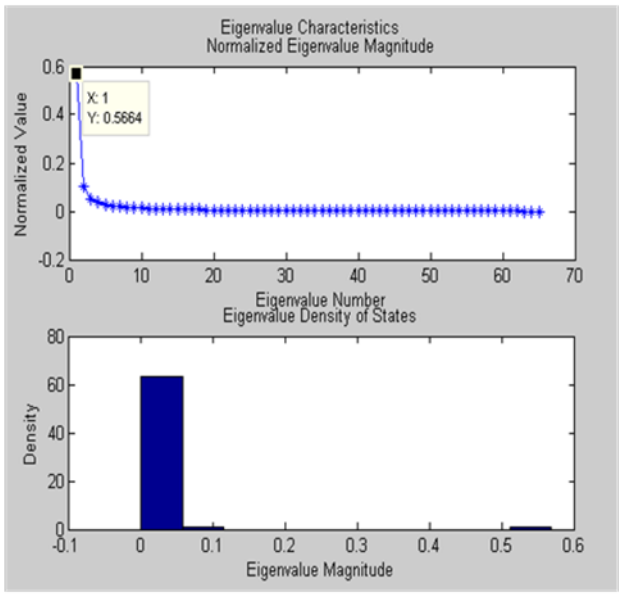
The market neutral portfolios are a set of vectors that are uncorrelated with the risk factors. The market neutral portfolios sits in the null space of the correlation factors, beta. This indicates that the number market neutral portfolio is equal to the difference between the rank of the correlation matrix and the number of risk factors. The above-described methodology results in the determination of the risk factors, the correlation coefficients (beta) of each time series with the risk factors, the residuals (idiosyncratic portion), and the market-neutral portfolios.

4) PCA Results

We have performed the Principal Component Analysis on quarterly data ranging from the 1st quarter of 2006 till the 2nd quarter of 2008. The PCA was performed on the price, q-alpha implied volatility and the historical volatility domains. We then analyzed the various plots from each PCA and made the following observations:

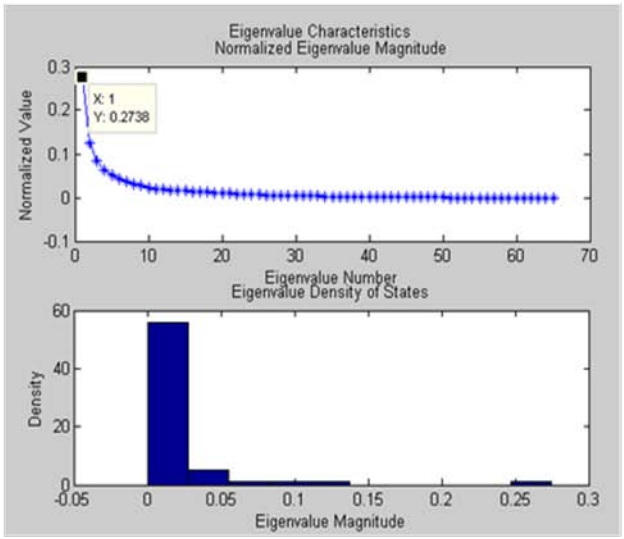
1) Maturity does not affect the spectrum of eigenvalues: We see that the eigenvector characteristics have very minor changes in the q-alpha implied volatility domain, when we observe data for the same quarter, but for different maturities. Out of the twelve unique quarters, in 11 quarters the normalized eigenvalue magnitude of the first eigenvector varies by less than 10% (absolute) as we vary the maturity date of the option. It can be also seen that the density of eigenvalue magnitudes is very similar for the same quarters in spite of changing the maturity dates. The following plot shows one such case where the absolute difference between the normalized eigenvalue magnitudes is 0.0824%:

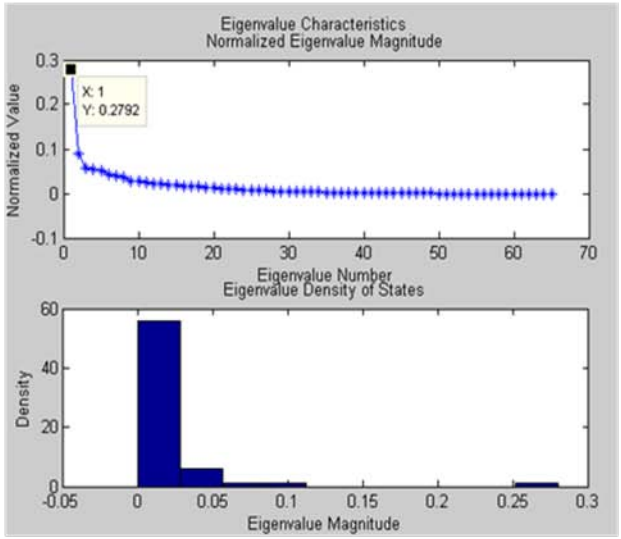




Qtr 2 06 Maturity 07, 08

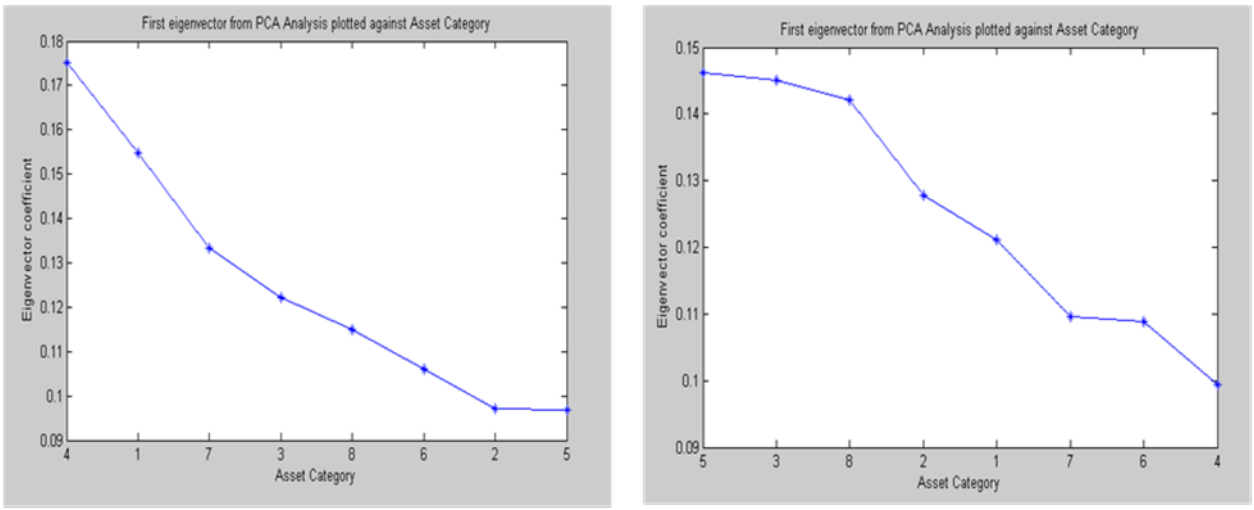
2) The normalized magnitude of the principal factor in the price space and volatility space are correlated: We also notice that the normalized magnitude of the principal factor in the price space and the q-alpha implied volatility space are very similar for the same quarter. Out of the 12 unique quarters this behavior can be observed for 11 quarters. The following plot shows one such case where the absolute difference between the normalized value of the principal factors in the price space and q-alpha implied volatility space is 0.0052%.





Q-alpha implied vol qtr1 2006, maturity 2007, Price qtr1 2006

3) Irrelevancy of Economic Indicators: We have plotted the first four eigenvectors from the PCA in the q-alpha implied volatility, historical volatility and price domains against the asset, market-cap, revenue and revenue to market cap categories (using the octile ranking method). We have observed no clustering of particular groups in any of the above plots. Further there seems to be no relation between the plots in the q-alpha implied volatility, historical volatility and price spaces for the same observation period. No relationship was established on either changing the maturity period for the same quarter or changing the observation period for the same maturity. This has led us to believe that economic indicators are irrelevant to the outcome of the PCA. The following plot shows a lack of relation as we transition from quarter 1 2006 to quarter 2 2006 for the same maturity(January 2007):

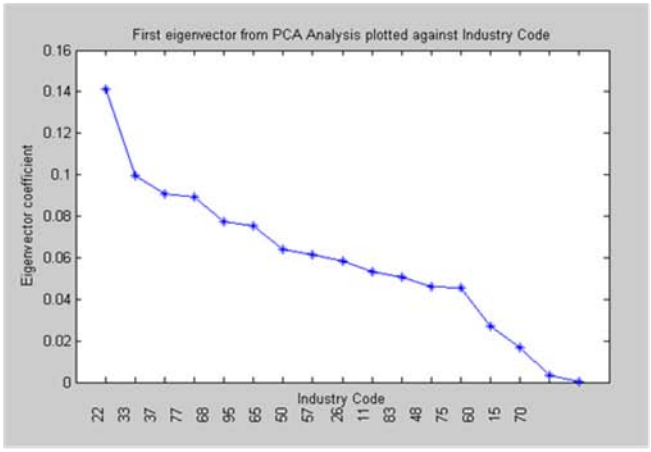


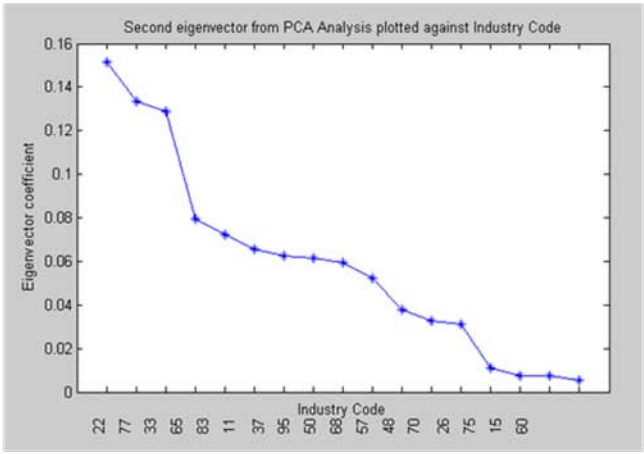
1st, 2nd Qtr 06 Maturity 07 q-alpha volatility

4) Impact of Industry: We have plotted the first four eigenvectors from the PCA in the q-alpha implied volatility, historical volatility and price domain against the industry category. The industry categorization was performed using the industry and sector categories listed on yahoo finance[2]. The detailed categorization can be found in Appendix 1. We observe two out the three industries – 77 (Oil), 22 (Banks) and 33 (Consumer Products) in the first three factors. This clustering occurs in the first four eigenvectors in all plots (historical volatility, q-alpha implied volatility and prices domain). If industry was a significant driving factor of the markets, if we observe a particular set of industries as principal factors in

[2] Yahoo Finance: http://finance.yahoo.com

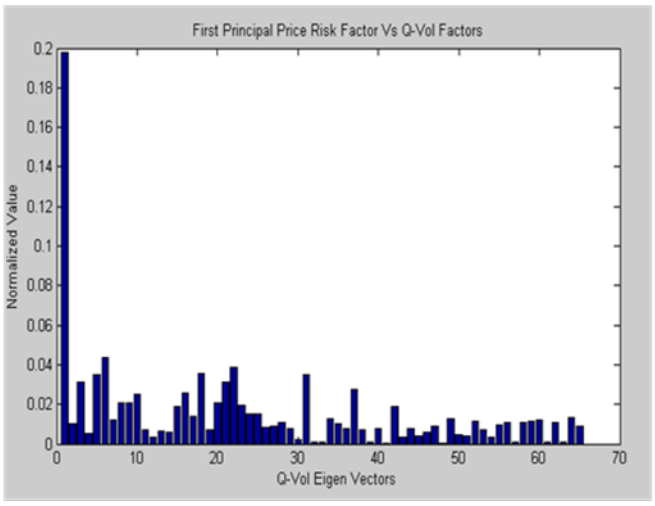
one eigenvector we would expect them to have minimal contribution in the other eigenvectors. Since the same industries are seen as major factors in both the principal as well as secondary eigenvectors we conclude that industry is not a sole driving factor of the market. The following plot shows banks (22), oil (77), consumer products (33) dominate in both the primary and secondary eigenvectors derived from the PCA done on q-alpha implied volatility observed during quarter 1 2006 (January 2007 Maturity):

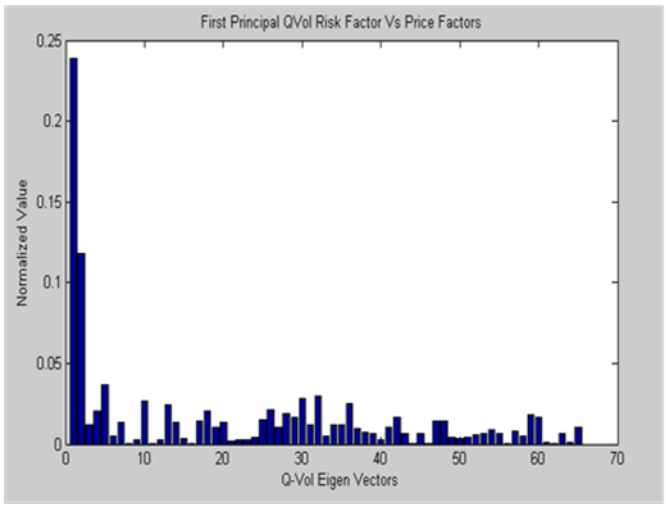




Qtr 1 06 Maturity 07

5) Relationship between the price and volatility risk factors: We plot the risk factors in the q-alpha volatility space in terms of the corresponding risk factors in the price space and vice versa. We observe that the first risk factor in the q-alpha implied volatility space accounts for a large portion of the first risk factor in the price space for 26 times out of the 48 total observations. This leads us to believe that the principal factor in the price space is correlated with the principle factor in the volatility space. The following plots show a significant correlation between the principal factor in the volatility space and the principal factor in the price space.





Qtr 3 07 Maturity 08

5) Residual Modeling:

5.1) Introduction

Financial time series data exhibit heteroskedastic behavior: The volatility (standard deviation) of the time series vary over the history of the data. This behavior is often evident in time history plots of the data where portions of the data are visibly more dynamic that other portions. This is termed volatility clustering. Note that data that does not exhibit volatility clustering may still be heteroskedastic. An appropriate model of financial data must have the capability of capturing the heteroskedastic nature of the data.

5.2) Residual Modeling Methodology

In our analysis, financial data are decomposed into a systematic component and an idiosyncratic component. The following reintroduces the model used in representing financial data and introduces the details of the residual modeling.

$$\phi_t = \sum_{j=1}^{M} \beta_j F_{jt} + \tilde{\phi}_t \text{ ----------(1)}$$

Where:

ϕ_t: Modeled quantity

F_j: Risk factor j

β_j: Correlation coefficient to risk factor j

$\tilde{\phi}$: Idiosyncratic component

In Equation $\phi_t = \sum_{j=1}^{M} \beta_j F_{jt} + \tilde{\phi}_t$ ----------(1 ϕ is a scalar quantity, indexed by time, which is to be modeled. The quantity is divided into a systematic component and an idiosyncratic portion. The systematic component is a linear combination of the correlations of quantity with the system's risk factors, F. The risk factors may be common to quantities of the same type. It is implied that the risk factors are the underlying principals that drives the core evolution of the modeled quantity.

The second term in Equation $\phi_t = \sum_{j=1}^{M} \beta_j F_{jt} + \tilde{\phi}_t$ ----------(1 is the idiosyncratic component of the quantity. It is peculiar to quantity being modeled (within the class of scalars), stochastic in nature, and encompasses the variance of the scalar.

$$\tilde{\phi}_t = \int_{t_o}^{t} \alpha_{t'} dt' + \left\langle \tilde{\phi} \right\rangle + \sigma_t dz \text{ ---------- (2)}$$

Where:

α: Residual drift

$\left\langle \tilde{\phi} \right\rangle$: Mean residual

σ^2: Conditional variance

dz: Independent identically distributed random variable of specific distribution

The drift term represents the stationarity of the data. It is often the case that the data is stationary, especially when the data is derived from the return of another time series

($\phi = \int_{t_o}^{t_o+dt} d\varphi / \varphi$), and hence α is zero. The stationarity of the data may quantified by calculating the integral time scale of the residual. Integral time scale quantifies the length, on average, for which the data is correlated. The integral time scale is a concept used in analysis of turbulent flows in fluid mechanics and it may be calculated using the equation below

$$T_{\text{int}} = \int_{0}^{t\_cross} C_{xx}(t)dt \Big/ C_{xx}(0) \quad \text{---------- (3)}$$

Where:

T_{int} : Integral time scale

C_{xx} : Autocovariance function of data

$C_{xx}(0)$ Zero-lag autocovariance of data (variance of data)

t_{cross} : Time at which the value of the autocovariance function first crosses zero

When the integral time scale, as calculated above, has a value that is lower than the time resolution, the data is stationary (for example, getting an integral time scale of 0.5 days on a daily return data). For all the data sets analyzed in this project the drift term is zero. The systematic component of the modeled quantity is derived using PCA analysis, the drift is typically zero, and the determination of the mean residual is trivial. Therefore the goals of the residual analysis, thereby completing the statistical model, are to quantify the conditional variance and identify the probability distribution.

The conditional variance and probability distribution of the random component are determined using the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) methodology (Engle 1982, 2001, Bollerslev 1986). The equation below states the GARCH model for conditional variance:

$$\sigma_t^2 = \kappa + \sum_{k=1}^{P} G_k \sigma_{t-k}^2 + \sum_{k=1}^{Q} A_k (\sigma dz)_{t-k}^2 \quad \text{---------- (4)}$$

Where:

σ_t^2: Conditional variance

κ: Variance constant term

dz: Independent identically distributed random variable

(σdz): Error term

G_k: Variance lag coefficients (GARCH coefficients)

A_k: Error lag coefficients (ARCH coefficients)

P, Q: Order of the GARCH model

The GARCH models considered in the context of this project are GARCH(1,1), GARCH(1,2), GARCH(2,1), and GARCH(2,2). The first index of the specification indicates the number of serial dependence of the conditional variance on past variances (the P terms or GARCH coefficients) and the second index the serial dependence of conditional variance on past error terms (the Q terms or ARCH coefficients). The probability distributions considered are the standard normal distribution and the Student-T distribution. An algorithm determines the appropriate GARCH model order and the probability distribution.

The model coefficients for a particular GARCH model order and distribution are determined by maximizing the of log-likelihood function. The procedure starts by assuming the appropriate distribution and GARCH model are the standard normal distribution and the GARCH(1,1) model, respectively. In statistical language, this set as the null hypotheses. Likelihood ratio test is used to determine whether or not to reject the null hypothesis that the standard normal distribution as the appropriate distribution. Then the algorithm successively tests whether or not to reject the

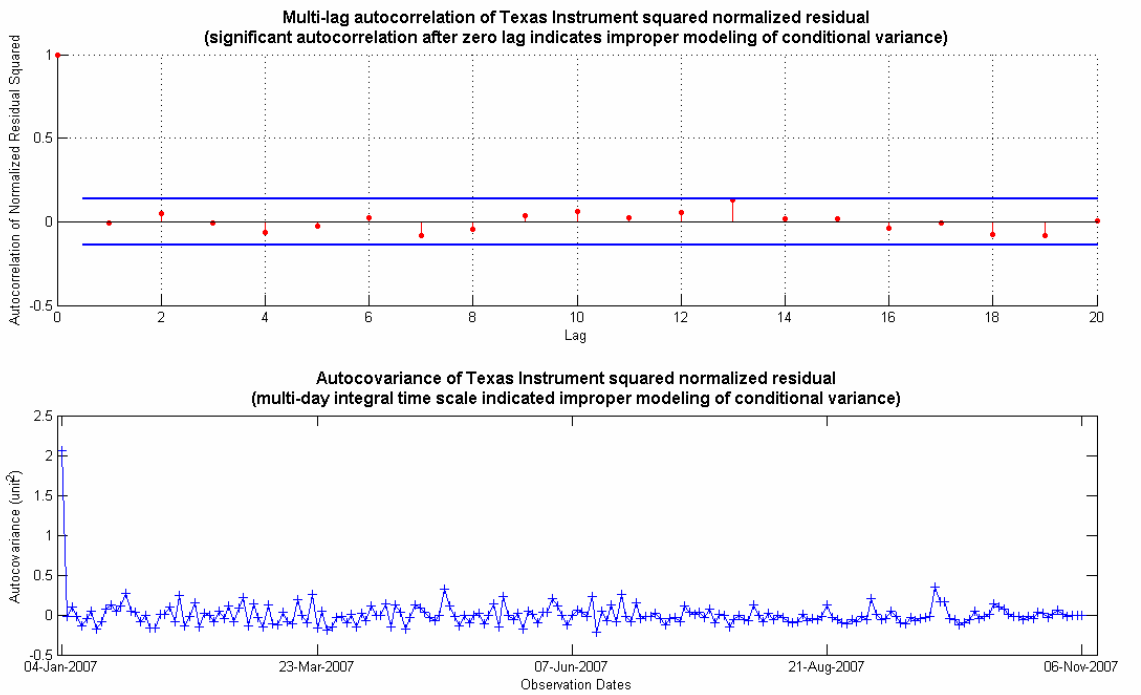
hypothesis of a lower order GARCH model, also using the likelihood ratio test. The hypothesis testing significance level used in the distribution test is 0.1, for GARCH(1,2) and GARCH(2,1) tests 0.05, and for the GARCH(2,2) test 0.02. At the end of this procedure, a GARCH model order and a probability distribution will have been selected and the model coefficients determined.

6) Residual Modeling Results

One could rearrange Equation

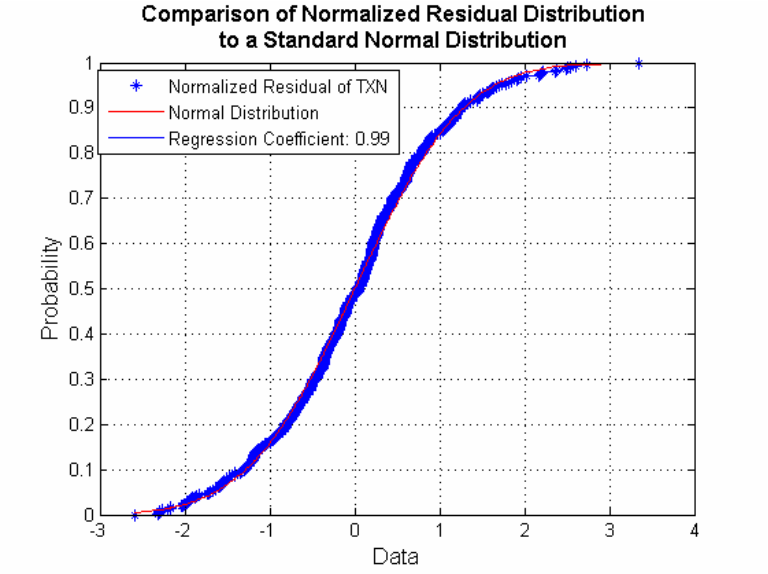
$$\tilde{\phi}_t = \int_{t_o}^{t} \alpha_{t'} dt' + \left\langle \tilde{\phi} \right\rangle + \sigma_t dz$$ ---------- (2 to isolate the normalized residual term: $dz = \left(\tilde{\phi}_t - \left\langle \tilde{\phi}_t \right\rangle\right) / \sigma_t$. Here it is assumed that the drift term is zero. This rearrangement indicates that a successful model of the residual will result in a normalized residual that has no serial correlation and is identically distributed with a specified probability distribution.

We conduct two tests to check the success of the residual modeling. The autocorrelation function of the squared normalized residual is calculated to check for serial dependence. Significant values of the autocorrelation function after the zero lag indicates serial correlation and hence improper modeling of the residual. For all the data analyzed in this work (price return, absolute q-alpha volatility, q-alpha volatility return, and absolute historic volatility), GARCH residual modeling was successful in removing serial correlations in the normalized residual. Example plots of the autocorrelation function are shown below



Autocorrelation plots of normalized residual of Texas Instrument price return from January to November 2007. The systematic component was calculated using PCA analysis of stocks in the S&P100 with 60% of the variance included. The conditional variance of residuals was modeled using GARCH(1,1) and the distribution was determined to be Gaussian.

The normalized residual is also tested for a goodness of fit to the determined probability distribution. The sample cumulative distribution function (CDF) is compared to the theoretical CDF of the probability distribution. The sample CDF at a sample space state is fraction of observations that have a lower state value than the state variable in consideration. As defined, the sample CDF will have a range of [0 1). The goodness-of-fit (gof) is calculated as $gof = 1 - \sum_{N} \left(CDF_{theoretical} - CDF_{sample} \right)^2 \Big/ N$. As defined, gof has a range of (0 1]. An example plot of the CDF comparison test is shown below



Comparison of the distribution of the normalized residual of Texas Instrument price return to a standard normal for data from January to November 2007. The systematic component was calculated using PCA analysis of stocks in the S&P100 with 60% of the variance included. The conditional variance of residuals was modeled using GARCH(1,1).

For a preponderance of the residuals analyzed, GARCH(1,1) and standard normal distribution are sufficient to properly model the residuals.

6.1) Persistence of Residual Model

Using the above-described methodology, stock price return and absolute q-alpha volatility data from January 2006 to June 2008 were segmented into quarterly bins and analyzed. The systematic components for the data for these 10 quarters were determined and the residuals analyzed. The residual model (GARCH parameter coefficients and normalized residual distribution) derived for each quarter is subsequently used to predict the conditional variance of the last quarter (April-June 2006). The performance of these predictions indicates the relative stability and persistence of the residual models. The figures below shows the result of this analysis

for Texas Instrument price return for the time frame stated above. It is seen that the residual model performs well a year out and starts to deviate after this.

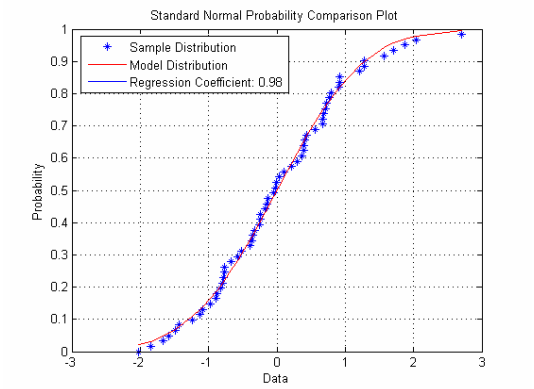


Figure A: Residual model parameters from quarter ten used to predict conditional variance of quarter 10

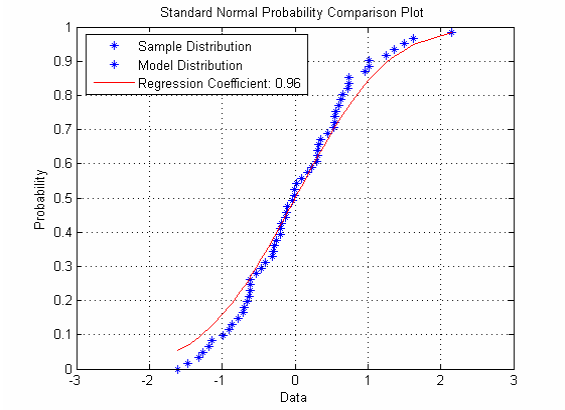


Figure B: Residual model parameters from quarter 9 used to predicts conditional variance of quarter ten

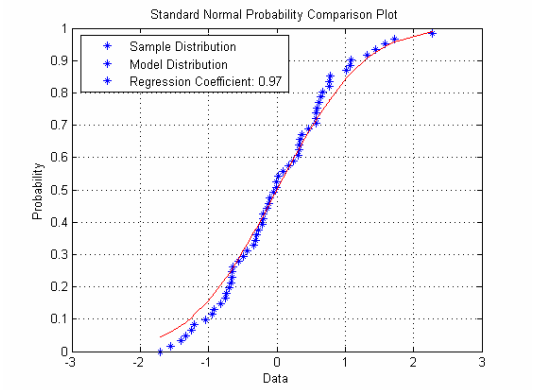


Figure C: Residual Model parameters from quarter 8 used to predict conditional variance of quarter ten

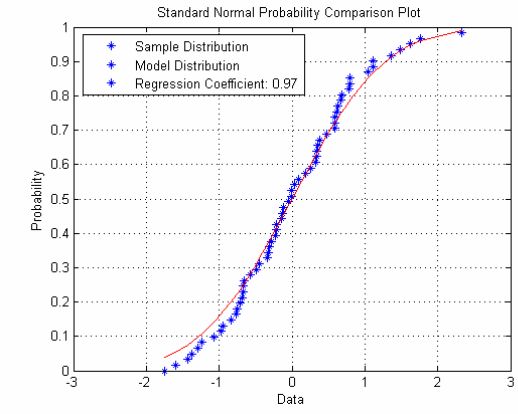


Figure D: Residual model parameters from quarter 7 used to predict conditional variance of quarter ten

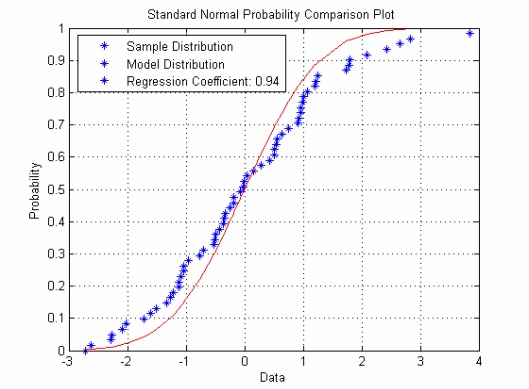


Figure E: Residual model parameters from quarter 6 used to predict conditional variance of quarter ten

7) Trading Strategy based on Residual Model

The above plots show that the residuals may be modeled properly and that the normalized residuals are well represented by a specified distribution (often the standard normal distribution). Hence the probability of an excursion of the normalized residual from its standard state may characterized properly. Therefore, one may devise a hedging strategy off this model. The hedging strategy will involve tracking individual stock returns normalized residuals (or volatility normalized residuals in the case of volatility trading) for excursions from the standard state. So for example if the normalized residual return of the TXN was 2, one could go long on systematic component of this stock and short the individual stock. This way the systematic component of the stock is hedged out and the strategy will capitalize on the stocks deviation from equilibrium.

8) Conclusions/ Recommendations

The following are the most important conclusions derived from the PCA plots and the residual modeling:

1) The normalized magnitude of the principal factor in the price space and volatility space are correlated.

2) The Banking, Retail and Energy sectors dominate the principal risk factors in the q-alpha implied volatility, historical volatility and prices space.

3) The conditional variance of the residual is adequately modeled by GARCH(1,1) and the normalized residual follow the standard normal distribution.

4) The conditional variance model parameters (GARCH model and coefficients) regressed from data from a quarter performs adequately in modeling the residuals one year out.

The following are the recommendations we have for future work built on this paper:

1) Derive a mathematical basis to justify the number of eigenvalues/ eigenvectors included in the PCA analysis.

2) Devise a more concrete method to quantify the relations observed in the PCA plots. A simple correlation between two factors could significantly improve our belief in the relations between different quantities.

3) Develop a trading /hedging strategy based on the modeling of the systematic component of the financial data, the prediction of the conditional variance, and the normalized residual model.

REFERENCES

Marco Avellaneda and Jeong-Hyun Lee. " Statistical Arbitrage in the U.S. Equities Market", 2008.

Tim Bollerslev. "Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics*, 31:307-327, 1986.

Lisa Borland. "A theory of non-Gaussian option pricing", QUANTITATIVE FINANCE VOLUME 2 (2002) 415–431.

Robert F. Engle. "Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation", *Econometrica* 50:987-1008, 1982.

Robert F. Engle. "GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics", *Journal of Economic Perspectives* 15(4):157-168, 2001.

APPENDICES

Appendix 1: Categorizations used for the PCA Plots

1) Economic Indicators - Mean and Standard Deviation Category:

KPI Category Criterion: Mean and Std Dev	
1	< (Mean - 1.5* Std)
2	(Mean - 1.5 * Std) to (Mean - 0.75 * Std)
3	(Mean - 0.75 * Std) to (Mean - Std/4)
4	(Mean - Std/4) to (Mean + Std/4)
5	(Mean + Std/4) to (Mean + 0.75 * Std)
6	(Mean + 0.75 * Std) to (Mean + 1.5 * Std)
7	> (Mean + 1.5 * Std)

2) Industry Categorization:

Code	Description
11	shipping/transportation
15	Aircraft
22	Commercial bank/ Financial Services/Securities
26	Insurance
33	Consumer Products
37	Restaurant/Retail
48	Lumber, wood and paper
50	Pharmaceutical
57	Hospital/ Healthcare
60	Entertainment
65	Computer software and Peripherals
68	Electrical Services/ Semi Conductor
69	Communication
70	Motors
75	Energy
77	Oil, Gas and Petrol
83	Plastics/Steel/Aluminum
95	Engineering Conglomerate
99	Others

Appendix 2: Programs used to extract data and model it. We give a brief description of the purpose, expected input and expected output for each program used. The detailed code can be provided on request.

1) Data_to_map_of_map_of_sets (C++):

    - Purpose: To convert given files to a map of map of sets(Maturities are keys to tickers which are themselves keys to a set of dates).
    - Input: Rows of data in the format -

      Current_Date Ticker Maturity_Date Q-alpha implied volatility
    - Output: Each unique combination of maturity date and ticker is mapped into a set of dates on which the q-alpha implied volatility for that ticker and maturity were recorded.

2) `Greedy_algorithm_to_get_maximal_data (C++):`

    - Purpose: To find a set of dates that has implied volatility data available for all tickers across at least four maturities
    - Input: The map of map of sets derived from the program Data_to_map_of_sets program.
    - Output: A list of tickers common for all maturities. A list of dates for each maturity where data is available for each of the common tickers and that maturity.

3) `Greedy_algorithm_to_pca_format (C++):`

    - Purpose: To extract the optimal data set from the master file and to record it in a format that the PCA routine recognizes.
    - Input: The optimal data derived from the Greedy Algorithm and the master file containing all the data.
    - Output: A Text file for each maturity and duration period, containing the data formatted such that each row has 65 columns one for each ticker. Each column

holds the q-alpha implied volatility for that date and each row represents a date for which the data is recorded for each maturity.

4) PCA_Analysis (matlab):

- Purpose: To perform PCA for a particular maturity and period of observation and to graph plots required to analyze the PCA.

- Input: PCA formatted files representing historical volatility, q-alpha implied volatility or price data. We also need the financial data for the 65 tickers for the period under consideration.

- Output: The PCA and the following plots:
  a) Eigen Vector Characteristics
  b) Evolution of Principal Eigen portfolio
  c) Eigenvector and Revenue to Market Cap
  d) Eigenvector and revenue category
  e) Eigenvector and Market Cap category
  f) Eigenvector and Asset category
  g) Eigenvector and industry category
  h) Eigenvector and ticker
  i) Eigenvector dependence between price and q-alpha implied volatility

5) Residual_Analysis (matlab):

- Purpose: Script calls functions that perform residual analyses and generate plots from the results.

- Inputs: Residual data from the PCA analysis.

- Outputs: Diagnostic plots.

6) PCA_Decomposition (matlab):

- Purpose: Function decomposes the variables in an observation space into systematic components and an idiosyncratic (uncorrelated) component using Principal Component Analysis (PCA). The total number of eigenvectors is used to decompose the space unless a limited explained variance is specified. This optional specification is described below.

- Input:

A This is an m-by-n matrix that contains the "m" observations for the "n" variables. Each column of this input matrix is a time series of a variable in the observation space

[jExplVarType] This flag specifies what option is used to determine the number of eigenvalues to use in the decomposition

0: Use all the e-values

1: Use a specified fraction of the total number of e-vlaues

2: Use the number of eigenvalues that account for a specified percentage of the area under the normalized e-value plot

3: Use the e-values that fall within a specified number of standard deviations from the largest e-value

[explVarFactor] - Optional input used with "jExplVarType", described above

0.01 - 1 : For option 1, 0.01 - 1 : For option 2, > 0 : For option 3

[iplot] - Diagnostic plotting flag.Set iplot > 0 to produce plots

- Output: eVectors, lambda, nEvecs, beta, residual, riskFactor, mnPortfolio (Market neutral portfolios)

7) GetIntegralTimeScale (matlab):

- Purpose: Function calculates the integral time scale, Taylor time scale, time to zero crossing, and autocovariance of time series.
- Input:

time: Time array (optional input). If data not present, code creates a linear time array.

y: Time series

iplot: Flag to generate plot (iplot =1)

- Output: Integral time scale, Taylor time scale, time to zero crossing, and autocovariance of time series.

8) CalcDistrFit (matlab):

- Purpose: This function calculates the goodness-of-fit of a sample data to a specified distribution. Default is the standard normal distribution.
- Input:

  x: Observed sample data

- Output:

  regrCoef: regression goodness-of-fit coefficient

  CDFx: Observed probability distribution

  CDFD: Theoretical probability distribution