Analysis of Causality in Stock Market Data

Chathra Hendahewa, Vladimir Pavlovic
Department of Computer Science
Rutgers University
Piscataway, NJ, USA
Email: {chathrah, vladimir}@cs.rutgers.edu

***Abstract*—Analyzing the changes in volatility is an important aspect in financial data analysis leading to effective estimation of risk and discovering underlying causes of such changes. While there is a rich literature in estimating implied and stochastic volatility in financial time series using traditional econometric methods, the application of machine learning methods such as sparse regression with temporal smoothness constraints is still in its infancy. In this paper, we propose a sparse, smooth regularized regression model to infer the volatility of the data while explicitly accounting for dependencies between different companies. Using real stock market data, we construct dynamic time varying graphs for different sectors of companies to further analyze how the volatility dependency between companies within sectors vary over time. We also show how our model captures the fluctuations in volatility over different economic conditions such as financial crisis periods. Further, based on these regression estimates we show how the proposed model assists in discovering useful correlations with external factors such as oil price, inflation, S&P500 index and also with various domestic trend indices.**

***Keywords*-Sparse Regression; Temporal Causal Graphs; Stock Market Analysis**

I. Introduction

Volatility based modeling has been widely used in financial data analysis since volatility is a major indicator of market stability. During periods of financial turbulence, the volatility of the stocks rises and thus indicates an anomaly to the common behavior. Analysis of causality between different stocks has also played an important role in identifying major market players and finding the cause and effect relationships among stocks, leading to insights in market dependency structure. Observing how these causal structures change over time could offer an alternative view of the underlying dynamics of stock price movements.

Tests of Granger causality [1], [2] have been used as a standard in analyzing linear systems of time series. Given two time series, Granger causality asserts whether one of the sequences is helpful in predicting the other by comparing the variance of the residuals of the linear predictive model. This model is based on the principle of time series forecasting that, cause must precede its effect in time. Later, it was also brought in to light that analyzing second order causality or causality in variance is more applicable to understanding the dynamics of the financial time series.The introduction of Auto Regressive Conditional Heteroskedasticity (ARCH) models by Engle in1982 [3] became pre-dominant in modeling such variance in financial time series. There have been various tests for causality such as Cross Correlation Function (CCF) Test, which estimates the cross-correlations between standardized (squared) residuals estimated for Generalized Auto Regressive Conditional Heteroskedasticity (GARCH) models [4]. When moving from univariate to multivariate GARCH models the number of parameters to be learnt begins to increase thus adding to the complexity of the estimation. In order to mitigate this problem, there have been attempts to develop regularization techniques based on penalized likelihood, which achieve sparse parameter estimation for multivariate GARCH model [5]. While these sparse models produce more succinct representation of dependencies they typically do not take into account their global structure. To address this problem, network-based methods such as Exponential Random Graph Models with smoothness constraints [6] and hidden Markov Random Field regression framework to model causal graph structures between multiple multivariate time series [7] have been proposed.

In this paper, we propose to infer causality of the financial time series using a sparse linear regression model with temporal smoothness, which can find the relationships between multiple variables based on the estimated volatility of the time series. By conducting this analysis over a period of time, we can discover how the causality structures change and hypothesize and assert their dependence on external events such as periods of financial crisis, economic booms, changes in sources such as oil price fluctuation and other economic indices. Although our proposed model also estimates the sparse regression coefficients based on the GARCH volatilities as proposed by Sun and Lin [5], unlike them we seek to induce temporal smoothness of sparse dependency estimates, while removing within-company local trends. We also analyze how the estimated causal structures relate to standard sector categories (e.g., Consumer Durables, Financials, Information technology) based on the US Global Industry Classification Standard, often used in financial decision making processes.

The rest of the paper is organized as follows. Section 2 provides the methods and learning framework of our model with GARCH based modeling and sparse regression based

causality analysis. In addition we describe the dynamic graph generation over time based on the regression estimates, which leads to analysis of causality structure changes within standard sectors. Section 3 describes the experimental settings and the results of our proposed model for real world stock market data on US based companies. Finally, we conclude our paper by discussing the advantages of our proposed model and summarizing the findings of causality from the real stock market data based analysis.

II. Causality Modeling

We propose to use a traditional Autoregressive Conditional Heteroskedasticity Model with a modification to its noise generating process combined with sparse regression to model the causality between different time series as explained in detail in the following sections.

A. Modified Auto Regressive Conditional Heteroskedasticity Model

In traditional econometric models, conditioning is often stated as regressions of the future values of the variables on the present and past values of the same variable. Using this generic model for autoregressive processes, the future value of a variable is regressed on the past and present values of the same variable where the noise is considered as Gaussian noise. Then the errors are modeled as a stochastic process where the variance is forecast as a moving average of past error terms.

$$x^w_{k,t} = \sum_{i=1}^{p} a^w_{k,i} x^w_{k,t-i} + \epsilon^w_{k,t}$$

$$where\ \epsilon^w_{k,t} \sim N(0, (\sigma^w_{k,t})^2),\ (\sigma^w_{k,t})^2 = \sum_{j=1}^{q} b^w_{k,j} (\epsilon^w_{k.t-j})^2 \tag{1}$$

In equation 1, $x^w_{k,t}$ refers to the normalized log return at time point t for a stock for company k inside a time window w. Time window w is the sliding interval $[t_s, t_s + \Delta T - 1]$ where $t_s = 1 + (w-1) * T_s,\ w = 0, 1, 2, ..., W$, ΔT is the window length and T_s is the offset step between successive windows. We typically set ΔT to span one month and choose the offset $T_s = 1$. The autoregressive terms are defined as $a^w_{k,i}$ for each lag i and each past observation of the normalized log returns for company k within that time window w is defined as $x_{k,t-i}$. An important aspect of model (1) is the fact that the noise variance $(\sigma^w_{k,t})^2$ is not stationary but, being modeled as the moving average of all errors, $\epsilon^w_{k,t}$ from the model for each company k for time window w.

Instead of the above formulation for modeling noise variance, we want to estimate the noise variance of a company to be dependent not only on its own past volatilities, but also on past volatilities of the other companies. The intuition behind this modification is to find how well the volatility of a single company could be explained by the volatility of other companies combined together. Therefore, we extend the volatility term estimated in model (1) to depend on other companies volatilities as shown in equation 2.

$$(\sigma^w_{k,t})^2 = \sum_j b^w_{kj} (\epsilon^w_{k,t-j})^2 + \sum_{l \neq k} \sum_j b^w_{lj} (\epsilon^w_{k,l,t-j})^2 \tag{2}$$

$$(\sigma^w_{k,t})^2 = \boldsymbol{\beta}^{w\top}_k . \boldsymbol{V}^w_k \tag{3}$$

Equation 3, is the same formulation of equation 2 expressed with a concise set of variables, where $\boldsymbol{\beta}^w_k$ corresponds to the the vector of all b^w_{kj} and b^w_{lj} coefficient terms and $\boldsymbol{V}^w_k$ represents the concatenated error terms defined in equation 2. Then we use sparse regression based method which is explained in the section which follows, to model these volatility based dependencies.

B. Sparse Linear Regression with smoothness constraint

Traditional Granger causality estimates the dependency between a pair of time series using linear regression although it does not facilitate sparsity in the dependency structure. Another way to find the dependency between different data points is to use sparse linear regression. Linear regression is an approach to modeling the relationship between a scalar dependent variable and one or more explanatory variables, which in this case are one stock's volatility and all other stocks' volatilities respectively. By introducing the sparsity constraint as in Lasso regression [8], we are ensuring that the coefficients of some of the explanatory variables are turned to zero, resulting in a sparse estimation. We are also imposing a smoothness constraint which ensures that the changes in coefficients between consecutive time windows is minimal, which enables smooth transition between consecutive time windows of estimation between w and $w-1$.

The constrained optimization problem of the training model for the sparse regression with smoothness constraints is shown in 4. The following optimization process has to be repeated independently for each company $k \in S$ in order to learn the total dependency structure for each company's volatility based on the other companies' volatilities.

$$\begin{aligned} &\forall_{k \in S} \{ min_{\beta^w_k} ||(\sigma^w_{k,t})^2 - \boldsymbol{\beta}^w_k (V^w_k)||^2_2 + \\ &(\lambda_{sp})^w_k |\boldsymbol{\beta}^w_k|_1 + (\lambda_{sm})^w_k ||\boldsymbol{\beta}^{w-1}_k - \boldsymbol{\beta}^w_k||^2_2 \} \\ &subject\ to\ (\lambda_{sp})^w_k \geq 0,\ (\lambda_{sm})^w_k \geq 0 \end{aligned} \tag{4}$$

C. Estimation and Learning

To estimate and learn the $\boldsymbol{\beta}^w_k$ coefficients for each company k for time window w specified in equation 4, Algorithm 1 is used and it defines how we tune the regularization terms $(\lambda_{sp})^w$ and $(\lambda_{sm})^w$ that corresponds to sparsity regularization term and smoothness regularization term respectively. This learning process has to be repeated independently for each company $k \in S$ to learn the total model for all companies in the data set, S. It should be noted that in

order to avoid over-fitting, we use cross validation. For a given window w, we separate the data points in to n equal sub windows and use 75% of the sub windows as training data and 25% of the sub windows as validation data.

Algorithm 1 Estimation and Learning

1: **for** $w = 1$ to W **do**
2: **if** $w == 1$ **then**
3: Set $(\boldsymbol{\Lambda_{sp}})_k^w = [\lambda_1, \lambda_2, .., .., .., .., \lambda_m]$
4: **else if** $w == 2$ **then**
5: Set $(\boldsymbol{\Lambda_{sm}})_k^w = [\gamma_1, \gamma_2, .., .., .., .., \gamma_n]$
6: **else**
7: Set $(\boldsymbol{\Lambda_{sp}})_k^w = [(\lambda_{sp})_k^{(w-1)*} \pm \Delta d]$
8: Set $(\boldsymbol{\Lambda_{sm}})_k^w = [(\lambda_{sm})_k^{(w-1)*} \pm \Delta e]$
9: **end if**
10: **for** $m = 1$ to $length(\boldsymbol{\Lambda_{sp}})_k^w$ **do**
11: **for** $n = 1$ to $length(\boldsymbol{\Lambda_{sm}})_k^w$ **do**
12: Estimate $\boldsymbol{\beta}_{k,(m,n)}^w$ for all possible $(\lambda_{sp})_k^w$ and $(\lambda_{sm})_k^w$ based on the training sub windows $||(\sigma_{k,t}^w)^2 - \boldsymbol{\beta}_{k,(m,n)}^w.(\boldsymbol{V}_k^w)||_2^2 + ((\boldsymbol{\Lambda_{sp}})_k^w[m])|\boldsymbol{\beta}_{k,(m,n)}^w|_1 + ((\boldsymbol{\Lambda_{sm}})_k^w[n])||\boldsymbol{\beta}_k^{w-1} - \boldsymbol{\beta}_{k,(m,n)}^w||_2^2$
13: **end for**
14: **end for**
15: For the validation sub window inside large window w estimate the best $(\boldsymbol{\beta}_{k,(m,n)}^w)^*$ which satisfies $\min_{(\boldsymbol{\beta}_{k,(m,n)}^w)^*}\{||(\sigma_{k,t}^w)^2 - \boldsymbol{\beta}_{k,(m,n)}^w.(\boldsymbol{V}_k^w)||_2^2 + ((\boldsymbol{\Lambda_{sp}})_k^w[m])|\boldsymbol{\beta}_{k,(m,n)}^w|_1 + ((\boldsymbol{\Lambda_{sm}})_k^w[n])||\boldsymbol{\beta}_k^{w-1} - \boldsymbol{\beta}_{k,(m,n)}^w||_2^2\}$
16: Set $\boldsymbol{\beta}_k^w = (\boldsymbol{\beta}_{k,(m,n)}^w)^*$
17: Set $(\lambda_{sp})_k^w = (\boldsymbol{\Lambda_{sp}})_k^w[m])^*$
18: Set $(\lambda_{sm})_k^w = (\boldsymbol{\Lambda_{sm}})_k^w[n])^*$
19: **end for**

Initially, we set the $(\boldsymbol{\Lambda}_{sp})_k^w$ and $(\boldsymbol{\Lambda_{sm}})_k^w$ as a vector of possible equidistant $(\lambda_{sp}^w)_k$ sparsity regularization parameter values and $(\lambda_{sm}^w)_k$ smoothness regularization parameter values respectively as shown in lines 2 - 5 in the algorithm 1. Then we estimate the corresponding $\boldsymbol{\beta}_{k,(m,n)}^w$ for each window w for each company k for each pair of sparseness and smoothness values for the training sub windows using the proposed model, as shown in lines 10-14. Then using the validation sub window we find which pair of sparseness and smoothness regularization values best fit the validation data which minimizing the residual of the proposed model as depicted in line 5. Then we set the current best estimated coefficients from the proposed model and the best pair of regularization parameters as the final estimations for the current window w (lines 16-18 in Algorithm 1). Although the learning is independently done for each company k for each window w, we need to ensure that smoothness is preserved in the estimated coefficients from the sparse regression model within consecutive windows w and $w-1$ for each company k. In order to facilitate this, we ensure the range of the best smoothness regularization term estimated at the previous window $(\lambda_{sm}^{w-1})^*$ is close to the smoothness regularization term estimated for the current window $(\lambda_{sm}^w)^*$, by only varying the latter value centered around the previous value in close proximity as shown in line 8 of algorithm 1. We are also ensuring that the estimated coefficients in consecutive time windows have a close sparsity constraint by varying the best estimated $(\lambda_{sp}^{w-1})^*$ at the previous window only by a small margin as shown in line 7 of the Algorithm 1. We continue this estimation and learning model until we have covered all the windows, $w \in W$ for each company k.

D. Dynamic Graph Construction

Once we estimate the lasso coefficients for each time window for each company $\boldsymbol{\beta}_k^w$, we use those to construct the adjacency matrix for all companies for each window w, $\boldsymbol{A}_{k\times k}^w$, where each entry in the matrix corresponds to $\boldsymbol{A}_{k,*}^w = |\boldsymbol{\beta}_k^w|$. For each time window, the graph nodes represent the companies and the edges between two nodes represent the magnitude of the coefficients estimated by the above model. The directionality of the edges correspond to which company's volatility is influencing the other company's volatility. All the coefficient weights with magnitude $\geq \eta$ (where η is typically set to 10^{-5}) are considered as edge weights in the graph and any coefficient value below this threshold is set to zero.

III. Experimental Results

In this section, we provide a detailed experimental analysis of real world stock market data using the proposed model.

A. Data

We use the historical stock market data for 990 publicly listed companies from the US Stock Exchange. For this analysis we are focusing on historical log returns of the stocks for the period of November 2006 to August 2009. We also obtain US Global Industry Classification Standard (US GICS) to categorize the companies into different sectors in order to provide sector wide analysis. The US GICS sectors used in this analysis are Consumer Durables (CD), Consumer Staples (CS), Energy (EN), Financials (FN), Health Care (HC), Industrials (IN), Information Technology (IT), Materials (MT), Telecommunications (TS) and Utilities (UT).

B. Within Sector Analysis

In this analysis, we perform dependency analysis within each sector based on the US GICS categorization. Then we construct month-long windows of daily log returns of each company. Using a sliding window approach, the time windows are lagged by 1 trading day at each iteration. For each company a GARCH model is estimated for each window and the volatilities are estimated. Then we model the dependency between each stock's volatility with respect to other stocks'

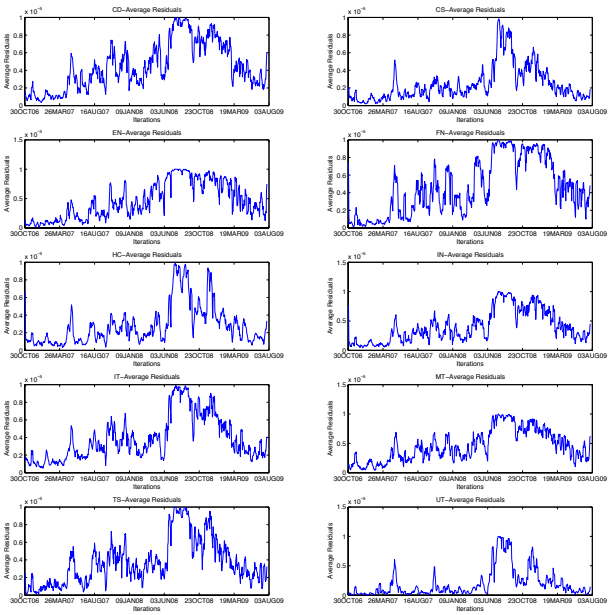

Figure 1. The sum of estimated residuals R_k^w for all companies within each sector over the time horizon of Nov 2006 - Aug 2009

volatility within the same sector using the proposed sparse, smooth regression model. We use sparse regression weights to construct time-varying dependency graphs, which we use to visualize the changing interactions among companies. In order to estimate the proposed sparse, smooth regression model, we used the SLEP: Sparse Learning Package [9].

To avoid over-fitting, we use cross validation in our learning process as mentioned in section 2 under the learning algorithm. For a given window w (a monthly-long window comprising of volatilities estimated for each trading day within that month), we separate the data points in to 4 equal sub windows, where each sub window corresponds to one week's volatilities estimated from the GARCH model. Then we use the first 3 sub windows as training data and the final sub window as the validation data. We set initial $(\mathbf{\Lambda}_{sp})_k^w = [0.001, ..., 0.01]$, 50 equally spaced values and at each subsequent window set the $(\lambda_{sp})_k^w$ to 5 equally spaced values centered around the best $(\lambda_{sp})_k^{w-1}$ estimated from the previous iteration. The same process is followed for smoothness parameter, $(\lambda_{sm})_k^w$ after the second time window. This ensures that our smoothness regularization term and the sparsity regularization term are smoothly changed within consecutive time windows (as shown in lines 7 and 8 in algorithm 1). Residual value in volatility estimation of company k, R_k^w is defined as the predictive error of that company's volatility in terms of other companies' volatilities, $R_k^w = ||(\sigma_{k,t}^w)^2 - \boldsymbol{\beta}_k^w(V_k^w)||_2^2$. These residual values provide an insight to how and when our proposed model finds it hard to estimate the dependency between each company's volatility between each sector. After summing up the residual values estimated for all the companies from each sector for each time point, we analyzed those plots as shown in Figure 1. We noticed across all sectors that within the period of June 2008 to January 2009, the residuals of the estimation was higher compared to other months during the period of late 2006 to late 2009. This phenomena is inline with the high fluctuations of volatilities in the stock market due to the financial crisis that occurred in the US in mid 2008. This increase in residual values estimated based on our proposed model, show that within such periods of economic turmoil, it is hard to predict the volatility of a company given other companies volatilities which is inline with the stock market behavior at that period. Figure 2 shows directed causal graphs estimated for all 10 sectors over the full 3 year period. The graph edges, $E(k,l)$ correspond to the total influence between two companies, k and l over the entire time span $w \in W$, i.e. $E(k,l) = \sum_w |\beta_{k,l}^w|$. Darker edges indicate high weights/dependencies between those node pairs whereas, lighter edges suggest weaker interactions. The size of node corresponds to the sum of all estimated residual values over the entire 3 years. The directionality represent the influence from node to node. It can be observed that the estimated graphs are sparse and certain companies in the sector are highly influenced by a typically small subset of the other companies in each sector.

C. Influence of External Factors

It is also interesting to analyze how the external economic factors such as the inflation rate, oil price affect the volatility of the stock market. Our analysis of within sector residual graphs in the previous section shown in Figure 1 highlighted their co-occurence with events such as the crisis in June 2008 to January 2009. This gives us an indication that this change in estimation might be due to other external influences that drive the stock market volatility to be higher within this period of economic downturn. Therefore, we collected daily/monthly data for oil prices, US inflation rate and S&P 500 index as external economic factors for the period in concern from 2006-2009. We also analyze how the search trends based on Google domestic trend indices (www.google.com/finance/domestic_trends) would have an impact on our estimations.

We summed up the residual values estimated from our proposed model for each month and whitened the data. We applied the same process to external factors related data to make both sequences comparable (since some of the external factors were available as monthly data). Since the external factor time sequences and the volatility estimated time sequences are not aligned, we aligned them using Dynamic Time Warping (DTW) method. We then calculated the correlation between each pair of aligned sequences. Figure 3 shows the estimated correlation weights for each corresponding pair of sector volatilities vs. the external factors in consideration. We find that all the sectors' volatilities are highly positively correlated with the oil prices and the US inflation rate, while most of the sectors show low or

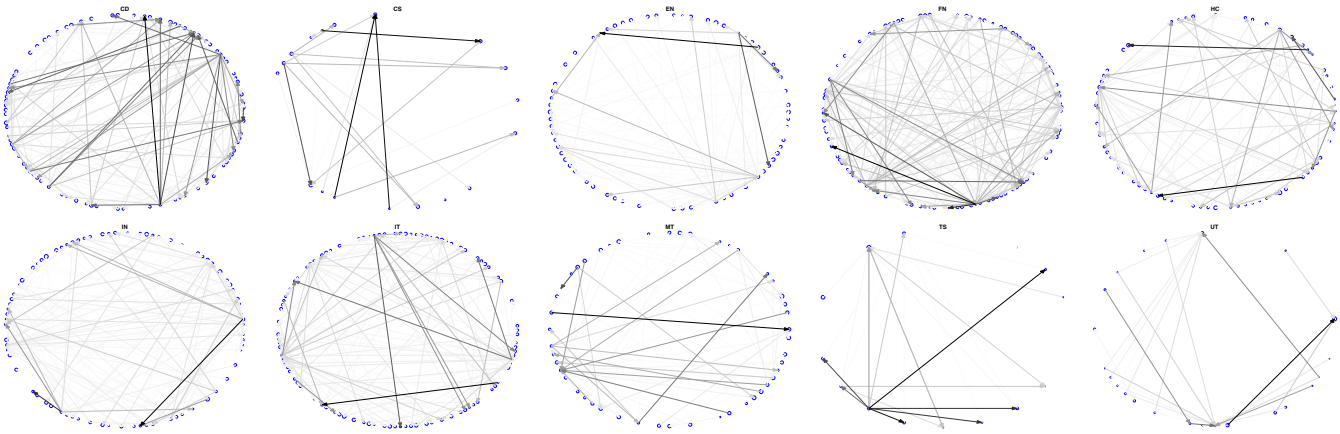

Figure 2. All sector residual value based graphs - First row (from left to right): Consumer Durables (CD), Consumer Staples (CS), Energy (EN), Financials (FN), Health Care (HC). Second row (from left to right): Industrials (IN), Information Technology (IT), Materials (MT), Telecommunications (TS), Utilities(UT).

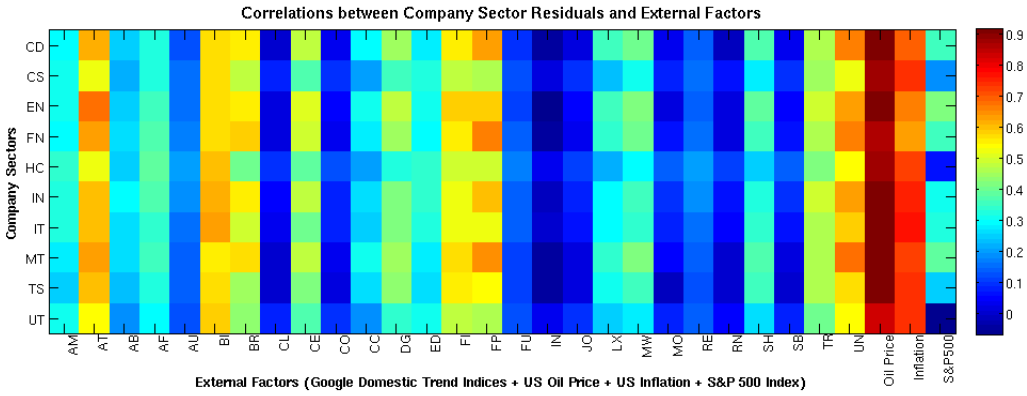

Figure 3. Correlation Coefficient between each sector's residuals estimated from the proposed model (in y-axis) and each external factor(27 Google Domestic trend indices, Oil price, US Inflation rate and S&P 500 index values are along x-axis)

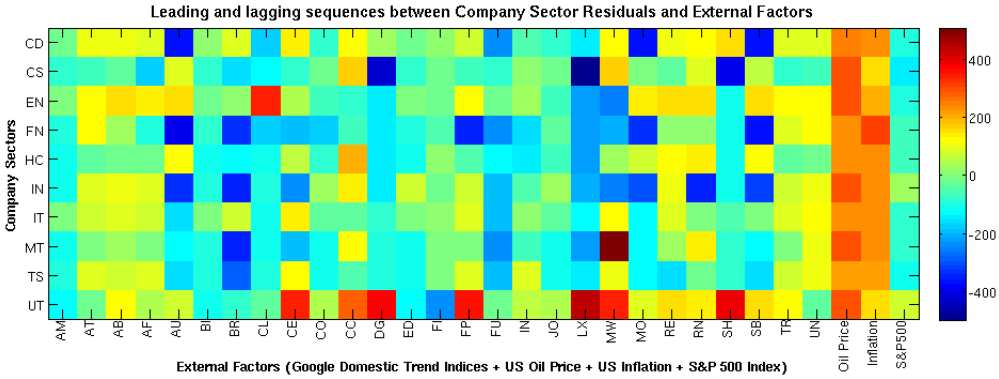

Figure 4. Leading and Lagging sequences for each sector's residuals estimated from the proposed model (in y-axis) and each external factor(27 Google Domestic trend indices, Oil price, US Inflation rate and S&P 500 index values are along x-axis). Here the positive values depicted by the range of colors from yellow to red means that the sector's volatility residual sequences are lagging and the external factor's sequences are leading while the negative values shown in colors ranging from blue to green show that the external factor sequences are lagging the sector residual sequences are leading.

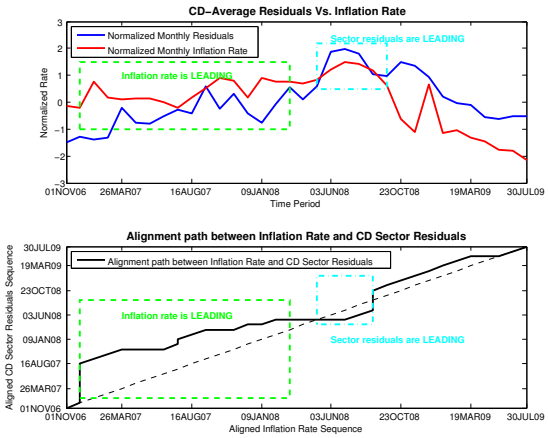

Figure 5. Top: Monthly Residuals for sector CD and Inflation rate, Bottom: Alignment path from DTW for the two sequences.

moderate positive correlation between S&P 500 index. Since inflation rate and oil prices are major economic indicators that affect stock market behavior it can be argued that when the rates of inflation and oil prices rise, the volatility of the stock market also increases since most of the companies operations would be influenced by those factors. Major constituents of S&P 500 index are coming from sectors such as Consumer Staples (CS), Energy (EN), Financials (FN) and Information Technology (IT) for 500 selected companies. We can observe that our volatility fluctuations are moderately positively correlated with S&P 500 index in sectors such as Energy (EN) and Consumer Durables (CD). When considering the correlations with Google's domestic trend indices, we observe that the trends such as Airline Travel (AT), Business and Industrials (BI), Financial Planning (FP) and Unemployment (UN) are highly correlated with most of the sector volatilities. This shows that people tend to search for terms belonging to those trend categories online, which correlate with the economic condition of the country within a period where the volatility of the stock market is high. This implies that when the domestic trend indices in the areas such as those mentioned above are searched more frequently, that particular time period would be followed by a stock market instability and higher volatility.

In addition to asserting the level of correlation between within sector dependencies and external indicators, it is also important to discover the temporal causal relationships among those factors. By analyzing whether an external factor precedes or lags the sector behavior we can assert the direction of causal dependency. For example, when we look at the normalized monthly inflation rate and the normalized monthly sector residuals we can observe which sequence is leading and how the leading/lagging sequences change over time. Such an example of leading/lagging sequences are illustrated in Figure 5 for one sector, Consumer Durables (CD) and the monthly inflation rate, along with the alignment path after using DTW method. Further, we provide analysis of leading and lagging sequences based on estimating the total distance from the DTW aligned paths to the diagonal path for each pair of sequences and illustrate this in Figure 4. The red squares indicate positive sum of distances between volatility of a sector and the corresponding external factor alignment path which implies that the volatility of sectors are lagging and the external factors are leading. Areas with blue squares imply that the volatility of sectors are leading and the external factors are lagging. Discovering that oil price, inflation rate and unemployment (UN) are leading the volatility sequences for almost all sectors show that these economic indicators and trends would be useful in predicting that the volatility of stock market is going to rise in the future.

IV. Conclusion

We have proposed a sparse, smooth regression framework and learning algorithm to estimate the dependency structure between volatility of a time series model. We have applied this framework to real world stock market data and have shown that one could discover meaningful insights about the volatility structure and how it changes over time for different sectors of companies. We have also analyzed the influence of external economic factors and user search query trends on the stock market behavior.

References

[1] C. Granger, "Testing for causality: A personal view- point," *Journal of Economic Dynamics and Control*, vol. 2, pp. 329–352, 1980.

[2] C. Granger., "Investigating causal relations by econometric models and cross-spectral method," *Econometrica*, vol. 37, 1969.

[3] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation," *Econometrica*, vol. 50, pp. 987–1007, 1982.

[4] C. M. Hafner and H. Herwartz, "Testing for causality in variance using multivariate garch models," *Econometric Institute Report EI*, 2004.

[5] Y. Sun. and X. Lin., "A sparse loading full-factor multivariate garch model," *S.Co*, 2009.

[6] M. Kolar., L. Song., A. Ahmed, and E. P. Xing, "Estimating time-varying networks," *Annals of Applied Statistics*, vol. 4, pp. 94–123, 2010.

[7] Y. Liu, A. Niculescu-Mizil, A. Lozano, and Y. Lu, "Learning temporal causal graphs for relational time-series analysis," *Proceedings of 27th the International Conference on Machine learning*, 2010.

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 58, pp. 267–288, 1996.

[9] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University, 2009. [Online]. Available: http://www.public.asu.edu/~jye02/Software/SLEP