Identifying and Predicting Market Reactions to Information Shocks in Commodity Markets

CS229 Project Report, Fall 2014

Eric Liu[†], Vedant Ahluwalia[‡], Deepyaman Datta[*], Dongyang Zhang[⁞]

[†] Computational and Mathematical Engineering, Stanford University, E-mail: ericql@stanford.edu
[‡] Computational and Mathematical Engineering, Stanford University, E-mail: vahluwal@stanford.edu
[*] Computer Science, Stanford University, E-mail: deepyaman.datta@utexas.edu
[⁞] NVIDIA, E-mail: dongyangz@nvidia.com

Abstract

This project proposes a three-stage time series model to identify the relationship between properties of news information shocks and patterns in market reactions to such news. Then, given a specific news update, the model predicts how market players will subsequently respond. We apply multivariate time series segmentation and clustering techniques on gold commodity futures, and then run various multi-class classification algorithms on relevant news articles.

1. Introduction

There exists ample evidence to suggest that financial market players often respond irrationally to news information. [1] For example, investors habitually overreact to adverse environmental and social news, resulting in an immediate negative return and a long-term trend reversal.[2] Therefore, exploring how different categories of news articles affect market behaviours in commodity prices may result in valuable findings.

1.1 Current Theory

The concept of leveraging news data to predict commodity price fluctuations has been extensively explored.[3] However, most current research exhibit at least one of the following two properties: the models tend to simplify the classification process either by dichotomizing the effect of news articles as only "good" or "bad," [4] or by pre-determining the categories under which the topics of these articles must fall.[5]

1.2 Our Approach

Instead of arbitrarily selecting the number of labels and dictating which news topics are relevant as input features, this project attempts to generate labels and features for the news data by first fitting models to the time series of market data. The entire model training process has three successive stages, which is depicted in Figure 1 below.

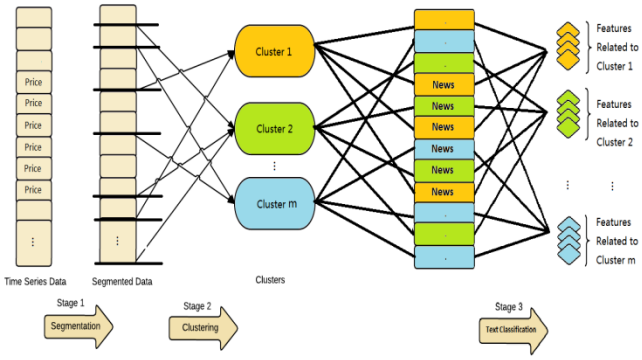

Figure 1: The Three Staged Model for Gold Price Prediction

In stage 1, we adopt a time series segmentation process that decomposes the time series data into time intervals where, within each interval, the data demonstrate similar behaviours. The endpoints of these time intervals are considered as structural breakpoints, which can hence serve as indicators to when specific news information shocks have occurred.

In stage 2, we apply various time series clustering techniques to categorize a number of distinct market behaviours following the structural breakpoints. This process replaces the classical "good" or "bad" labelling approach, and results in a more advanced multi-class labelling strategy.

In stage 3, we employ the usual text mining methods to generate relevant features for the news articles, including sentiment analysis, profile of mood states, bag-of words models, and LDA. Then, SVM and other supervised learning methods are applied to characterize rules of association between news articles features and clusters of market reactions.

[1] Bondt & Thaler (1985), Owen (2002), Ma, Tang, & Hasan (2005).
[2] Länsilahti (2012).
[3] Roache & Rossi (2010), Kilian & Vega (2011).
[4] Maheu & McCurdy (2004), Gidofalvi & Elkan (2001).
[5] Fang & Peress (2009), Schumaker & Chen (2009).

2. Models

In the following sections, both the underlying theory and relevant metrics of the model applied in each stage are explained in greater detail.

2.1 Stage 1 - Time Series Segmentation

The objective of time series segmentation is to identify time intervals that dissect the dataset into homogenous sections. Given a multivariate time series dataset $\{x_i\} \in \mathbb{R}^p, 1 \leq i \leq n$. Suppose we fix the number of segments desired, D, and wish to find the time intervals $S_d = [i_{d-1}, i_d], 1 \leq d \leq D$, with $i_0 = 1$, and $i_D = n$ (i_d 's are referred to as structural breakpoints). We may choose the time intervals that minimize the following cost function

$$C_D(S) = \sum_{d=1}^{D} \left(\frac{1}{i_d - i_{d-1} + 1} \sum_{i \in S_d} dist(f_d(x_i), x_i) \right)$$

where $\mathrm{S} = \{\mathrm{S_d}\}_{1 \leq \mathrm{d} \leq \mathrm{D}}$ is a specific segmentation of the time series with D segments, $f_d(x_i)$: $\mathbb{R}^p \to \mathbb{R}^p$ denotes a fitted model based on the data points $\{x_i\}_{i \in S_d}$, and $dist(a, b)$: $\mathbb{R}^{p \times p} \to \mathbb{R}$ is some measure of distance.

The cost function $C_D(S)$ is a weighted average of each segment's error residuals, with weights proportional to the length of each segment. Theoretically, by methods of dynamic programming, one may obtain the solutions D^* and S^* to the global minimum cost. However, due to the impracticality of this approach with large datasets, we adopt an iterative top-down greedy algorithm that solves the optimal segmentation S at each iteration. The algorithm is outlined below and further explained in Bankó (2011).

Initially, we set $D = 1$ and fit only one model across the entire dataset. Then, at each iteration step k, we set $D = k + 1$, and intend to select an additional structural breakpoint i_{d_k}. To find the optimal k^{th} breakpoint, we fit models f_d on each segment S_d, $1 \leq d \leq k$, for every possible value of i_{d_k}, and choose the breakpoint that gives the most cost reduction. The iterations terminate either when we reach the pre-determined number of segments D_{target}, or when the benefit of including an additional structural breakpoint falls below some threshold b_{limit}.

2.2 Stage 2 – Time Series Clustering

In this stage, we wish to categorize the how market reacts after each structural point i_d. For each date i_d, we generate a higher dimensional vector

$$y_{i_d}^T = \left[x_{i_d}^T, x_{i_d+1}^T, \ldots, x_{i_d+r}^T\right] \in \mathbb{R}^{\mathrm{p(r+1)}}$$

by appending $x_{i_d}^T$ the market behaviour from the date of the structural breakpoint till r days later, $x_{i_d+r}^T$. Then, we follow a wavelet based multivariate time series clustering method as proposed by D'Urso (2012). The main advantage of applying the wavelet based technique is that the approach does not require stationarity in the data.

To determine the appropriate number of clusters, we use two diagnostic measures given by Nieto-Barajas and Contreras-Cristán (2014), namely, the heterogeneity measure (HM) and the logarithm of pseudo marginal likelihood ($LPML$).

Suppose the data has been grouped into m clusters, $\{G_1, \ldots, G_m\}$. HM computes the aggregate of variability within each cluster, and is given by

$$HM(G_1, \ldots, G_m) = \sum_{k=1}^{m} \left(\frac{2}{n_k - 1} \sum_{i<j \in G_k} \left\| x_i - x_j \right\|_2^2 \right)$$

where $n_k = |G_k|$ is the number of observations in each cluster G_k. Necessarily, if the clustering algorithm is robust, a model with higher number of clusters will have a lower HM value. Hence, m should be chosen such that both m and HM have relatively small magnitudes.

The $LPML$ measure is the sum of log-transforms of the conditional predictive ordinate (CPO) statistics for each x_i, which is given by

$$\widehat{LPML} = \sum_{i=1}^{m} \log\left(\widehat{CPO}_i\right)$$

$$= \sum_{i=1}^{m} log\left(\frac{1}{L} \sum_{l=1}^{L} \frac{1}{f\left(x_i \middle| \alpha^{(l)}, \gamma^{(l)}, \sigma_\epsilon^{(l)}\right)} \right)^{-1}$$

where the conditional likelihood is given as

$$f\left(x_i | \alpha^{(l)}, \gamma^{(l)}, \sigma_\epsilon^{(l)}\right) = \frac{\mathrm{b + am}}{\mathrm{b + n}} N(x_0 | \, Z\alpha_0^{(l)}, W_0)$$

$$+ \sum_{j=1}^{m} \frac{n_j^* - a}{b+n} N(\mathrm{x_0} | \, Z\alpha_0^{(l)} + \mathrm{X}\beta_\mathrm{j}^{*(l)} + \theta_\mathrm{j}^{*(l)}, \sigma_{\epsilon_0}^{2(l)} I).$$

In the above equation, the estimated coefficients arise from an assumed Bayesian model with a Poisson-Dirichlet process prior, with a general sampling model that follows the distribution

$$x_i = Z\alpha_i + X\beta_i + \theta_i + \epsilon_i \ , 1 \leq i \leq n$$

with $\theta_i = \rho\theta_{i-1} + \nu_i$, which exhibits an autoregressive behaviour, $\epsilon_i \sim N\left(0, \sigma_{\epsilon_i}^2\right)$, and $\nu_i \sim N\left(0, \sigma_\theta^2\right)$.

Technical details of the derivations of these posterior conditional likelihood estimates and other model specifications are given in Nieto-Barajas (2014). Intuitively, $\widehat{CPO}$ is a Monte Carlo estimate of the conditional likelihood (hence the name "pseudo" conditional likelihood) for each observation x_i, and we choose the number of clusters m that maximizes the log likelihood of the entire dataset.

2.3 Stage 3 – Text Classification

From Stage 1, we obtain a set of dates of structural breakpoints in the time series data, which we assume the significant change in market behaviour is due to some news information shock. In stage 2, we categorize the market reactions to these information shocks into a smaller number of clusters. Now, we have made all necessary preparations for the ultimate task: discovering rules of associations between news articles and these market reactions.

First, we assign a label $l_j \in \{0,1,\dots,\mathrm{m}\}$ to each news article ω_j. If the date of news ω_j, i, is not one of the structural breakpoints i_d, we assign the label $l_j = 0$. Otherwise, we assign $l_j = k$, where the cluster G_k contains the observation x_i. In words, we assign each news article the label of the market reacted (or did not react) to the news information shocks (or a lack thereof) on the day when the article was published.

Then, we apply to the collection of news article data two classical sentiment analysis techniques—Profile of Mood States (POMS) and Lydia Sentiment Analysis Systems (LSAS)—as outlined by Han (2012):

$$\mathrm{P}(t) \to m \in \mathbb{R}^6 = [\|w \cap p_1\|, \|w \cap p_2\|, \dots, \|w \cap p_6\|]$$

$$m_d = \frac{\sum_{\forall t \in T_d} \frac{m}{\|m\|}}{\|T_d\|}$$

$$\theta_{m_d}[i,k] = [m_i, m_{i+1}, \dots, m_{i+k}]$$

$$\overline{m}_i = \frac{\frac{m}{\|m\|} - \bar{x}(\theta[i,\pm k])}{\sigma(\theta[i,\pm k])}$$

where $\overline{m}_i$ is the normalized mood vector. Again, we refer the technical details to Han's paper and focus on the intuitive explanation. Both POMS and LSAS can be seen as extra features we generate for each news article, which gives a holistic sense of the significance of the entire article as opposed to individual words or stems within each article.

We use both generative and discriminative algorithms to uncover the most relevant features for each cluster label. The models we employed include: Multinomial Naïve Bayes, Nearest Centroids classifier, Linear SVC (with no regularization, L1 regularization, and L2 regularization), SGD Classifier, KNN classifier, Passive Aggressive classifier, and Ridge classifier. Finally, we select the most appropriate approach by comparing the F1 score and the confusion matrix for each model.

3. Data Collection

Daily prices of gold futures are collected from 2000 to 2014, including each day's high, low, opening, and closing price. For each day i, we capture the market behaviours by constructing a 3-dimensional vector, $x_i = [x_{1i}, \dots, x_{3i}]^T$ with the following components:

x_{1i} ~ opening price jump = $\frac{open_i - close_{i-1}}{close_{i-1}}$

x_{2i} ~ intra-day price movement = $\frac{close_i - open_i}{open_i}$

x_{3i} ~ intra-day volatility = $\frac{high_i - low_i}{last_i}$

and obtain a normalized vector $\hat{x}_i$ by dividing each component x_{ji} by its sample standard deviation, s_j. Below are graphs of daily gold prices and normalized intra-day price movements ($\hat{x}_{2i}$) from Jan. 3, 2001 to Mar. 23, 2014.

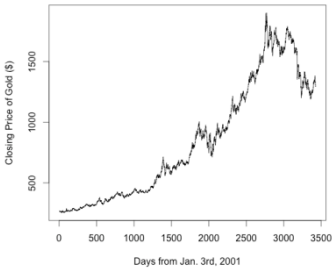

Fig 2: Historical Gold Prices

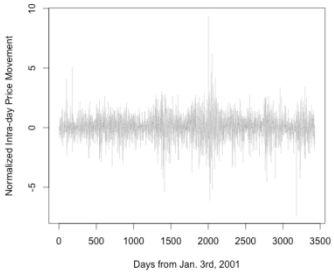

Fig 3: Normalized intra-day price movements

The news articles we collected include all news entries from the Wall Street Journal Online from 2010 to 2014, which amounts to 297,616 articles in 1,232 days (an average of 239.6 articles per day). The total number of features generated is 121,349, which is the size of the vocabulary list after removing stop-words and filtering with minimum and maximum document frequency.

4. Results

We applied time series segmentation on the normalized market behaviour vectors using the top-down algorithm. Choosing the Euclidean norm as our distance function $dist(a,b) = \|a-b\|_2^2$, we calculated the minimum cost $C_D(S)$ at each step $D = k$, and have reproduced some of the values in Table 1 below.

D	$C_D(S)$	$\Delta C_D(S)$
50	0.9586	$1.2206*10^{-3}$
100	0.9195	$0.8040*10^{-3}$
200	0.8776	$0.5192*10^{-3}$
300	0.8348	$0.4377*10^{-3}$
400	0.7954	$0.3936*10^{-3}$
500	0.7638	$0.3721*10^{-3}$

Table 1: Minimum Costs for Various Number of Segments

It is evident that $C_D(S)$ is negatively correlated with D. Because D is also the number of data points for Stage 2 clustering, we ought to choose some number of segments that is not too small as to undermine the accuracy of the clustering in Stage 2, but also not too large as to label structural breakpoints on days that exhibited no significant change. Having these two goals in mind, we choose the number of segments $D = 350$, because the change in minimum cost for each additional iteration, $\Delta C_D(S)$, seems to reach very close to 0 for $D > 350$. Figure 4 below shows the historical gold prices and the segments with $D = 350$.

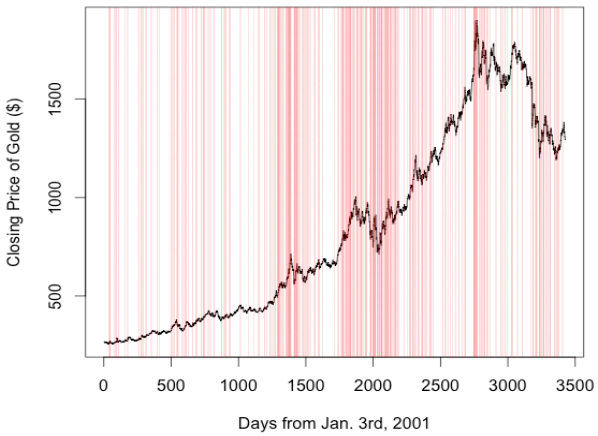

Figure 4: Gold Prices with 350 Segments

Next, we apply the wavelet based multivariate time series clustering with the appended vectors $\{y_i\}_{1\le i\le 350}$, with each y_i initiating from the structural breakpoint i_d. We alter the parameter m, which is the number of desired clusters, and calculate both the heterogeneity measure and the logarithmic pseudo marginal likelihood for each value of m. Table 2 below illustrates some selected results.

m	HM	$LPML$
2	5.00	7.3983
4	4.18	8.3793
6	3.32	10.0374
7	3.25	9.9652
8	3.24	8.2387
10	2.98	8.8162

Table 2: Number of Clusters and Diagnostic Measures

As expected, the HM decreases with the number of clusters, but we observe that after 7 clusters, each additional cluster does not reduce variance by a significant amount. The LPML measure also suggests the appropriate number of clusters is between 6 and 7. We choose 7 to be the number of clusters (and hence 8 labels) for the text classification problem in Stage 3.

After labelling each news article data with its corresponding cluster, we ran the text classification algorithms on the training set (every three observations from four consecutive data points). Then, we verify our trained model on the test set (every fourth data points) and compute the F1 scores, training time, and test time for each algorithm. The results are shown in Figure 5 below.

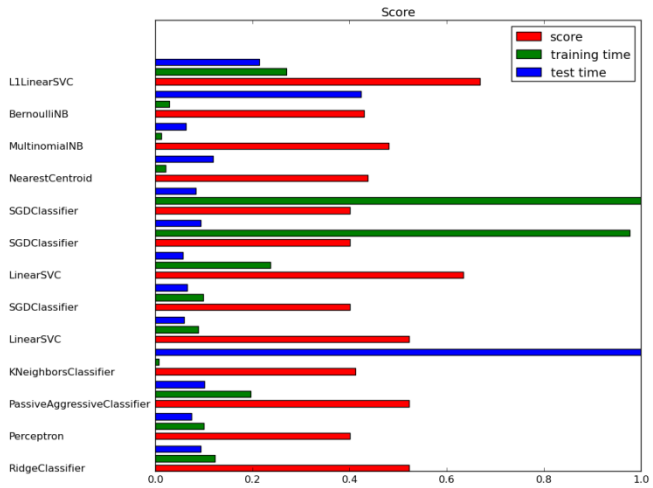

Figure 5: F1 Scores, Training Time, and Test Times

From the results, the linear SVC with L1 regularization had the highest F1 score, which is 0.67. It is worth noting that this number is considerably high, because our problem is multi-class (with 8 labels), and hence correct prediction is much more difficult than a usual two-class classification task.

5. Conclusion & Future Research

The main contribution of this three-stage model to the current scholarly discourse is that it provides a more data-driven approach in generating the labels for news articles. This methodology can be readily applied to other commodities as well as to stocks and equity derivatives.

Because the clustering process consists only of unsupervised learning techniques, some clusters may have no economic explanations, hence resulting in less significance when regressed against the news article data. Further improvements on the model could consider incorporating economic theories in a mixture model approach. Also, due to the presence of heteroskedasticity, exploring a GARCH framework may prove beneficial. Finally, instead of modeling each commodity individually, a DPCA analysis could be conducted to fit all commodities simultaneously, and hence take the effect of covariances between different commodities into consideration.

References

Argiento, R., Cremaschi, A. and Guglielmi, A. (2012). A bayesian nonparametric mixture model for cluster analysis. *Technical report Quaderno Imati CNR, 2012(3*).

Bankó, Z., Dobos, L., & Abonyi, J. (2011). Dynamic Principal Component Analysis in Multivariate Time-Series Segmentation. *Conservation, Information, Evolution-towards a sustainable engineering and economy, 1*(1), 11-24.

Bondt, W. F., & Thaler, R. (1985). Does the stock market overreact?. *The Journal of finance*, *40*(3), 793-805.

Boyd, J. H., Hu, J., & Jagannathan, R. (2005). The stock market's reaction to unemployment news: why bad news is usually good for stocks. *The Journal of Finance*, *60*(2), 649-672.

Chib, S., & Greenberg, E. (1996). Markov Chain Monte Carlo Simulation Methods in Econometrics. *Econometric Theory, 12*(03), 409-431.

D'Urso, P., & Maharaj, E. A. (2012). Wavelets-based clustering of multivariate time series. *Fuzzy Sets and Systems*, *193*, 33-61.

Fang, L., & Peress, J. (2009). Media coverage and the cross‐section of stock returns. *The Journal of Finance*, *64*(5), 2023-2052.

Fung, G. P. C., Yu, J. X., & Lu, H. (2005). The Predicting Power of Textual Information on Financial Markets. *IEEE Intelligent Informatics Bulletin*, *5*(1), 1-10.

Gidofalvi, G., & Elkan, C. (2001). Using news articles to predict stock price movements. *Department of Computer Science and Engineering, University of California, San Diego*.

Han, Z. (2012). Data and text mining of financial markets using news and social media (Unpublished doctoral Dissertation). University of Manchester, Manchester.

Ikenberry, D. L., & Ramnath, S. (2002). Underreaction to self‐selected news events: The Case of Stock Splits. *Review of Financial Studies*, *15*(2), 489-526.

Kaya, M. Y., & Karsligil, M. E. (2010). Stock price prediction using financial news articles. *Information and Financial Engineering (ICIFE), 2010 2nd IEEE International Conference,* 478-482.

Kilian, L., & Vega, C. (2011). Do energy prices respond to US macroeconomic news? A test of the hypothesis of predetermined energy prices. *Review of Economics and Statistics*, *93*(2), 660-671.

Länsilahti, S. (2012). Market reactions to Environmental, Social, and Governance (ESG)-news: evidence from European markets.

Ma, Y., Tang, A. P., & Hasan, T. (2005). The stock price overreaction effect: Evidence on Nasdaq stocks. *Quarterly Journal of Business and Economics*, 113-127.

Maheu, J. M., & McCurdy, T. H. (2004). News arrival, jump dynamics, and volatility components for individual stock returns. *The Journal of Finance*, *59*(2), 755-793.

Mukhopadhyay, S. and Gelfand, A.E. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association 92,* 633-639.

Nieto-Barajas, L. E., & Contreras-Cristán, A. (2014). A Bayesian nonparametric approach for time series clustering. *Bayesian Analysis*, *9*(1), 147-170.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, *9*(2), 249-265.

Owen, S. (2002). Behavioural finance and the decision to invest in high tech stocks. *University of Technology*. 1-22.

Roache, S. K., & Rossi, M. (2010). The effects of economic news on commodity prices. *The Quarterly Review of Economics and Finance*, *50*(3), 377-385.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems. (TOIS)*, *27*(2:12), 1-19.

Veronesi, P. (1999). Stock market overreactions to bad news in good times: a rational expectations equilibrium model. *Review of Financial Studies*, *12*(5), 975-1007.