A Security Price Data Cleaning Technique: Reynold's Decomposition Approach

Rachel V. Mok
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
rmok@mit.edu

Wai Yin Mok, Kit Yee Cheung
College of Business Administration
University of Alabama in Huntsville
Huntsville, Alabama, 35899
mokw@uah.edu, kityeemok@gmail.com

Abstract

We propose a security price data cleaning technique based on Reynold's decomposition that uses T_0 (the time period of integration) to determine the de-noise level of the price data. The goal of this study is to find the optimal T_0 that reveals an underlying price trend, possibly indicating the intrinsic value of the security. The DJIA (Dow Jones Industrial Average) Index and the thirty companies comprising the index are our fundamental interest. Preliminary results suggest that the graphs of α (a key percentage measure) versus T_0 of the thirty companies and the DJIA Index exhibit at least two properties: (1) α drops exponentially as T_0 increases when $T_0 \lessapprox$ order of magnitude of 100 days, and (2) α drops linearly as T_0 increases when $T_0 \gtrapprox$ order of magnitude of 100 days. For the DJIA Index itself, T_0 is less than order of magnitude of 100 days. The result of applying our technique to each component stock of the DJIA parallels the result of the technique applied to the DJIA Index itself.

1 Introduction

Understanding and analyzing financial data in order to forecast and make cost-effective decisions is challenging because of the complex and volatile nature of security prices. The most recent financial market meltdown in 2008-09 casted doubts on financial data analysis and forecasting. Inability to recognize or acknowledge financial distress signaled by pertinent financial data was a significant factor leading to these catastrophic economic results (Kaur, 2015). Thus, veracity of financial data takes priority in any data driven decision making. Like any big data infrastructure, veracity includes validation, noise level, deception, detection, relevance and ranking of data collected (Goes, 2014). Depending on how collected financial data are captured and processed in an analysis, generated assessments can vary greatly from real financial market performance. One has to look no farther than the recent settlement of $77 million between the SEC and Standard & Poor credit rating agency to see an example of how data analysis can be misleading (http://www.sec.gov/news/pressrelease/2015-10.html).

Several financial computation models that deal with cleaning financial data employ similar methodologies, such as candlestick strategies (Detollenaere and Mazza, 2014), multiple-stage algorithm for detecting outliers in ultra high-frequency financial market data (Verousis and ap Gwilym, 2010), financial data filtering (http://www.olsendata.com) and data-cleaning algorithm (Chung et al., 2004a; Chung et al., 2004b). Most data cleaning methodologies involve the detection, distribution and/or the removal of outliers (Shamsipour et al., 2014; Sun et al., 2013). However removing outliers in the dataset may have a statistical distortion effect on the dataset itself (Dasu and Loh, 2012).

To this end, we propose a data cleaning technique based on Reynold's decomposition in order to decompose the price data into a mean part and a fluctuating part. Fluctuations in stock prices are perpetual and irrational in time because the weak form of market efficiency and different types of market participants create a complex dynamic of behavioral finance (Verheyden et al., 2015). Nevertheless, our approach could minimize part of the effect of irrational price fluctuations by incorporating and averaging fluctuation points (i.e., outliers) within a moving time period of integration, T_0. In essence, the length of T_0 in the analysis determines the level of veracity, with the larger the T_0, the lesser the influence of the fluctuation points will be. We believe our data cleaning tech-

nique is particularly applicable to security prices due to the intense nature of security price changes in relatively short periods, and it allows the user to gauge different moving time periods of integration to produce a unique set of statistical data for targeted analysis.

2 Reynold's Decomposition

In the study of turbulence in fluid dynamics, each component of the velocity is characterized by fluctuations over time. One method to study the dynamics in this regime is to perform a Reynold's decomposition such that the mean part of the velocity is separated from the fluctuations. We propose that this technique could also be used to study financial data. In other words, we propose that the price as a function of time, $p(t)$, can be decomposed into the following:

$$p(t) = \bar{p}(t) + p'(t) \tag{1}$$

where $\bar{p}(t)$ is the mean portion and $p'(t)$ is the fluctuating portion of the price. We define $\bar{p}(t)$ to be a moving time-average that can be found by performing the following integral

$$\bar{p}(t) = \frac{1}{T_0} \int_{t-T_0/2}^{t+T_0/2} p(t')dt' \tag{2}$$

where T_0 is the time period of integration. T_0 must be a time period that is greater than the time period of the fluctuations, τ, and less than the time period of interest, T. T is dependent on each particular analysis; for example, T could be weeks, months, or years. Thus, $\tau < T_0 < T$. Furthermore, the time-averaged value of the fluctuating portion over the entire time period of interest is zero (Müller, 2006; Mills, 1999). As the time period of integration increases, $\bar{p}(t)$ is farther away from the actual $p(t)$ and the magnitude of $p'(t)$ increases. Thus, the goal of this research is to find the optimal time period of integration, T_0, that excludes the miscellaneous fluctuations and captures the essential trend of the price data.

3 Methods

In this study, we focus on the thirty companies comprising the Dow Jones Industrial Average (DJIA) as of May 13, 2015, and the DJIA Index because, being the second oldest financial index, the DJIA is the benchmark that tracks financial market performance as a whole. Thus, it represents a broad market, and its validity is intensely scrutinized and followed by at least 10 Wall Street analysts (Lee and Swaminathan, 1999; Moroney, 2012; Stillman, 1986). The ticker symbols for the thirty companies that were studied in this analysis are as follows: GS, IBM, MMM, BA, AAPL, UTX, UNH, HD, DIS, CVX, NKE, TRV, JNJ, MCD, CAT, XOM, PG, AXP, WMT, DD, V, JPM, MRK, VZ, MSFT, KO, PFE, INTC, CSCO, and GE. Because different companies can comprise the DJIA Index at any point in time, we only focus on the index as a whole when performing the analysis for the DJIA Index itself.

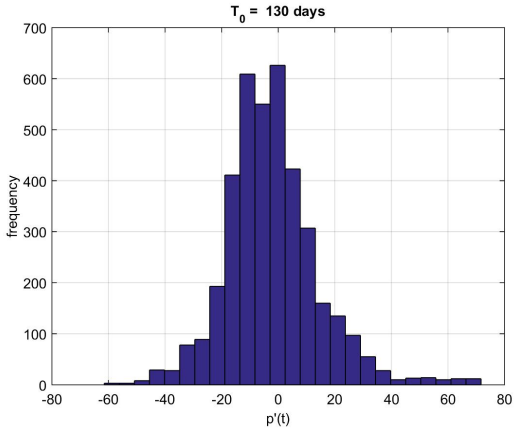

Figure 1: The histogram for GS with T_0 = 130 days.

Daily adjusted close stock price data for the thirty Dow Jones companies listed above from the time of inception of the company to May 13, 2015, are obtained from Yahoo! Finance. For the DJIA Index, the daily adjusted close stock price data from Jan. 29, 1985, to May 13, 2015, are also obtained from Yahoo! Finance. The adjusted close stock price is used because it accounts for stock dividends and splits. Only days in which the stock price is provided, i.e., business days, are considered in this study. Thus, the time from Friday to Monday is taken as only one (business) day.

We estimate the time period of fluctuations to be a day, $\tau \sim 1$ business day, and the time period of interest to be the total number of business days since the inception of the stock, $T \sim 260 \times n$ business days, where n represents the number of years since the inception of the stock. Further, we chose the following time periods of integration, T_0, for this study: 4 days, 10 days, 20 days, 30 days, 64 days, 130 days, 194 days, 260 days, 390

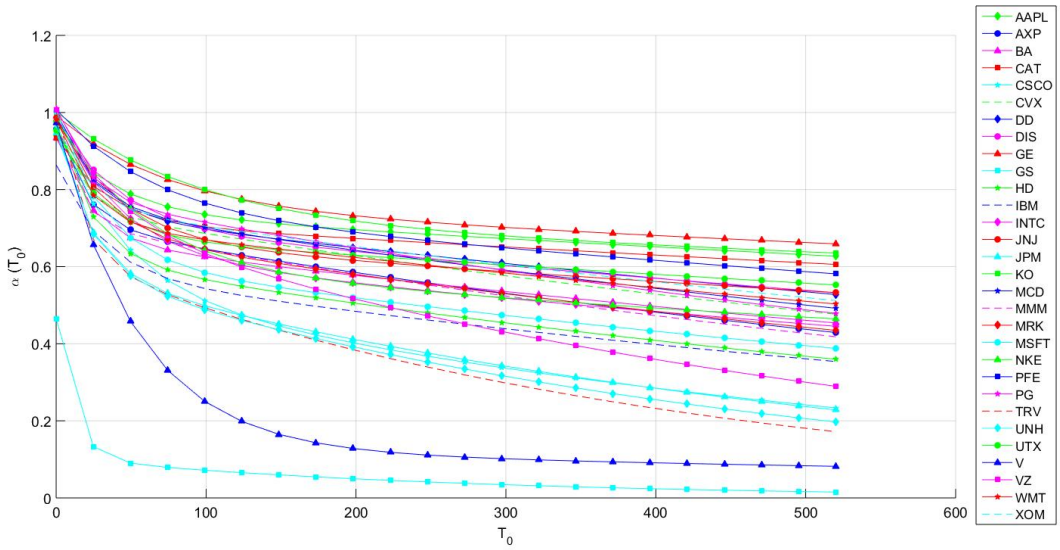

Figure 2: Fitted curve $\alpha(T_0) = a_1\ e^{b_1\ T_0} + c_1\ e^{d_1\ T_0}$ for the listed thirty stocks. $a_1, b_1, c_1,$ and d_1 are curve fitting parameters. The lowest goodness-of-fit measure R^2 among the thirty stocks is 0.9909.

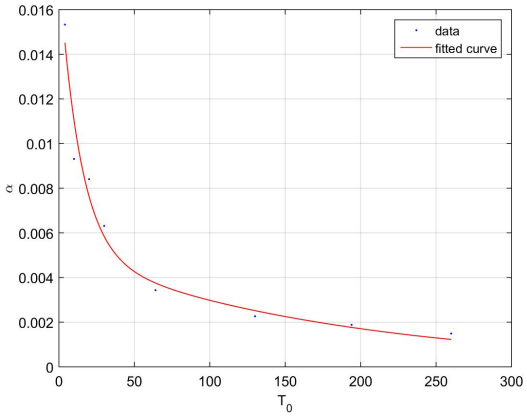

Figure 3: α vs. T_0 for the DJIA Index. The goodness-of-fit measure R^2 is 0.9707 for the fitted curve.

days, and 520 days, which roughly represent the following time periods: one week, two weeks, one month, one-and-a-half months, a quarter of a year, half of a year, three-quarters of a year, one year, one-and-a-half years, and two years, respectively.

$\bar{p}(t)$ is calculated by only considering the analysis time period from $T_0/2$ after the day of inception to $T_0/2$ before May 13, 2015, such that for each day $\bar{p}(t)$ is calculated, the full time period of integration is used. To exemplify, consider the case where $T = 1000$ days and $T_0 = 100$ days. Then the first 50 days (day 1 to day 50) are not included in the analysis, and neither are the last 50 days (day 951 to day 1000). For each day in the analysis time period, the integration stated in Eq. (2) is performed numerically to find $\bar{p}(t)$ for that day. $p'(t)$ is found by subtracting $\bar{p}(t)$ from $p(t)$, the actual price, for that day.

For each specific T_0, the statistics of $p'(t)$ are analyzed. Specifically, a histogram with 25 bins of $p'(t)$ is created for each T_0. As an example, Fig 1 shows a histogram for GS (The Goldman Sachs Group Inc). Note that like Fig 1, most of the histograms are centered around 0, which suggests that most of the fluctuations for the stocks are nearly zero. Therefore, the actual stock price is near or nearly equal to the local time-average for most of the time period analyzed. For most stocks, as T_0 increases, the maximum height achieved by the histogram decreases and the histogram tails become heavier. Thus, as T_0 increases, there are more observations away from the center of the distribution. This is observed because as the time period of integration increases, more points are considered in the average. Therefore, there is a greater likelihood that $\bar{p}(t)$ is different from the actual price.

To measure the fidelity of $\bar{p}(t)$ to $p(t)$, the number of data points of $p'(t)$ that are within 1 dollar from zero are counted and divided by the total number of data points in the analysis period. We will call this percentage measure α, and this measure should be as close as possible to 100% to reflect that $\bar{p}(t)$ is a good approximation of $p(t)$.

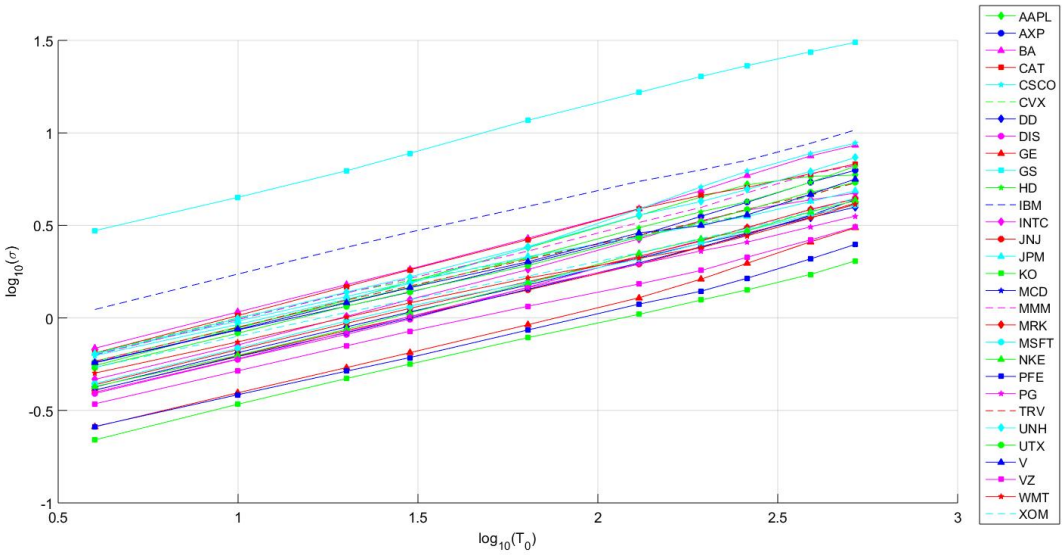

Figure 4: $\log_{10}(\sigma)$ vs. $\log_{10}(T_0)$ for the listed thirty stocks. Because each stock is a line in this plot, a power law relationship exists between σ and T_0.

As stated previously, if $p'(t)$ is near zero, that means that $\bar{p}(t)$ is close to $p(t)$ because $p(t) = \bar{p}(t) + p'(t)$. As T_0 increases, α decreases because the mean is farther away from the actual price when the integration period is larger. Using the MATLAB® curve fitting tool, it is found that for all of the thirty stocks the relationship between α and T_0 is best represented by the following equation

$$\alpha(T_0) = a_1\, e^{b_1\, T_0} + c_1\, e^{d_1\, T_0} \tag{3}$$

where $a_1, b_1, c_1,$ and d_1 are curve fitting parameters. In fact, the lowest goodness-of-fit measure R^2 among all thirty stocks is 0.9909. As an example, the curve fitting parameters for GS are $a_1 = 0.3613, b_1 = -0.09018, c_1 = 0.1034,$ and $d_1 = -0.003687$. The first derivative of this equation is

$$\frac{d\alpha}{dT_0} = a_1\, b_1\, e^{b_1\, T_0} + c_1\, d_1\, e^{d_1\, T_0} \tag{4}$$

and the second derivative is

$$\frac{d^2\alpha}{dT_0^2} = a_1\, (b_1)^2\, e^{b_1\, T_0} + c_1\, (d_1)^2\, e^{d_1\, T_0} \tag{5}$$

For most of the stocks, it was discovered that when T_0 is fewer than 100 days, the measure α drops exponentially as T_0 increases. However, the second derivative (Eq. (5)) becomes near zero in a range from 96 days to 387 days for the thirty stocks analyzed, with the most common being approximately 125 days. Thus, when T_0 is at least an order of magnitude of 100 days, α starts to decrease linearly for nearly all of the stocks analyzed. Fig 2 plots the curve fitted $\alpha(T_0)$ for all thirty analyzed stocks. As we can see, the general trend among the thirty stocks is that α drops exponentially when T_0 is fewer than 100 days, but α drops linearly when T_0 is greater than 100 days. An appealing fact is that the graph of α against T_0 for the DJIA Index, Fig 3, also exhibits similar trends in α, as shown in Fig 2. Note the different scales of the vertical axes of Fig 2 and Fig 3, which means that Fig 3 is much flatter than Fig 2.

Mathematically, we will define the point where the slope is constant by the following

$$\lim_{T_0 \to t_c} \frac{d^2\alpha}{dT_0^2} = 0 \tag{6}$$

where t_c is the time period of integration at which the second derivative of α approaches zero. Thus, for the thirty stocks analyzed, t_c is in the following range 96 days $< t_c <$ 387 days. Therefore, for time periods of integration larger than t_c, the change in α will be relatively small.

The standard deviation σ of the fluctuations $p'(t)$, defined as

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(p'(t_i) - \bar{p'}(t))^2}, \tag{7}$$

is also analyzed where N is the total number of data points and $\bar{p'}(t)$ is the total time average of the

fluctuations. A large σ of the fluctuations reflects that $\bar{p}(t)$ is not equal to $p(t)$. To remove the miscellaneous fluctuations of the price, σ should be as large as possible. As indicated by the straight lines in the log-log plot in Fig 4, σ and T_0 are related by a power law where the slope of the line indicates the exponent.

Using the MATLAB® curve fitting tool, we fit the following equation for each of the thirty stocks:

$$\sigma = a_2 \, T_0^{b_2} + c_2 \tag{8}$$

where $a_2, b_2,$ and c_2 are curve fitting parameters. As an example, $a_2 = 2.194, b_2 = 0.4317,$ and $c_2 = -1.425$ for GS. The lowest goodness-of-fit measure R^2 among the thirty stocks is 0.9849. b_2, the exponent, varies from 0.35 to 0.69 for all thirty stocks. The average exponent is 0.5.

4 Results and Conclusions

This paper demonstrates preliminary results of an ongoing security price data cleaning research. We found that the graphs of α versus T_0 of the thirty companies and the DJIA Index exhibit at least two properties: (1) α drops exponentially as T_0 increases when $T_0 \lessapprox$ order of magnitude of 100 days, and (2) α drops linearly as T_0 increases when $T_0 \gtrapprox$ order of magnitude of 100 days. Thus, the optimal T_0 for the thirty companies studied is approximately 100 days. For the DJIA Index itself, the optimal T_0 appears to be less than 100 days. One of the possible explanations is that the DJIA Index might show the counter measure effect of fluctuation points among the thirty companies since the DJIA is a composite of the thirty companies that collectively provide a balance view of the market. As a result, T_0 might be even smaller for the second derivative to approach zero. We also found that σ and T_0 are related by a power law. As for future research, we plan to define mathematical metrics in our study of security price valuations and trading strategies.

References

Kee H. Chung, Chairat Chuwonganant, and D. Timothy McCormick. 2004a. Order preferencing and market quality on NASDAQ before and after decimalization. *Journal of Financial Economics*, 71(3):581–612.

Kee H. Chung, Bonnie F. Van Ness, and Robert A. Van Ness. 2004b. Trading costs and quote clustering on the NYSE and NASDAQ after decimalization. *Journal of Financial Research*, 27(3):309–328.

Tamraparni Dasu and Ji Meng Loh. 2012. Statistical distortion: Consequences of data cleaning. *Proceedings of the VLDB Endowment*, 5(11):1674–1683.

Benoit Detollenaere and Paolo Mazza. 2014. Do Japanese candlesticks help solve the trader's dilemma? *Journal of Banking and Finance*, 48:386–395, November.

Paulo B. Goes. 2014. Big data and IS research. *MIS Quarterly*, 38(3):iii–viii, September.

Inderjit Kaur. 2015. Early warning system of currency crisis : Insights from global financial crisis 2008. *IUP Journal of Applied Economics*, 14(1):69–83, January.

Charles M. C. Lee and Bhaskaran Swaminathan. 1999. Valuing the Dow: A bottom-up approach. *Financial Analysts Journal*, 55(5):4–23, September.

Anthony F. Mills. 1999. *Basic Heat and Mass Transfer*. Prentice Hall, second edition.

Richard Moroney. 2012. What we're thinking add it up: Dow has further upside. *Dow Theory Forecasts*, 68(9):2–3, February.

Peter Müller. 2006. *The Equations of Oceanic Motions*. Cambridge University Press.

Mansour Shamsipour, Farshad Farzadfar, Kimiya Gohari, Mahboubeh Parsaeian, Hassan Amini, Katayoun Rabiei, Mohammad Sadegh Hassanvand, Iman Navidi, Akbar Fotouhi, Kazem Naddafi, Nizal Sarrafzadegan, Anita Mansouri, Alireza Mesdaghinia, Bagher Larijani, and Masud Yunesian. 2014. A framework for exploration and cleaning of environmental data - Tehran air quality data experience. *Archives of Iranian Medicine*, 17(12):821–829, December.

Richard Joseph Stillman. 1986. *Dow Jones Industrial Average : history and role in an investment strategy*. Irwin Professional Pub.

W. Sun, B. Whelan, AB. McBratney, and B. Minasny. 2013. An integrated framework for software to provide yield data cleaning and estimation of an opportunity index for site-specific crop management. *Precision Agriculture*, 14(4):376–391, August.

Tim Verheyden, Lieven De Moor, and Filip Van den Bossche. 2015. Towards a new framework on efficient markets. *Research in International Business and Finance*, 34:294–308, May.

Thanos Verousis and Owain ap Gwilym. 2010. An improved algorithm for cleaning ultra high-frequency data. *Journal of Derivatives & Hedge Funds*, 15(4):323–340.