Building Factors using Natural Language Processing on Regulatory Filings

Alexander P. Jermann [*]
Dept. of Industrial Engineering & Operations Research
Columbia University
New York, 10027, NY
alexander.jermann@columbia.edu

1 Introduction

Every year (and quarter) every U.S. public company has to disclose a comprehensive summary of their financial standing to the U.S. Securities and Exchange Commission ("SEC") to comply with disclosure requirements. This report is public and highly standardized and includes a description of the business, its corporate structure and governance, risk factors, legal proceedings, financial statements, and more.

Empirical evidence shows that these regulatory reports are generally not entirely rewritten year-to-year but rather edited to fit the company's current financial standing. This means that there is a high textual similarity between reports. Cohen et al. (2018) made the discovery that the textual similarity between consecutive reports from the previous to the current year was correlated to short-term subsequent financial performance of the company. In other words, if a company changed a lot of the text from their previous to their most recent regulatory filing the subsequent financial performance was lower relative to companies with fewer text changes. Cohen et al. (2018) showed that the results were robust over longer periods of time from 1995-2014. This result is surprising as the efficient market hypothesis postulates that any public information is priced into the asset.

For this project, I provide a short introduction by giving a possible economic rationale for the existence of such factors and analyze the robustness and risks of said factor. Using the theoretical and empirical findings by Cohen et al. (2018), I then build a natural language processing ("NLP") model in Python that automatically compared the documents by filing year (and quarter) to derive statistics. I subsequently use those statistics as a signal to build an equity long/short portfolio to verify the validity of the theory in the years after the paper was published. In this project, I found that the factors provided robust results during the time period of 2015 to 2016 providing 6-7% yearly returns and that factors have since lost on their predictive power only rendering 2% returns.

2 Background and Relevant Theory

2.1 Regulatory Filings

Every U.S. public company is required to disclose information to shareholders information on an ongoing basis to comply with disclosure requirements. The annual and quarterly reports provide a comprehensive overview of the company's business and financial standing ranging from information on the business model, risk factors, legal proceedings, financial statements, corporate governance, and more (U.S. Securities and Exchange Comission). Companies try to be strategic as to what information is included in the report and how it is included. There is an inherent trade-off between complying with disclosure requirements and revealing as little information as possible that could negatively affect the firms subsequent financial performance as shown by (Dyer et al., 2017) and Lee (2012).

[*] alternate email: alexander.jermann@gmail.com

Preprint. Under review.

Zhang (2009) and Easton and Zmijewski (1993) find evidence of investor underreaction to information released in 10-K reports and a corresponding lag in the reaction of the underlying asset pricing. They identify that trading volumes and stock price movements happen during the weeks before the release of the official 10-K report and are sluggish after the release of the 10-K report and only get priced over the next few weeks. They suggest that this is due to the fact that metrics like earnings per share, dividends, sales growth are disclosed and made public weeks before during earnings calls. As a result, investors might view the filing as superfluous and largely ignore it. This fact is supported by the number of downloads of the 10-K reports on the S.E.C. online system EDGAR.

Lee (2012) explores an alternative explanation for the delayed response in pricing. He argues that while the semi-strong form of the efficient market hypothesis (see Fama (1970) for the hypothesis) states that capital markets incorporate all public information into security pricing in a timely and efficient manner, judgement and decision research suggests otherwise. Hirshleifer et al. (2003) find that investors, analysts have bounded rationality in the amount of information that they can process and can thus not attend to all information made available. Measuring complexity of documents in terms of word count, empirical evidence suggests that historically more complex documents are more heavily mispriced and have larger price adjustments in subsequent trading days.

Finally, Cohen et al. (2018) suggests companies are lazy and consistently use previous 10-K reports as templates for current reporting requirements. The author finds evidence that companies tend to only edit sections to reflect the current companies business and financial health. The authors find that companies that do not change their financial reports by a lot ("non-changers") tend to relatively outperform companies that make a lot of changes ("changers"). They find that a long/short portfolio that goes long on the top 20% ("non-changers" as in high textual similarity between documents) and short on the bottom 20% ("changers") by an average of 5-7% annually, supporting the assumption that companies only report what is necessary.

In summary, this section provided evidence that there is quantifiable evidence of investor inattention to 10-K reports, that the information released in 10-K and 10-Q reports has a significant lag and is only priced into the underlying security over the following weeks through gradual information processing or more immediate news reporting. This section has also shown that textual changes of subsequent reports have historically provided a signal of future financial performance of companies. Arguably, it thus makes sense to use a systematic approach to quantify changes of financial reports as a factor for investing.

2.2 Factor Investing

Factor investing is an investment approach where quantifiable characteristics or "factors" are used to explain differences in stock returns. Typical factors that were first published 30 years ago include: size, value, momentum, asset growth, and leverage (Value, 2016). The earliest paper on factors was published by Ross (1976). In his paper, Ross proposes an arbitrage pricing theory that holds that the expected return of assets can be modeled as the linear combination of various factors as follows:

$$r_j = a_j + \lambda_{j1} f_1 + \lambda_{j2} f_2 + ... + \lambda_{jn} f_n + \epsilon_j \tag{1}$$

where

- a_j is a constant for asset j
- f_n is a systematic factor
- λ_{jn} is called the factor loading and represents the relative sensitivity of the j-th asset to the n-th factor
- ϵ_j is the j-th asset idiosyncratic risk with mean 0.

More rigorously we define returns as $r \in \mathbb{R}^m$, factor loadings as $\Lambda \in \mathbb{R}^{m \times n}$, and factors as $f \in \mathbb{R}^n$ and is:

$$r = r_f + \Lambda f + \epsilon, \epsilon \sim \mathcal{N}(0, \Psi) \tag{2}$$

where ϵ follows a multivariate normal distribution with mean 0. In this model, we assume that

$$f \sim \mathcal{N}(\mu, \Omega) \tag{3}$$

where μ is the expected risk premium vector and Ω is the factor covariance matrix. We thus get the expected returns of:

$$\mathbb{E}(r) = r_f + \Lambda \mu, \; Cov(r) = \Lambda \Omega \Lambda^T + \Psi \tag{4}$$

It is assumed that the factors are known in a given model. In the subsequent sections we will use the similarity measures of the documents as the factors.

3 Methodology

To measure how effective textual changes between subsequent regulatory filings of individual companies are in predicting future returns of companies, we first download all the 10-K and 10-Q reports of the past few years for every U.S publicly traded company, define and compute the similarity measures between every document of a given company, and then map the factor to the historical asset price and then analyze how well the factor predicts future returns. As a final step we use the factor to rank assets and build a long/short portfolio that goes long on assets that have high factor values (top quintile or 20% and goes short on the bottom quintile or 20%). The next few subsections outline the methodology in more detail.

3.1 Data Collection & Scraping

In order to build the necessary sample dataset of 10-K and 10-Q reports, we had to scrape the SEC's Electronic Data Gathering, Analysis, and Retrieval ("EDGAR") site. For context, the average length of a 10-K report in 2017 was 60'000 words long which is the length of the Book Lord of the Flies by William Golding. The final dataset of 10-K and 10-Q reports between 2011 and 2018, contained 260'425 documents from all U.S. companies listed on the NYSE, AMEX, or NASDAQ in that time-period. This resulted in a dataset of approximately 80 gigabytes large. Since we were only interested in textual analysis of these documents we discard supplementary material such as images, tables, PDF files, excel files, etc. All the necessary code and documentation on how to download, pre-process, and get the dataset into the right format is linked in the code section 5.1 at the end of this paper.

3.2 Difference Measures

To measure the similarity between two documents we use two metrics that are common in the literature of textual analysis and natural language processing. Namely the Jaccard Similarity Measure and the Cosine similarity measures.

3.2.1 Jaccard Similarity Measure

The Jaccard Similarity measure compares members of two sets and compares which members are shared and which are distinct.Intuitively, the Jaccard Similarity measure is the size of the intersection, divided by the size of the union of the words. The value it returns will be between 0 and 1, where zero would be no words and common, and 1 would indicate that both sets have the same words.

Definition: Let A and B be words, then the Jaccard Similarity score is

$$J(A, B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} \tag{5}$$

where $0 \geq J(A, B) \geq 1$. If both $A, B = \emptyset$ we define $J(A, B) = 1$.

In the context of comparing documents, it is a somewhat naive measure as it does not take into consideration word counts and frequencies of words as opposed to the cosine measure in the following section. Further, it does not take into consideration word similarity and contextual meaning of words. A project by Kai et al. (2017) showed that using word-embeddings to capture contextual information did not increase the predictive power of the factor. Intuitively, this makes sense since a 60'000 word document is being condensed into a single number between 0 and 1. So much information is lost in this mapping that the extra information captured by contextual information does not significantly improve the predictive power. We note however that we find that word count and frequency does improve the accuracy as we will see in the result section.

3.2.2 Cosine Similarity Measure

Let A and B be words. We map A and B into the vector space S, where S has dimension of the union between the word sets A and B

The dimensions of the vector-space are the set of the union between A and B. The vector $A \in$

$$C(A, B) = \frac{\mathbf{A} \cdot \mathbf{B}}{|| \ \mathbf{A} \ || \cdot || \ \mathbf{B} \ ||} \tag{6}$$

where the nominator is the dot product between the vectors $\mathbf{A}$ and $\mathbf{B}$ and the denominator is the Euclidean norm.

3.3 Statistical Factor Analysis

In this section, we introduce the tools and measures for analyzing a given alpha factors effectiveness at predicting future returns. As discussed in section 2.2, alpha factors express a predictive relationship between the set of information and future returns. It is important to note that these measure merely indicate a statistical relation and have to be rigorously back-tested before being able to implement them.

3.3.1 Information Coefficient

The information coefficient ("IC") provides a measure for the predictiveness of the alpha factor. It describes the correlation between predicted and actual stock returns. The values of the IC range between $[-1; 1]$ where -1 describes no correct prediction and +1 describes perfect prediction of future returns. As a rule of thumb any value above 0 is satisfactory and a mean value above 0.1 is excellent.

3.3.2 Return Analysis

To analyze the factors ability to generate returns, we analyze the historical performance using historical stock prices of assets. Since historical pricing information is public and easily accessible, and U.S. public companies are required by law to submit quarterly reports (which we base our factors on), the universe of stocks is large enough universe of stocks that we can use for analysis.

Our general methodology is to use the factors to rank the assets in 30 day periods and divide them into quintiles where we long the best performing quintile (the stocks with the highest similarity and thus the least changes) and short the bottom quintile (the stocks with the lowest similarity and thus many changes). We use mean period-wise return by factor quintiles, and cumulative return measures.

4 Results and Discussion

In this section we present the results of the experimentation. The main aim of this paper was to explore the factor predictiveness of textual similarity measures in predicting future returns after the main paper by Cohen et al. (2018) was published in 2018 using data from 1994 - 2015. In this paper we, apply the same factors to a dataset of U.S. public companies' 10-K and 10-Q reports from 2015 to 2018. The results are presented in two subsections, one for the Jaccard Similarity measure and one for the Cosine similarity measure. Please see the appendix in section 5 for the documentation and code used to compute the graphs in the following subsections.

4.1 Jaccard Similarity

See figures 1 - 5

4.2 Cosine Similarity

See figures 6 - 10

4.3 Discussion

Analyzing the results from sections 4.1 and 4.2 we can make several observations. Both, the jaccard similarity factor and the cosine similarity factor demonstrated relatively good information coefficients

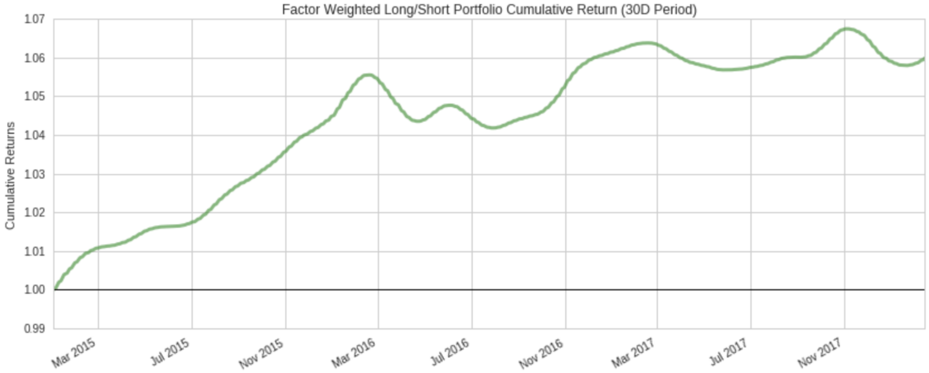


Figure 1: This graph shows the cumulative return of the factor weighted long/short portfolio over 30 day periods. We see that between March of 2015 and March 2016 the cumulative return is 6%. This result corresponds to the average range of cumulative returns of the factor in the years 1994-2015 as found by Cohen et al. (2018). However the subsequent year starting in March of 2016, the returns seem to flatten out to 2%.

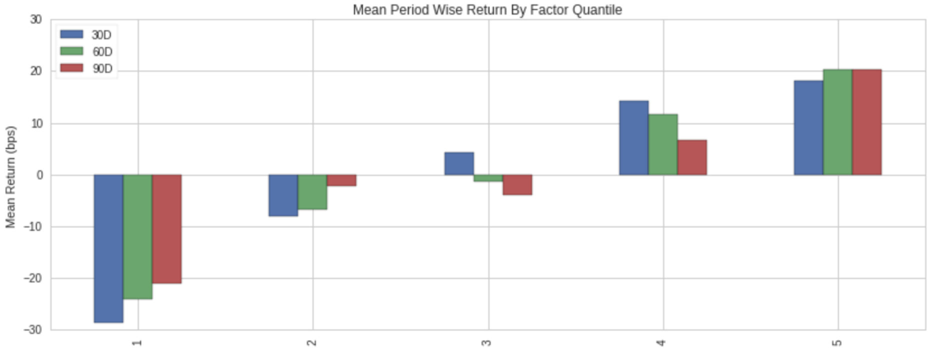


Figure 2: On this bar plot we see the mean period-wise return by factor quantile, where the quantile number 1, represents the stocks that had the lowest similarity scores, meaning that they had the largest changes. We note that the absolute value of the mean return for the bottom quantile (around 25 basis points) exceeds the returns from going long the stocks with the highest factor values (20 basis points).

and subsequently were able to generate 6-7% return in the time period of March 2015 to March 2016 when using a long/short strategy. After March 2016 their performance of the two factors diverged, in that the cosine factor performed better than the jaccard similarity factor but still not as well as they did in the previous year or as found in the main paper by Cohen et al. (2018). We hypothesize that the cosine similarity factor performs better because it also takes into account word count and term frequencies, which adds more fine-grained information especially when considering negative words. In conclusion, we can hypothesisze that the predictive power of the textual similarity factor has subsided since the publishing of the paper. However, to confirm this we would have to continue to check the factor effectiveness in the coming years.

For future work, I would want to further explore the textual similarity between subsections of the regulatory reports, since Cohen et al. (2018) found that the Risk section was especially indicative of future performance.

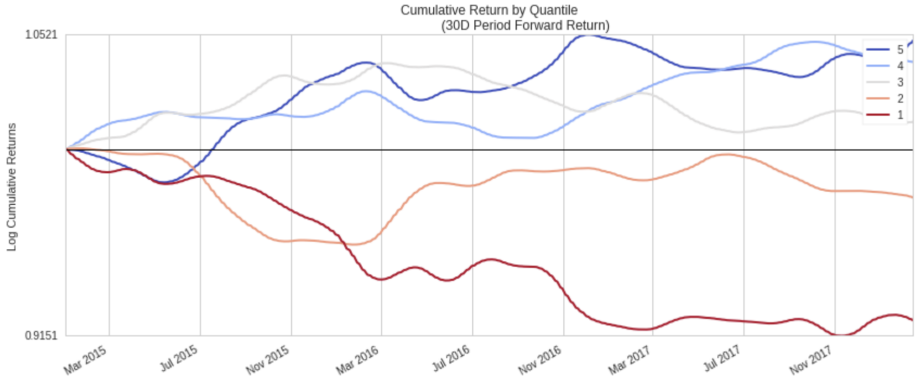
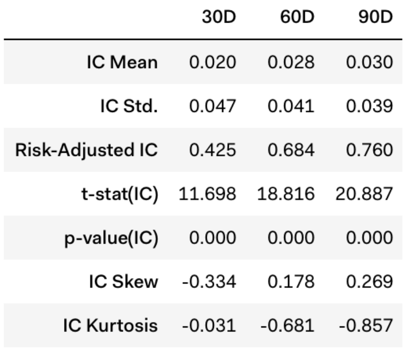


Figure 3: This graph shows the cumulative return of the factor weighted long/short portfolio in 30 day period by quantile. The graph shows nicely how there is a large spread between the bottom quantile (red) goes down, and the top quantile that goes up (Blue).



	30D	60D	90D
IC Mean	0.020	0.028	0.030
IC Std.	0.047	0.041	0.039
Risk-Adjusted IC	0.425	0.684	0.760
t-stat(IC)	11.698	18.816	20.887
p-value(IC)	0.000	0.000	0.000
IC Skew	-0.334	0.178	0.269
IC Kurtosis	-0.031	-0.681	-0.857

Figure 4: The table illustrates the information coefficient mean and distribution over time periods of 30, 60, and 90 days. We note that a number above zero indicates the factors ability to predict future returns. As a rule of thumb a value above zero is satisfactory and a value above 0.1 is exceptional.

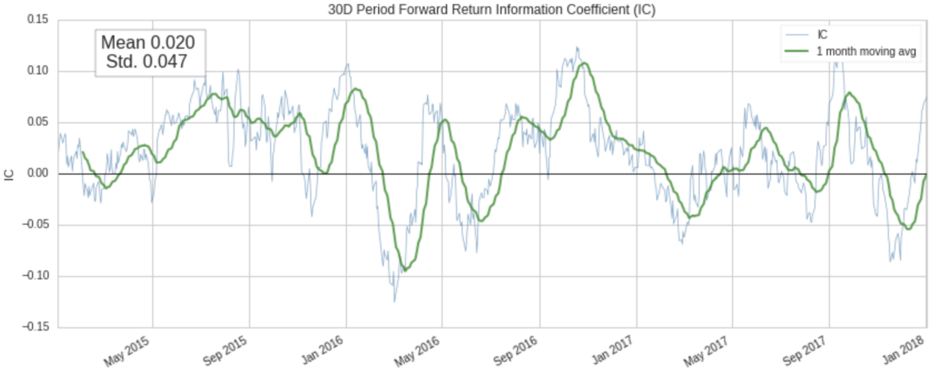


Figure 5: The time series shows the the 30 day revolving mean and standard deviation over time.

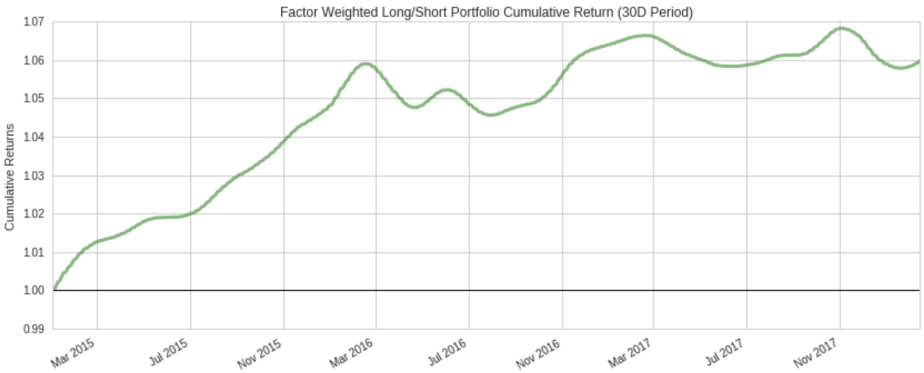


Figure 6: On the graph we see the cumulative return of the factor weighted long/short portfolio over 30 day periods for the cosine similarity measure. We see that between March of 2015 and March 2016 the cumulative return is 6%. This result corresponds to the average range of cumulative returns of the factor in the years 1994-2015 as found by Cohen et al. (2018).

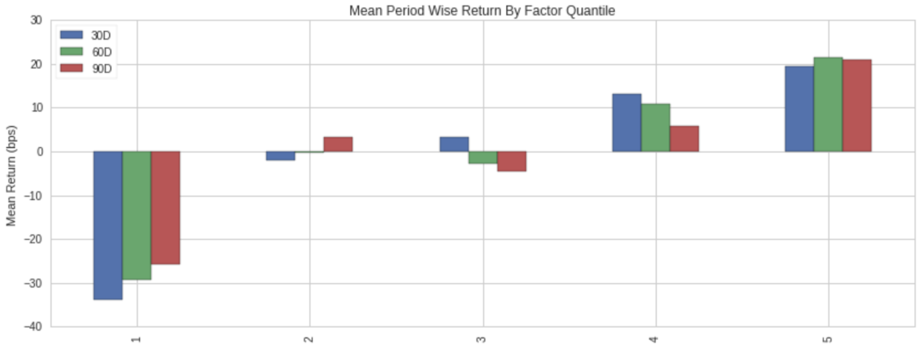


Figure 7: On this bar plot we see the mean period-wise return by factor quantile, where the quantile number 1, represents the stocks that had the lowest similarity scores, meaning that they had the largest changes. We note that the absolute value of the mean return for the bottom quantile (around 25 basis points) exceeds the returns from going long the stocks with the highest factor values (20 basis points).

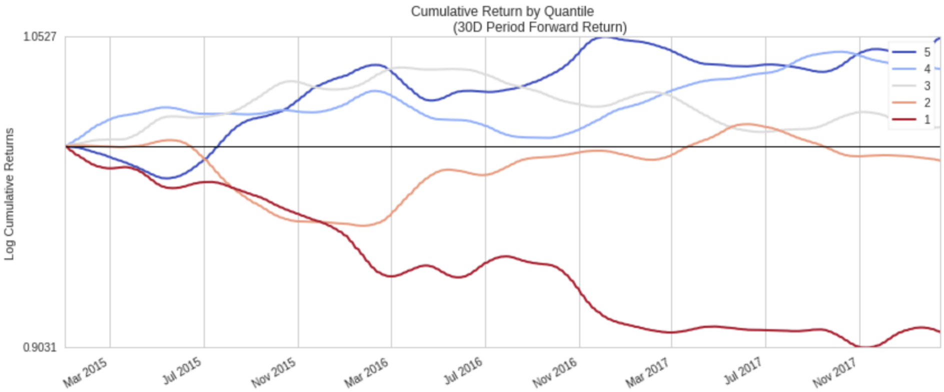
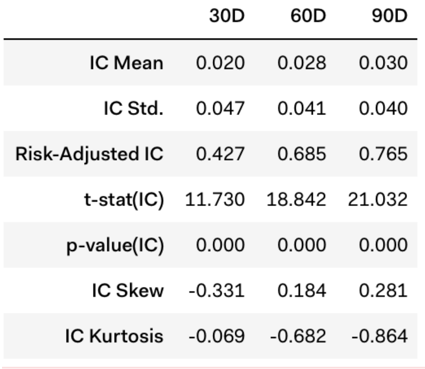


Figure 8: This graph shows the cumulative return of the factor weighted long/short portfolio in 30 day period by quantile. The graph shows nicely how there is a large spread between the bottom quantile (red) goes down, and the top quantile that goes up (Blue).



	30D	60D	90D
IC Mean	0.020	0.028	0.030
IC Std.	0.047	0.041	0.040
Risk-Adjusted IC	0.427	0.685	0.765
t-stat(IC)	11.730	18.842	21.032
p-value(IC)	0.000	0.000	0.000
IC Skew	-0.331	0.184	0.281
IC Kurtosis	-0.069	-0.682	-0.864

Figure 9: The table illustrates the information coefficient mean and distribution over time periods of 30, 60, and 90 days. We note that a number above zero indicates the factors ability to predict future returns. As a rule of thumb a value above zero is satisfactory and a value above 0.1 is exceptional.

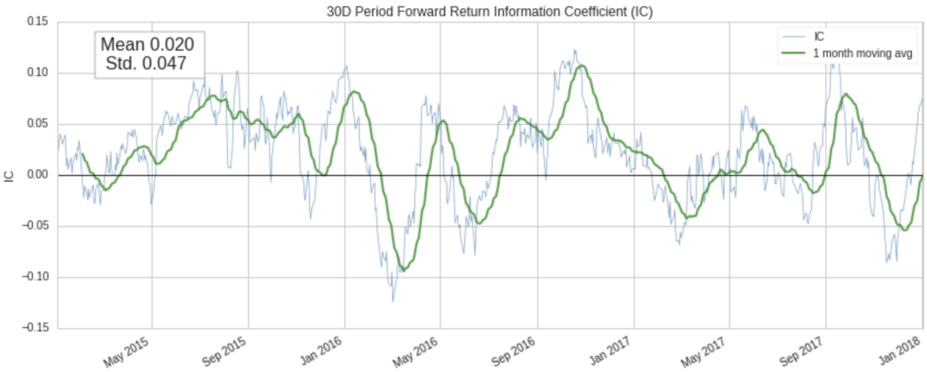


Figure 10: The time-series graph shows the mean IC measure over time.

References

Lauren Cohen, Christopher Malloy, and Quoc Nguyen. Lazy Prices. Working Paper 25084, National Bureau of Economic Research, sep 2018. URL `http://www.nber.org/papers/w25084`.

Travis Dyer, Mark Lang, and Lorien Stice-Lawrence. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, 64(2):221–245, 2017. ISSN 0165-4101. doi: https://doi.org/10.1016/j.jacceco.2017.07.002. URL `http://www.sciencedirect.com/science/article/pii/S0165410117300484`.

Peter D Easton and Mark E Zmijewski. SEC Form 1OK / 1OQ Reports and Annual Reports to Shareholders : Reporting Lags and Squared. 31(1):113–129, 1993.

Eugene F Fama. Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, 25(2):383–417, 1970. ISSN 00221082, 15406261. doi: 10.2307/2325486. URL `http://www.jstor.org/stable/2325486`.

David Hirshleifer, Siew Hong Teoh, Natasha Burns, Dick Dietrich, John Fellingham, Gerry Garvey, Jack Hirshleifer, Jack Hughes, Eugene Kandel, Sonya Seongyeon Lim, Bruce Miller, Ro Verrecchia, and Jerry Zimmerman. Limited attention , information disclosure , and financial reporting. 36: 337–386, 2003. doi: 10.1016/j.jacceco.2003.10.002.

Kuspa Kai, Victor Cheung, and Alex Lin. Lazy Prices: Vector Representations of Financial Disclosures and Market Outperformance, 2017.

Yen-Jung Lee. The Effect of Quarterly Report Readability on Information Efficiency of Stock Prices*. *Contemporary Accounting Research*, 29(4):1137–1170, dec 2012. ISSN 0823-9150. doi: 10.1111/j.1911-3846.2011.01152.x. URL `https://doi.org/10.1111/j.1911-3846.2011.01152.x`.

Stephen A Ross. The Arbitrage Theory of Capital Asset Pricing. pages 341–360, 1976.

U.S. Securities and Exchange Comission. Form 10-K & 10-Q. URL `http://www.sec.gov/edgar.shtml`.

Combining Value. Combining value and momentum. 14(2):33–48, 2016.

Haifeng You Æ Xiao-jun Zhang. Financial reporting complexity and investor underreaction to 10-K information. pages 559–586, 2009. doi: 10.1007/s11142-008-9083-2.

5 Appendix

5.1 Code

All the necessary code is available under the following link: https://github.com/jermann/gcm_project. The code and all the graphs is also made available in the appendix of this paper.