

Lecture 8

Machine learning in trading

Alan Kwan, PhD



THE UNIVERSITY OF HONG KONG
Faculty of Business and Economics

Machine learning

- A lot is said about machine learning in quantitative trading
- To be clear, there are some very advanced things going on, much of which will never surface, but most people do very simple things
- Today we will provide a simple overview of machine learning in finance
- For a better class, take it from Luo Ye

Different objectives of machine learning in trading

- One type is to improve predictive regressions
- Another type is to extract features from data natural language processing
 - Probably this is the thing you have most commonly
- Compress data into a few dimensions principal components analysis is one such example, network analysis is another
- Dealing with unstructured data trying to use a machine to clean data – can be viewed as part of alternative data

Natural language processing

➤ Sources of information

- Traditional data: Company filing, analyst report, conference call transcripts
- Alternative data: Reports from social media, government filings / newsletters that are not intended for a financial audience
- In between traditional and alternative: news media

➤ Try to make a machine understand the article contents

- Tone – is this good news or bad news?
- Subject – what company does this refer to?
- Topic – is this a factory fire? Is there a deep story that is hard to detect?

➤ If the subject is a stock, the topic is important, and the news is good or bad, then there may be a trading opportunity

- To understand NLP more deeply, please take Matthias Buehlmaier's course (spelling?)

High versus low frequency

- On the one hand, NLP can be used to read lots of information a human may not be able to read
 - In this case, you have to be able to trade fast.
 - You can either read the news yourself or buy a service that does this for you. **Ravenpack** is an example of a dataset
- On the other hand, NLP might be used to read information **faster than a human**
- Both are different ways one could make money, but
 - The faster the horizon you have, the simpler the technology you have to apply

NLP – types of strategies

➤ Tone analysis:

- We have found bad news, so let's short, or this is good news

➤ Narrative:


- The managers are trying to delay the revelation of bad news
- The manager is giving early warning signs

➤ Sustainability

- Regardless of whether or not there is alpha, can we try to track companies based on their track record of sustainability?
- Requires identifying subject (is it a publicly listed company?), topic (is it a sustainability concern?), and tone (is this good news or bad news?)

Focus will be corporate filings

[ABOUT](#) | [DIVISIONS](#) | [ENFORCEMENT](#) | [REGULATION](#) | [EDUCATION](#) | [FILINGS](#) | [NEWS](#)



EDGAR Search Tools

Latest Filings

Company Filings

Mutual Funds

Variable Insurance Products

Daily Filings by Type

Boolean Archive Search


Full Text (Past 4 Years)

CIK Lookup

Confidential


EDGAR | Company Filings

Free access to more than 21 million filings

Company Name 

SEARCH

More Options ▶

Fast Search 

SEARCH

Ticker symbol or CIK is the fastest way to find company filings.

Guides

How to Research Public Companies

Learn [how to quickly research](#) a company's operations and financial information with EDGAR search tools.

Form Types

Review [reference versions of EDGAR forms](#) filed by companies, funds, and individuals.

Search Tools

CIK Lookup Tool

Look up the [central index key \(CIK\)](#) of an EDGAR filer. Searching by CIK is the most accurate way to view filings.

Save Your Search

Want to get updates on new filings? Learn how to [save your search](#) by subscribing to EDGAR RSS feeds.

7

Focus will be corporate filings

Tesla, Inc. CIK#: 0001318605 (see all company filings)

SIC: 3711 - MOTOR VEHICLES & PASSENGER CAR BODIES

State location: CA | State of Inc.: DE | Fiscal Year End: 1231

formerly: TESLA MOTORS INC (filings through 2017-01-27)

(Assistant Director Office: 5)

Get **insider transactions** for this **issuer**.

Business Address

3500 DEER CREEK RD

PALO ALTO CA 94304

650-681-5000

Mailing Address

3500 DEER CREEK RD

PALO ALTO CA 94304

Filter Results: Filing Type: Prior to: (YYYYMMDD) Ownership? ☐ include ☒ exclude ☐ only Limit Results Per Page 40 Entries Search Show All

Items 1 - 40  [RSS Feed](#)

Next 40

Filings	Format	Description	Filing Date	File/Film Number
CT ORDER	Documents	Confidential treatment order Acc-no: 9999999997-19-000570 (33 Act) Size: 94 KB	2019-02-25	333-164593 19627948
CT ORDER	Documents	Confidential treatment order Acc-no: 9999999997-19-000569 (34 Act) Size: 91 KB	2019-02-25	001-34756 19627947
CT ORDER	Documents	Confidential treatment order Acc-no: 9999999997-19-000568 (34 Act) Size: 91 KB	2019-02-25	001-34756 19627946
SC TO-T	Documents	Tender offer statement by Third Party Acc-no: 0001193125-19-044863 Size: 103 KB	2019-02-20	
S-4	Documents	Registration of securities, business combinations Acc-no: 0001193125-19-044857 (33 Act) Size: 3 MB	2019-02-20	333-229749 19617204
10-K	Documents Interactive Data	Annual report [Section 13 and 15(d), not S-K Item 405] Acc-no: 0001564590-19-003165 (34 Act) Size: 30 MB	2019-02-19	001-34756 19613254
SC 13G/A	Documents	[Amend] Statement of acquisition of beneficial ownership by individuals Acc-no: 0001193125-19-040787 (34 Act) Size: 42 KB	2019-02-14	005-85943 19606156
SC 13G	Documents	Statement of acquisition of beneficial ownership by individuals Acc-no: 0001104659-19-008581 (34 Act) Size: 170 KB	2019-02-14	005-85943 19604164
SC 13G	Documents	Statement of acquisition of beneficial ownership by individuals Acc-no: 0001422849-19-000129 (34 Act) Size: 8 KB	2019-02-14	005-85943 19599678
SC 13G/A	Documents	[Amend] Statement of acquisition of beneficial ownership by individuals Acc-no: 0000315066-19-001389 (34 Act) Size: 8 KB	2019-02-13	005-85943 19595203
SC 13G/A	Documents	[Amend] Statement of acquisition of beneficial ownership by individuals Acc-no: 0000315066-19-001389 (34 Act) Size: 8 KB	2019-02-11	005-85943 19595203

Downloading filings is easy

Package 'edgarWebR'

May 25, 2018

Title SEC Filings Access

Description A set of methods to access and parse live filing information from the U.S. Securities and Exchange Commission (SEC - <<https://sec.gov>>) including company and fund filings along with all associated metadata.

Version 1.0.0

Maintainer Micah J Waldstein <micah@waldste.in>

Date 2018-05-24

Depends R (>= 3.4.0)

Imports xml2, methods, http

Suggests covr, ggplot2, knitr, purrr, rmarkdown, httptest, tokenizers, devtools, dplyr, tidyr, roxygen2

VignetteBuilder knitr

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

python-edgar 2.5

```
pip install python-edgar
```



Download the SEC filings index from EDGAR since 1993

Navigation

Project description

Release history

Download files

Project description

[[Build Status]](<https://travis-ci.org/edouardswiac/python-edgar.svg?branch=master>)

[[PyPI]](<https://img.shields.io/pypi/v/python-edgar.svg>)

[[PyPI - License]](<https://img.shields.io/pypi/l/python-edgar.svg>)

[[PyPI - Python Version]](<https://img.shields.io/pypi/pyversions/python-edgar.svg>)

Build a master index of SEC filings since 1993 with `python-edgar`

Tone analysis

- Tone: if it sounds negative, then it may be conveying negative news
- Simple, naïve form:
 - Pick a list of bad words and good words
 - Assign a score to a document based on the fraction of bad words or good words
- There are much more complicated tone analysis models
- Having been to quant conferences in Asia, it seems like basic tone analysis may still work in the Asian context or in non-English languages

Word lists

- What is a negative word, what is a positive word?
 - For financial analysis, it may not be the case words that are normally negative are negative in the context of financial analysis
 - Different words apply to different languages
 - Moreover, there is formal speech, but a tweet, may not mean the same thing
 - \$WAG could be a reference to Walgreens, or a rapper's reference to SWAG - money or goods taken by a thief or burglar! 😊
- We will study two ways people develop word lists

What to do with word lists

- Once you develop a word list, you simple score documents for tone

- $$\frac{\#negativeWords}{\#words}$$

- $$\frac{\#positiveWords}{\#words}$$

- $$\frac{\#negative - \#positive\ words}{\#words}$$

McDonald-Loughran Analysis

- They first downloaded all 10-Ks from 1995-2008
 - Extracted Management Discussion and Analysis
- Develop a list of all common words (5% of 10-Ks or more)
 - Remove stupid words like “and”, “or”, “the” etc. these are called stop words
 - “To create the Fin-Neg, Fin-Pos, Fin-Unc, and Fin-Lit word lists, we first develop a dictionary of words and word counts from all 10-Ks filed during 1994 to 2008. We carefully examine all words occurring in at least 5% of the documents, to consider their most likely usage in financial documents (including inflections). Words that we include beyond the 5% level are typically inflections of root words that made the original cut.”
- Used the Harvard dictionary, which defined negative words
 - Sometimes words the Harvard dictionary defines as negative or neutral are not that in the context of finance. “Negative”, “interest”, etc.
 - Liability is a negative word
- Finally, they validate by showing their word dictionary correlates to negative announcement returns
 - That is, if they are really capturing negative news, then more negative words should explain negative news

Python

- [EDGAR_DownloadForms_v2.1.py](#)

Program to download [EDGAR](#) files from the SEC site by form type.

- Dependencies (i.e., modules you must download that are accessed by the program):

[EDGAR_Forms_v2.1.py](#) - module that can be imported to provide convenient lists of form variants.

[EDGAR_Pac_v2.1.py](#) - module containing utility subroutines to facilitate downloading the EDGAR master index files.

[General_Uilities_v2.1.py](#) - module with generic utilities, in this case used for downloading files to a string or to a file.

- [Generic_Parser.py](#)

Program to generate sentiment counts for all files contained within a specified folder. Sentiment counts are based on the Loughran-McDonald dictionary.

- Python Dependencies (i.e., modules you must download that are accessed by the program):

[Load_MasterDictionary.py](#) - module to load Loughran-McDonald master dictionary.

- Data Dependencies:

[LoughranMcDonald_MasterDictionary_2014.csv](#) - data file with LM Dictionary

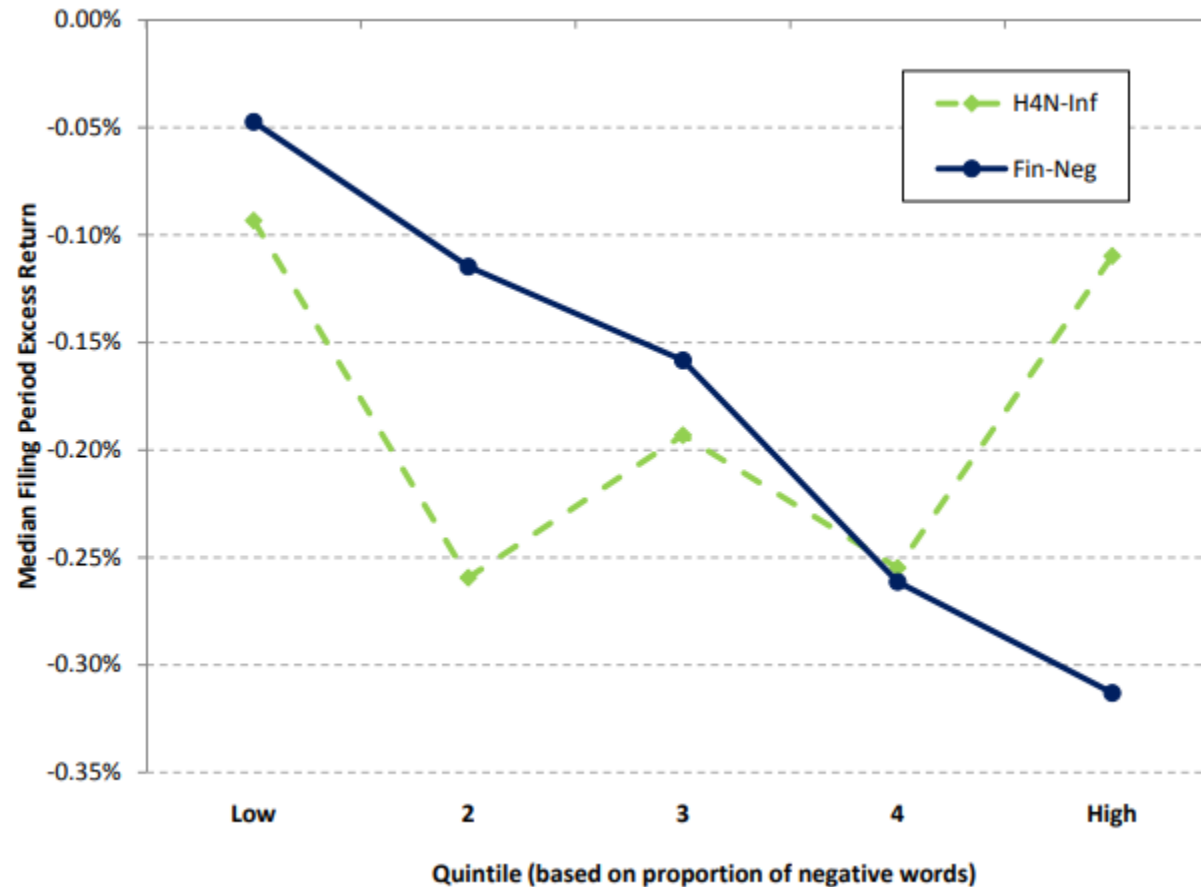
MD word list – some examples

[illegible]

Some bad words

```
> b$Word %>% sample(80) %>% sort
[1] "ABDICATIONS"      "ABNORMALITIES"    "ABROGATING"       "ADVERSARY"        "AGGRAVATIONS"
[6] "ANTITRUST"        "BOYCOTTING"       "BREAK"            "CANCEL"           "CIRCUMVENTING"
[11] "COERCES"          "CONDEMN"         "CONFESSED"        "CRIMES"           "CRITICIZING"
[16] "CULPABLE"         "DANGEROUSLY"      "DECEIVE"          "DEFAMED"          "DEFENDS"
[21] "DEGRADATIONS"     "DEGRADING"        "DEPRESSES"        "DEPRESSING"       "DETERIORATE"
[26] "DIMINISH"         "DISAGREEMENTS"    "DISAGREES"        "DISASTERS"        "DISGORGEMENT"
[31] "DISHONORABLE"     "DISLOYALLY"       "DISPARAGES"       "DISPARAGINGLY"    "DISTURBING"
[36] "DIVESTS"          "EMBEZZLEMENTS"   "EVASIVE"          "EXPLOITS"         "FALSELY"
[41] "HALT"             "HARSHLY"         "HINDERING"        "HINDRANCES"       "INACTIVATING"
[46] "INACTIVATION"     "INDICTABLE"       "INJURIOUS"        "INTENTIONAL"      "INTERRUPTS"
[51] "KICKBACKS"        "LIQUIDATING"      "MISCALCULATE"     "MISJUDGMENT"      "MISPRICING"
[56] "NEGATIVE"         "OBSTRUCTIONS"     "OPPOSITIONS"      "OVERCHARGED"      "OVERESTIMATION"
[61] "PENALTY"          "PREMATURELY"     "PREVENTION"       "RELUCTANT"        "REPARATION"
[66] "SENTENCED"        "SERIOUSNESS"     "SHORTAGES"        "SOLVENCIES"       "STOPPAGE"
[71] "SUBJECTING"       "SUFFER"          "SUSPICION"        "UNDERMINES"       "UNDERPAYMENTS"
[76] "UNDERPRODUCED"    "UNECONOMICAL"    "UNFORESEEABLE"    "UNLAWFUL"         "UNLICENSED"
> |
```


McDonald-Loughran Analysis



When a corporate filing is available on EDGAR

A useful literature review



Original Article

Textual Analysis in Accounting and Finance: A Survey

TIM LOUGHRAN, BILL MCDONALD

First published: 26 April 2016 | <https://doi.org/10.1111/1475-679X.12123> | Cited by: 34



PDF



TOOLS



SHARE

ABSTRACT

Relative to quantitative methods traditionally used in accounting and finance, textual analysis is substantially less precise. Thus, understanding the art is of equal importance to understanding the science. In this survey, we describe the nuances of the method and, as users of textual analysis, some of the tripwires in implementation. We also review the contemporary textual analysis literature and highlight areas of future research.

Seeking alpha paper

- This paper scrapes all articles on Seeking Alpha over the period 2005-2012

Seeking alpha paper

- This paper scrapes all articles on Seeking Alpha over the period 2005-2012
 - They do some analysis to detect the company + ticker
 - Does it mention a company in the first paragraph?
 - Is it only one company that it mentions?
- Then, what is the tone?
- Does the tone of the article, and of the *comments* predict future stock returns?
 - “Wisdom of crowds” story – pundits should provide useful information.
 - Moreover, disagreement by the crowd with the author should predict future negative returns

Seeking alpha paper

A “Negative” Article about Google (12 negative words, 494 total words, NegSA = 2.43%):

“Does Google Uphold ‘Do No Evil’ with shareholders?”

January 12, 2010 | about: GOOG

Author: Ravi Nagarajan (<http://seekingalpha.com/author/ravi-nagarajan>)

Article URL: <http://seekingalpha.com/article/182037-does-google-uphold-do-no-evil-with-shareholders>

As we discussed recently in an article on Ken Auletta’s new book, *Googled: The End of the World as We Know It*, the story of Google’s (GOOG) founding and astounding growth is one that has a secure place in the history books. A major part of Google’s success has been attributed to its unique way of doing business. The motto “Do No Evil” has been enshrined into Google’s core philosophy. Google has been positioned by its founders as more than just a business but as an institution that seeks to promote a better world for society.

This type of pronouncement from a corporation was always certain to bring about a great deal of skepticism. After all, Google is now a large corporation presumably seeking to maximize shareholder wealth. Or is it?

Wonderful Timing, Just Not For Shareholders

As the Wall Street Journal reminds us Monday, in early 2009 Google re-priced a large number of options at much lower strike prices. 7.6 million options with an average strike price of \$522 were exchanged for an equivalent number exercisable at \$308.57. This narrowly missed the low for the year of \$282.75. Google now trades at just under \$600.

Google’s founders were supposedly influenced by Warren Buffett when they published an “owner’s manual” shortly before Google’s IPO. It is, therefore, even more surprising that management reacted to what proved to be a temporary share price decline by massively re-pricing options at the expense of Google’s shareholders.

Seeking alpha paper

4. Commenter: joshpritchard (<http://seekingalpha.com/user/758651/comments>) (15 negative words, 287 total words, *NegSA-Comment* = 5.23%):

“@rubicon59

The trial has progressed quite a bit from then. It was broken into three phases: Copyright, Patents, then Damages. The copyright component, which would have been where Oracle had its chance to win an injunction, is over. The ruling is out. The jury found Google to have infringed by copying 9 lines of code, out of millions. That's it. They said Google did infringe the SSO of the API, but couldn't decide whether it was covered by Fair Use. Then Oracle brought a motion to the judge to decide on Fair Use as a matter of law. He rejected the motion yesterday.

The judge in the Oracle v Google case stated yesterday that the max damages for the 9 lines of Rangecheck code would be \$150K (the max for statutory damages). Now the court is on the patent phase of the trial, with only 2 of the remaining 7 patents in suit still standing. All three court experts (Oracle, Google, and the Court have all provided one) say that the combined value of the 2 patents is ~5M at most. Even if Oracle were to win 3x treble for willful infringement, you're still looking at damages <\$20M. An injunction *is* off the table. Oracle has almost definitely spent more on the case than they can hope to recoup in damages.

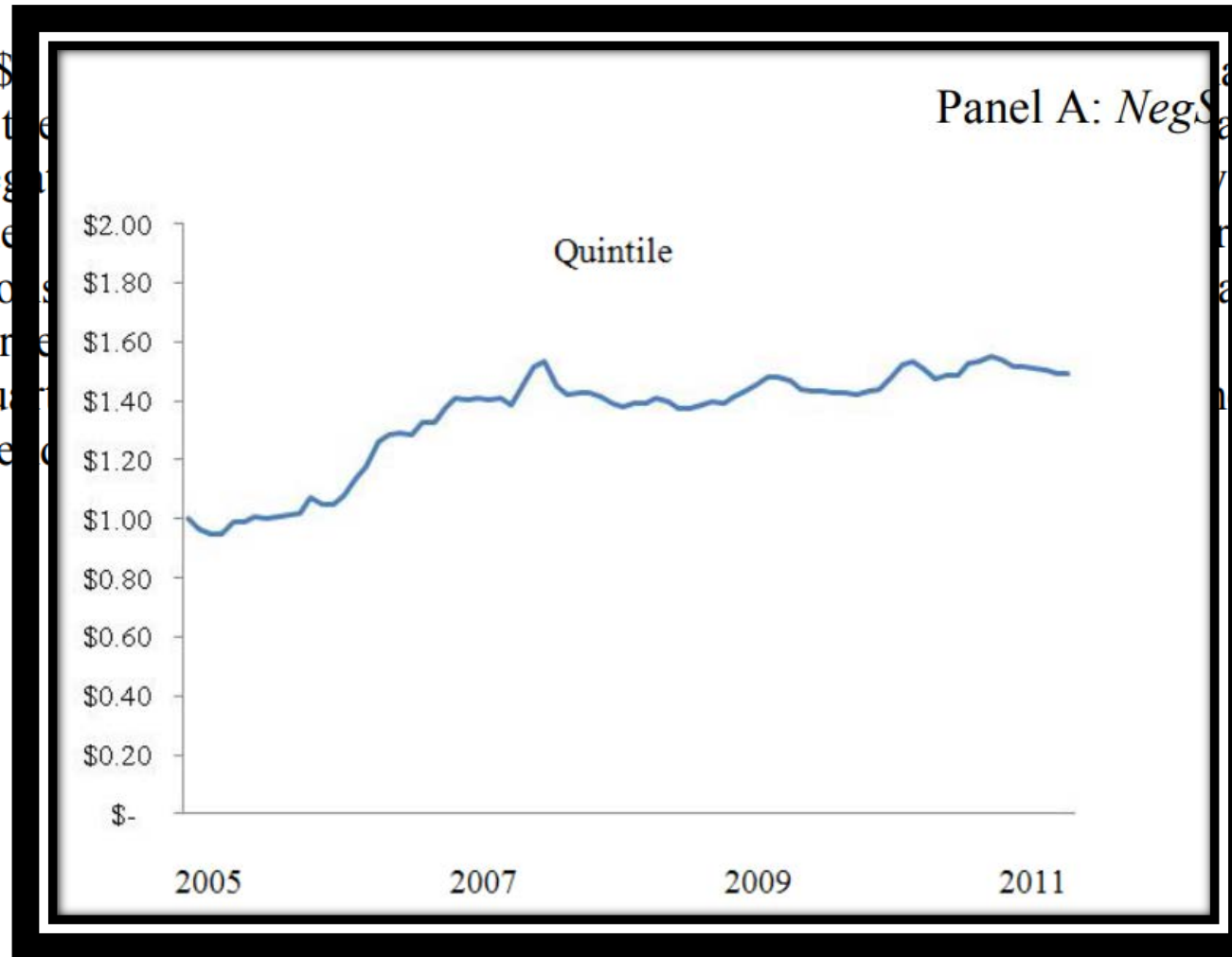
See groklaw.net for court filing and transcripts... it was just yesterday (Thurs) that Judge Alsup said explicitly that Oracle was going to get a max of \$150K in damages for the copyright infringement, and suggested they settle instead of putting before the jury (which will almost certainly be a more cost effective move for Oracle at this point, though they'd lose a lot of face for their baseless case).”

Strategy: form long / short based on tone

This figure depicts how \$1 invested in a simple calendar-time trading strategy would have evolved. The trading strategy is as follows: At the end of each trading day t , we assign stocks into quintile (quartile) portfolios based on the average fraction of negative words across all articles published on SA about company i on day t ($NegSA_{i,t}$); we also form quintile (quartile) portfolios based on the average fraction of negative words across SA comments posted over days t to $t+1$ in response to the SA articles ($NegSA-Comment_{i,t}$). We skip two days and hold each stock in its respective portfolio for three months. Based on the daily returns of a long-short portfolio, where we go long stocks in the bottom quintile (quartile) and short stocks in the top quintile (quartile), we plot how much \$1 would have grown/shrunk through calendar time.

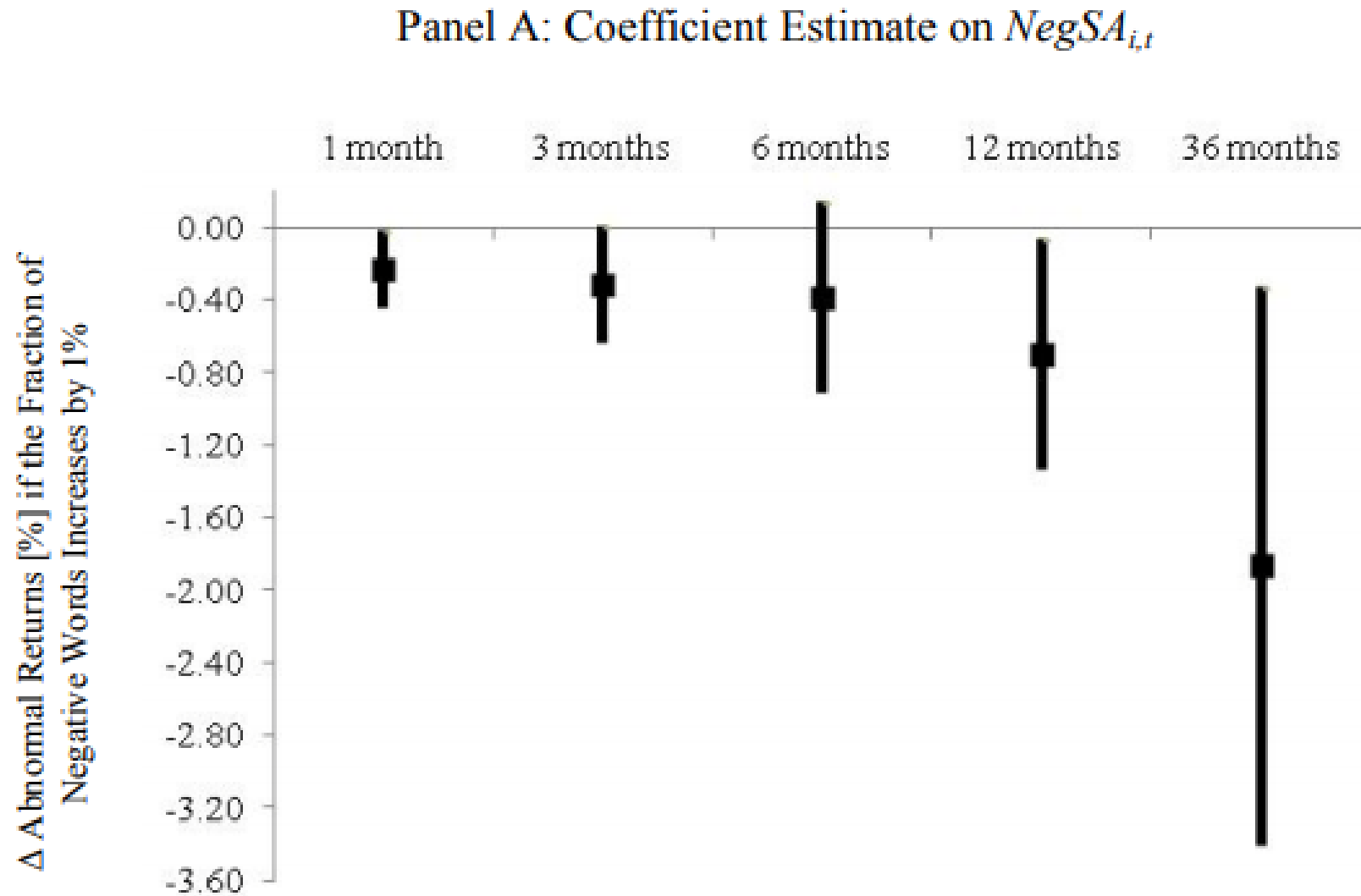
Strategy: form long / short based on tone

This figure depicts how \$ have evolved. The trading strategy is as follows: At the end of each month, we calculate the average fraction of negative sentiment comments posted on day t ($NegSA_{i,t}$); we then form quintile (quartile) portfolios based on $NegSA_{i,t}$; we go long stocks in the top quintile (quartile) and hold each stock in its respective portfolio for three months. We then repeat the process where we go long stocks in the bottom quintile (quartile) and hold each stock in its respective portfolio for three months. The figure shows how much \$1 would have



have evolved. The trading strategy is as follows: At the end of each month, we calculate the average fraction of negative sentiment comments posted on day t ($NegSA_{i,t}$); we then form quintile (quartile) portfolios based on $NegSA_{i,t}$; we go long stocks in the top quintile (quartile) and hold each stock in its respective portfolio for three months. We then repeat the process where we go long stocks in the bottom quintile (quartile) and hold each stock in its respective portfolio for three months. The figure shows how much \$1 would have

Seeking alpha paper



Sentiment analysis is a deep topic

- One of the key problems with sentiment analysis in the prior example is that it assumes all of the words should be weighted equally
- From what I understand from industry practitioners, as of today is relatively unexplored in quant equity in Asia
 - Naïve word dictionaries have yet to see their equivalent in Chinese
- But, in reality, some words matter more than others
 - How do we define which words matter? What are possible ways we could assess that?
 - Ideally, we'd need to know words we say matter, empirically do matter
- Idea: why don't we run a regression of returns on words?
 - Problem: we need instances where the words really do matter

Sentiment analysis is a deep topic

- The ideal measure ...
 - Score positively related to the number of occurrences of a positive word
 - Scores negatively related to the number of occurrences of a negative word
 - Score positively related to the strength of a positive word
 - Score negative related to the strength of a negative word
 - Score inversely related to the number of total words in the documents
- That is, the more words there are, each word should matter less

Word power – Jegadeesh and Wu (2011)

- This paper seeks to solve a few problems
 - when a document comes out, we don't know if the document is the only piece of news
 - We don't know what words are the most important
- Earnings announcement days: when a company announces earnings, we can be pretty sure it's the only event that matters
- They try to figure out : what words matter correlate the most to stock returns in-sample
 - Out of sample, can you trade?

Word power – Jegadeesh and Wu (2011)

- Create a word vector $F_{i,j}$ where F is the frequency (count) of the word j in document i
- Then score according to the weight of each word e.g. how “powerfully negative or positive” the word is
 - $\frac{\sum_{\text{across all words}} (w_j F_{i,j})}{\text{number_distinct_words}}$
 - Where w is determined from a regression
- Okay, what is this regression?

Word power – Jegadeesh and Wu (2011)

- Time period: January 1995 to December 2010
 - Take all Form 10-Ks for the year from the SEC EDGAR database
- Stock price at least \$3.00 - a bit odd choice?
 - Exclude financial firms. Why? Words such as “risk” and “casualty” may not be negative words for financial firms
 - This is common in a lot of papers, for reasons you and I may not really understand
- Do not count positive/negative words accompanied by a negator within 3 words
 - “not”, “unless”

Outcome: returns [filing date, filing date + 3]

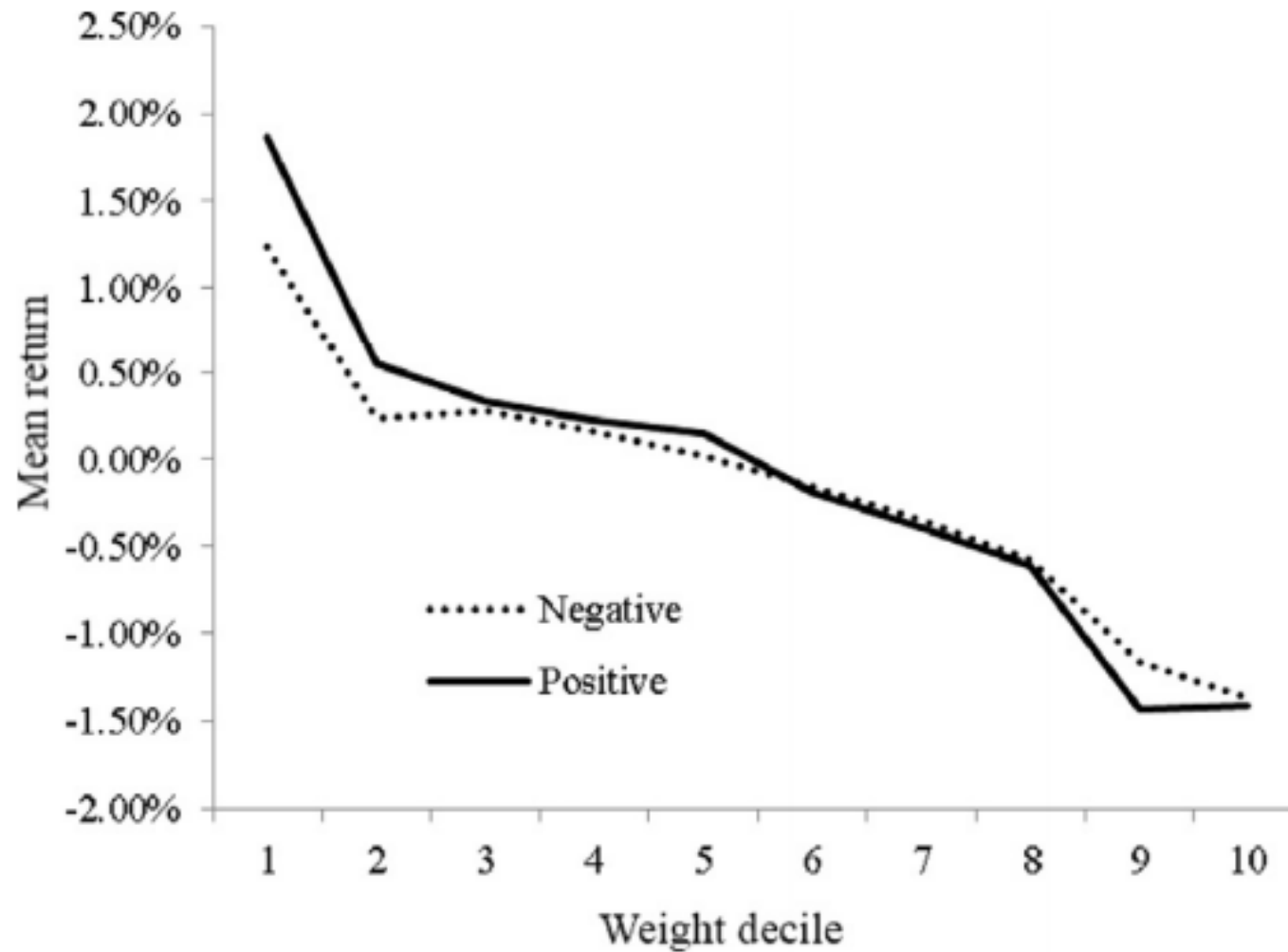
$$r_i = a + b \left(\sum_{j=1}^J (w_j F_{i,j}) \frac{1}{a_i} \right) + \epsilon_i$$
$$= a + \left(\sum_{j=1}^J (bw_j F_{i,j}) \frac{1}{a_i} \right) + \epsilon_i$$

- There are a bit more normalization schemes in the paper

Word power – Jegadeesh and Wu (2011)

			Least impactful words	
Most impactful words			WP rank	idf Rank
	WP rank	idf rank		
Panel A: Positive words				
ingenuity	1	14	lucrative	123
acclaimed	2	7	tremendous	122
influential	3	26	worthy	121
revolutionize	4	19	happy	120
optimistic	5	42	spectacular	119
enthusiasm	6	29	beautiful	118
excited	7	48	smooth	117
courteous	8	20	conducive	116
regain	9	39	receptive	115
incredible	10	3	proactive	114
Panel B: Negative words				
imperil	1	18	dispossess	718
disavow	2	22	ridicule	717
insubordination	3	20	mischief	716
bailout	4	31	derogatory	715
dismal	5	10	disorderly	714
untruthful	6	39	disassociate	713
unwelcome	7	5	immoral	712
turbulent	8	140	irreconcilable	711
vitiate	9	38	disgrace	710
undocumented	10	55	extenuating	709

Word power – Jegadeesh and Wu (2011)



- More positive + less negative (e.g. normalized to point the same direction)

Making money?

- The dependent variable is the abnormal returns computed within the “Event Windows” specified at the top of the respective columns. The independent variables in all regressions are the Word Power score calculated using lists of positive and negative words. We compute the word power weights for each year using Regression (6) and Eq. (7) over the sample period prior to the filing of 10-Ks, and compute positive and negative WP scores for each 10-K using Eq. (4). The estimates use a sample of 45,860 10-Ks over 1995 to 2010. The independent variables are standardized to a mean of 0 and standard deviation of 1.

Panel A: Positive Words

	Event Windows		
	+5 to +9	+5 to +14	+5 to +26
<i>Dependent Variable</i>			
Market-adjusted returns	0.132 (2.06)	0.200 (1.81)	0.228 (0.07)
Size-adjusted returns	0.093 (1.98)	0.123 (1.80)	0.130 (0.25)

Panel B: Negative Words

	Event Windows		
	+5 to +9	+5 to +14	+5 to 26
<i>Dependent Variable</i>			
Market-adjusted returns	0.101 (1.93)	0.132 (1.51)	0.191 (0.83)
Size-adjusted returns	0.111 (1.90)	0.127 (1.44)	0.144 (0.45)

NLP – more examples

- NLP is one of the most commonly used tools in finance
- Not just tone analysis, but many creative analyses
 - Textual similarity between two documents
 - “Topic” analysis – unsupervised
 - “Topic” analysis – supervised
 - Relatedly, “event detection”
 - Word-weightings – what matters?
 - Narrative:
 - Who are you calling on in a conference call? Are you trying to delay the revelation of good news?
 - Are you blaming other people instead of your own company’s mis-management?

NLP – more examples

- Not just tone analysis, but many creative analyses
 - Textual similarity between two documents
 - “Topic” analysis – unsupervised
 - “Topic” analysis – supervised
 - Relatedly, “event detection”
 - Narrative:
 - Who are you calling on in a conference call? Are you trying to delay the revelation of good news?
 - Are you blaming other people instead of your own company’s mis-management?

Cohen, Malloy and Quoc (2016)

- This paper focuses on public companies, and show deviations from former language or well-codified text has information about future outcomes of the firm
- Default choices / boilerplate rarely change
 - They are not top-line numbers
 - But the fact that they are changed could be meaningful
- Firms may change repetitive, seemingly harmless information
 - But such changes

Motivating example – NetApp

- American company which focuses on computer storage, big data, data visualization and overall data management

NetApp, Inc. (NTAP) [☆ Add to watchlist](#)
NasdaqGS - NasdaqGS Real Time Price. Currency in USD

34.98 -0.66 (-1.85%) **35.01** 0.03 (0.09%)
At close: 4:00 PM EDT After hours: 4:41 PM EDT

[Summary](#) [Conversations](#) [Statistics](#) [Profile](#) [Financials](#) [Options](#) [Holders](#) [Historical Data](#) [Analysts](#)

NetApp, Inc.

495 East Java Drive
Sunnyvale, CA 94089
United States
[408-822-6000](tel:408-822-6000)
<http://www.netapp.com>

Sector: **Technology**
Industry: **Data Storage Devices**
Full Time Employees: **12,030**

Key Executives

Name	Title	Pay	Exercised	Age
Mr. George Kurian	Chief Exec. Officer, Pres and Director	1.32M	N/A	49
Mr. Ronald J. Pasek	Chief Financial Officer, Principal Accounting Officer and Exec. VP	31.94k	N/A	55
Mr. Jeffrey K. Bergmann	VP of Corp. Fin.	459.98k	N/A	N/A
Mr. Matthew K. Fawcett	Sr. VP, Gen. Counsel and Corp. Sec.	670.66k	N/A	48
Mr. Joel D. Reich	Exec. VP of Product Operations	635.21k	N/A	57

Amounts are as of December 31, 2016 and compensation values are for the last fiscal year ending on that date. Pay is salary, bonuses, etc. Exercised is the value of options exercised during the fiscal year. Currency in USD.

Form 10-K Annual reports

- All companies with more than \$10 million in assets and 500 owners are required by the SEC to file a Form 10-K, which is the firm's annual report. A briefer 10-Q
- Contains 4 parts and 15 schedules

Form 10-K Annual reports

- All companies with more than \$100 million in assets must file a Form 10-K, which is the firm's annual report to the SEC.
- Contains 4 parts and 15 schedules

TABLE OF CONTENTS		
PART I		
Item 1	Business	3
Item 1A	Risk Factors	15
Item 1B	Unresolved Staff Comments	25
Item 2	Properties	25
Item 3	Legal Proceedings	25
Item 4	Mine Safety Disclosures	25
PART II		
Item 5	Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities	26
Item 6	Selected Financial Data	29
Item 7	Management's Discussion and Analysis of Financial Condition and Results of Operations	30
Item 7A	Quantitative and Qualitative Disclosures About Market Risk	49
Item 8	Financial Statements and Supplementary Data	51
Item 9	Changes in and Disagreements with Accountants on Accounting and Financial Disclosure	89
Item 9A	Controls and Procedures	89
Item 9B	Other Information	89
PART III		
Item 10	Directors, Executive Officers and Corporate Governance	90
Item 11	Executive Compensation	90
Item 12	Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters	90
Item 13	Certain Relationships and Related Transactions, and Director Independence	90
Item 14	Principal Accounting Fees and Services	90
PART IV		
Item 15	Exhibits, Financial Statement Schedules	90
	Signatures	91

Split the document into sections

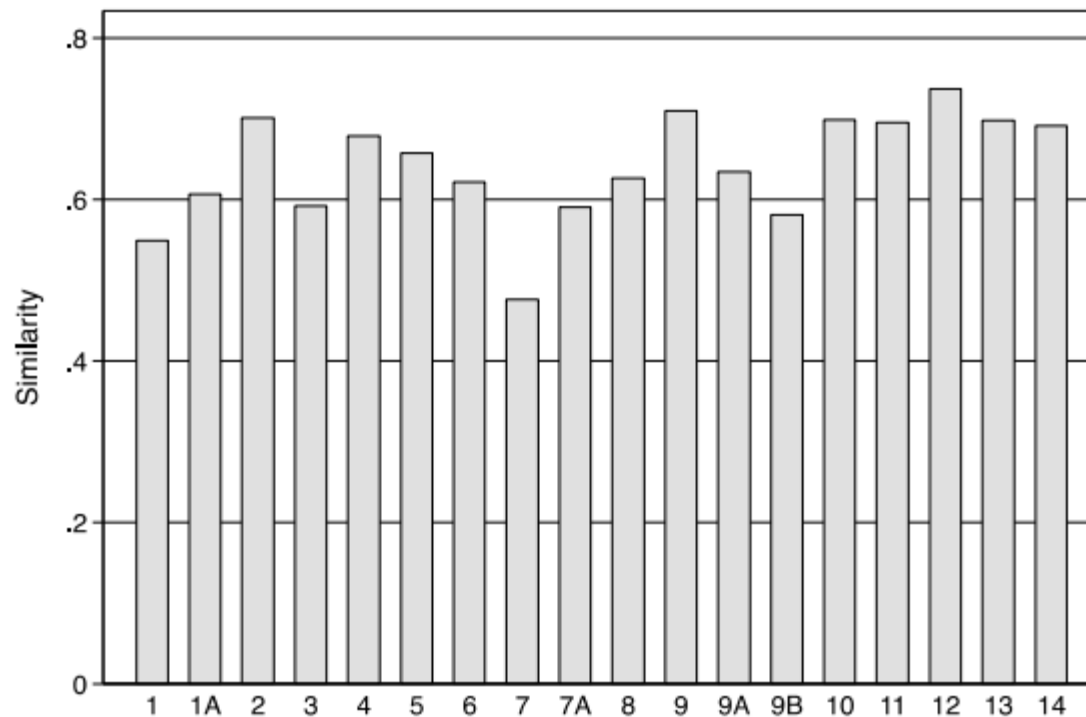
- Suppose we could split sections
- In general, section-by-section, most sections stay the same

Split the document into sections

- Suppose we could split sections
- In general, section-by-section, most sections stay the same

Figure 4: Which Sections Change the Most – 10K

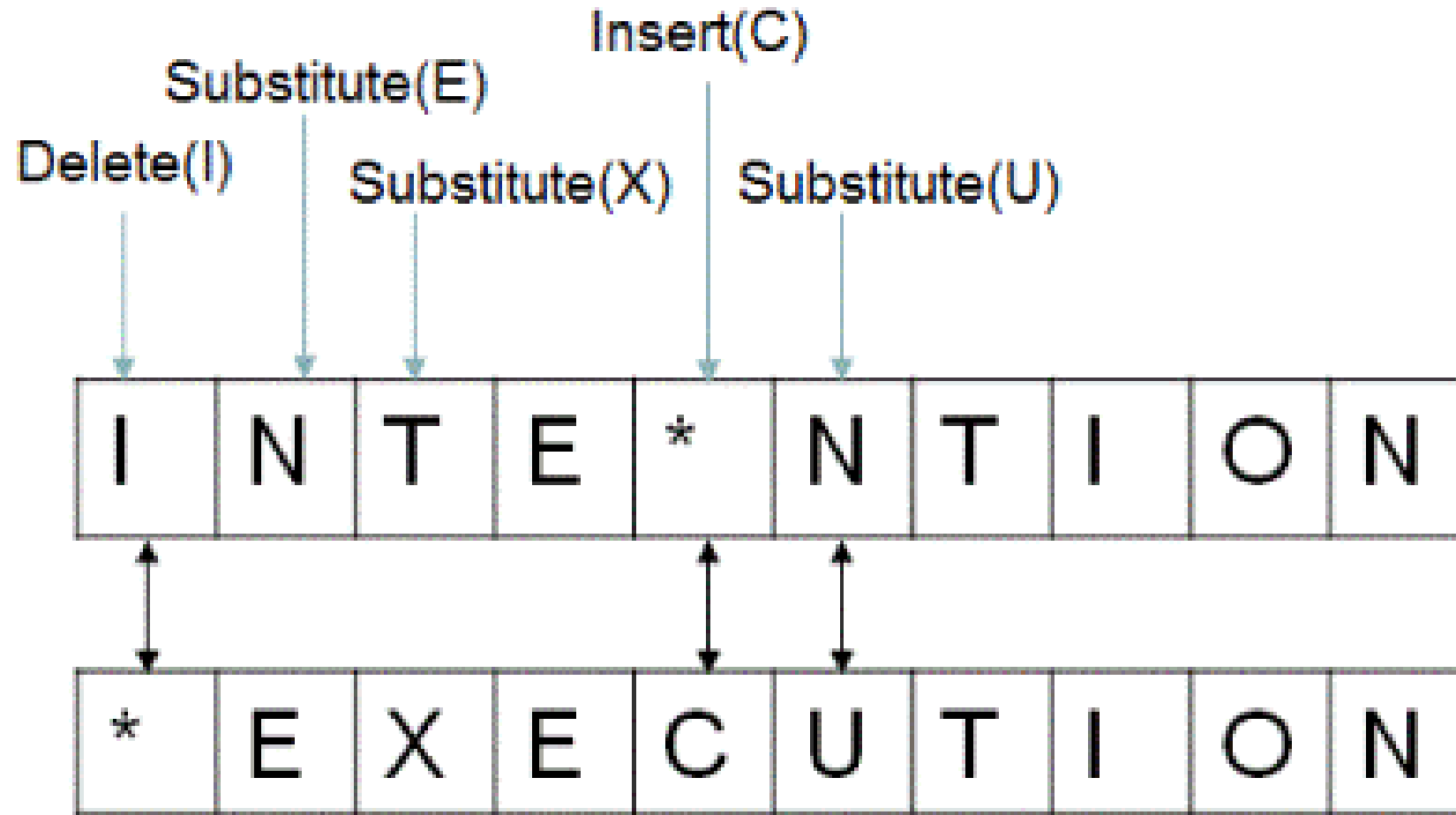
This figure reports the average similarity score for different items of firms' 10-Ks.



How do we calculate similarity?

- Cosine similarity $\frac{A*B}{||A|| ||B||} = \frac{\sum A_i B_i}{\sqrt{\sum A_i^2} \sqrt{\sum B_i^2}}$ probably most popular
- There are others, such as minimum edit distance and Jaccard distance.
 - Minimum edit distance is the number of letters necessary to go from one to the next
 - Jaccard distance

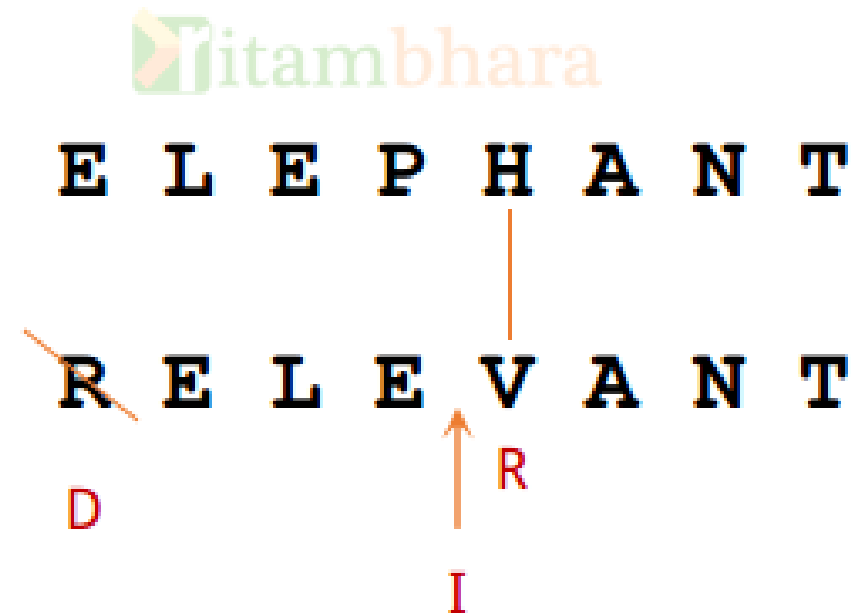
Minimum edit distance



Minimum edit distance

		E	L	E	P	H	A	N	T
	0	1	2	3	4	5	6	7	8
R	1	1	2	3	4	5	6	7	8
E	2	1	2	2	3	4	5	6	7
L	3	2	1	2	3	4	5	6	7
E	4	3	2	1	2	3	4	5	6
V	5	4	3	2	2	3	4	5	6
A	6	5	4	3	3	3	3	4	5
N	7	6	5	4	4	4	4	3	4
T	8	7	6	5	5	5	5	4	3

itambhara



Min Edit Distance: 3

How do we calculate cosine similarity?

$$\rho_{xy} = \frac{\frac{1}{n} \sum ((x_i - \bar{x}) * (y_i - \bar{y}))}{\sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} * \sqrt{\frac{1}{n} \sum (y_i - \bar{y})^2}}$$

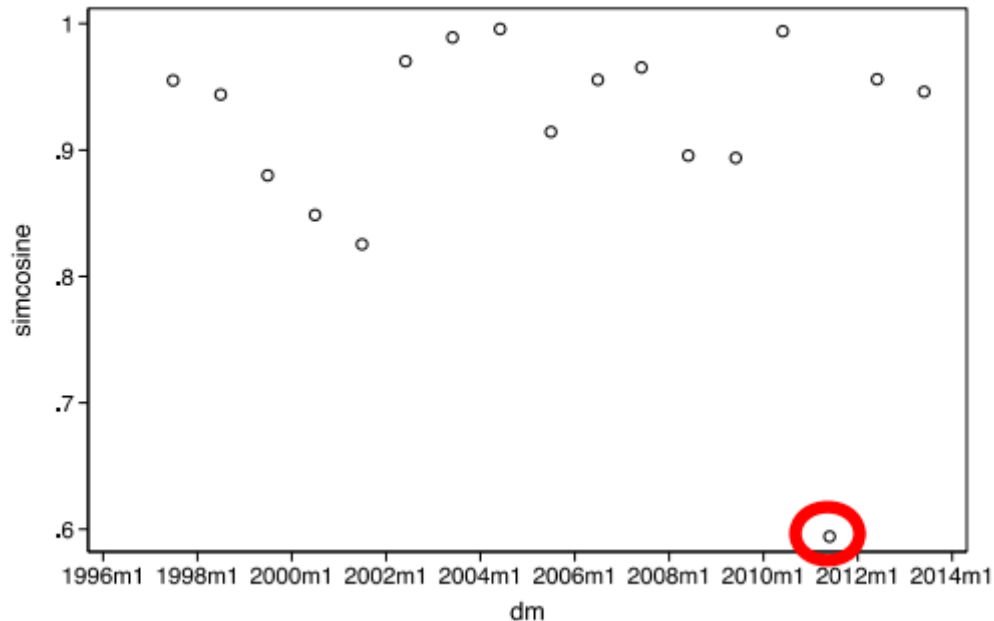
$\frac{1}{n}$ can be dropped

$$\rho_{xy} = \frac{\sum ((x_i - \bar{x}) * (y_i - \bar{y}))}{\sqrt{\sum (x_i - \bar{x})^2} * \sqrt{\sum (y_i - \bar{y})^2}}$$

$$\cos(\theta) = \frac{\sum x_i * y_i}{\sqrt{\sum x_i^2} * \sqrt{\sum y_i^2}}$$

Split the document into sections

- Suppose we could split sections
- In general, section-by-section, most sections stay the same
- In the case of NetApp... something happened in 20011



Six months later

- NetApp flooded the news with links to §

- 1) **November 3rd, 2011:** "Syria Crackdown Gets Italy Firm's Aid With U.S.-Europe Spy Gear", reported that Syrian intelligence agents have contracted Area SpA, an Italian surveillance company, to complete a highly sophisticated system that tracks Internet activity using NetApp equipment.

<http://www.bloomberg.com/news/articles/2011-11-03/syria-crackdown-gets-italy-firm-s-aid-with-u-s-europe-spy-gear>

- 2) **November 9th, 2011:** "NetApp Role in Syria Spy Project Spurs Demands for U.S. Inquiry"

<http://www.bloomberg.com/news/articles/2011-11-10/netapp-role-in-syria-spy-project-spurs-demands-for-u-s-inquiry>

→ *Senators Mark Kirk, a Republican from Illinois, and Robert Casey, a Democrat from Pennsylvania, will send a letter today to the State and Commerce departments requesting an investigation into two U.S. companies whose technology has been used to "monitor activities of Syrian citizens," according to a draft of the letter. One of the companies is NetApp, whose role in the Internet surveillance system was detailed in a Nov. 3 article by Bloomberg News.*

In addition, Representative James McGovern, a Democrat from Massachusetts and co-chairman of the Tom Lantos Human Rights Commission in the House, said he has instructed his staff to follow up with government agencies regarding NetApp to make sure U.S. sanctions against Syria are being enforced.

"I find it unconscionable that a U.S.-based company's technology is being sent to Syria to help spy on peaceful citizens," McGovern said.

In their letter, Senators Kirk and Casey ask that pending conclusion of an investigation, officials consider suspending all U.S. government work with NetApp, which received more than \$111 million in U.S. contracts since 2001.

- 3) **November 14th, 2011:** Senators Mark Kirk (R-IL), Robert Casey (D-PA) and Christopher Coons (D-DE) sent the following letter to the secretary of state and secretary of commerce, asking the administration officials to look into the matter:

<http://www.casey.senate.gov/newsroom/releases/casey-urges-administration-to-investigate-companies-allegedly-aiding-syrian-regime>

So what happened?

- Newspapers allege that the Syrian government had been using
- NetApp equipment to conduct intelligence-gathering and surveillance activities.
- Equipment was not directly purchased from NetApp, but rather acquired through a Italian re-seller (Area SpA)
- Prior to this, the US government had placed sanctions on doing business with the Syrian government.
- The journalists find evidence (including e-mails) prior to 2011, before the filing of NetApp's 10-K, showing that NetApp knew of the involvement.
- The U.S. government uses NetApp for government work as well, and prominent senators asked the government to consider suspending all contracts with NetApp.

Textual similarity

- Event A: NetApp drastically changes the wording of their annual reports.
- Event B: Six months later, their relation to business dealings with a sanctioned foreign government is exposed.
- Was it a coincidence, or were these events linked? Did something about changes to the Form 10-K hint at the portending government inquiry?
- Let's see how the language of the annual reports changed.

Textual similarity

Panel A:

2010 (Old)

The failure to comply with U.S. government regulatory requirements could subject us to fines and other penalties, which could have a material adverse effect on our revenues, operating results and financial position.

2011 (New)

Failure to comply with U.S. government regulatory requirements **by us or our reseller partners** could subject us to fines and other penalties, which could have a material adverse effect on our revenues, operating results and financial position.

Panel B:

2010 (Old)

We are a party to lawsuits in the normal course of our business, including our ongoing litigation with Sun Microsystems which was recently acquired by Oracle Corporation. Litigation can be expensive, lengthy and disruptive to normal business operations. Moreover, the results of complex legal proceedings are difficult to predict. An unfavorable resolution of a particular lawsuit could have a material adverse effect on our business, operating results, or financial condition.

2011 (New)

We may be a party to lawsuits and other claims in the normal course of our business from time to time, including intellectual property, commercial, product liability, employment, class action, whistleblower and other litigation and claims, and **governmental and other regulatory investigations and proceedings**. Litigation can be expensive, lengthy and disruptive to normal business operations. Moreover, the results of complex legal proceedings are difficult to predict. An unfavorable resolution of a particular lawsuit could have a material adverse effect on our business, operating results, or financial condition.

Textual similarity

Panel C:

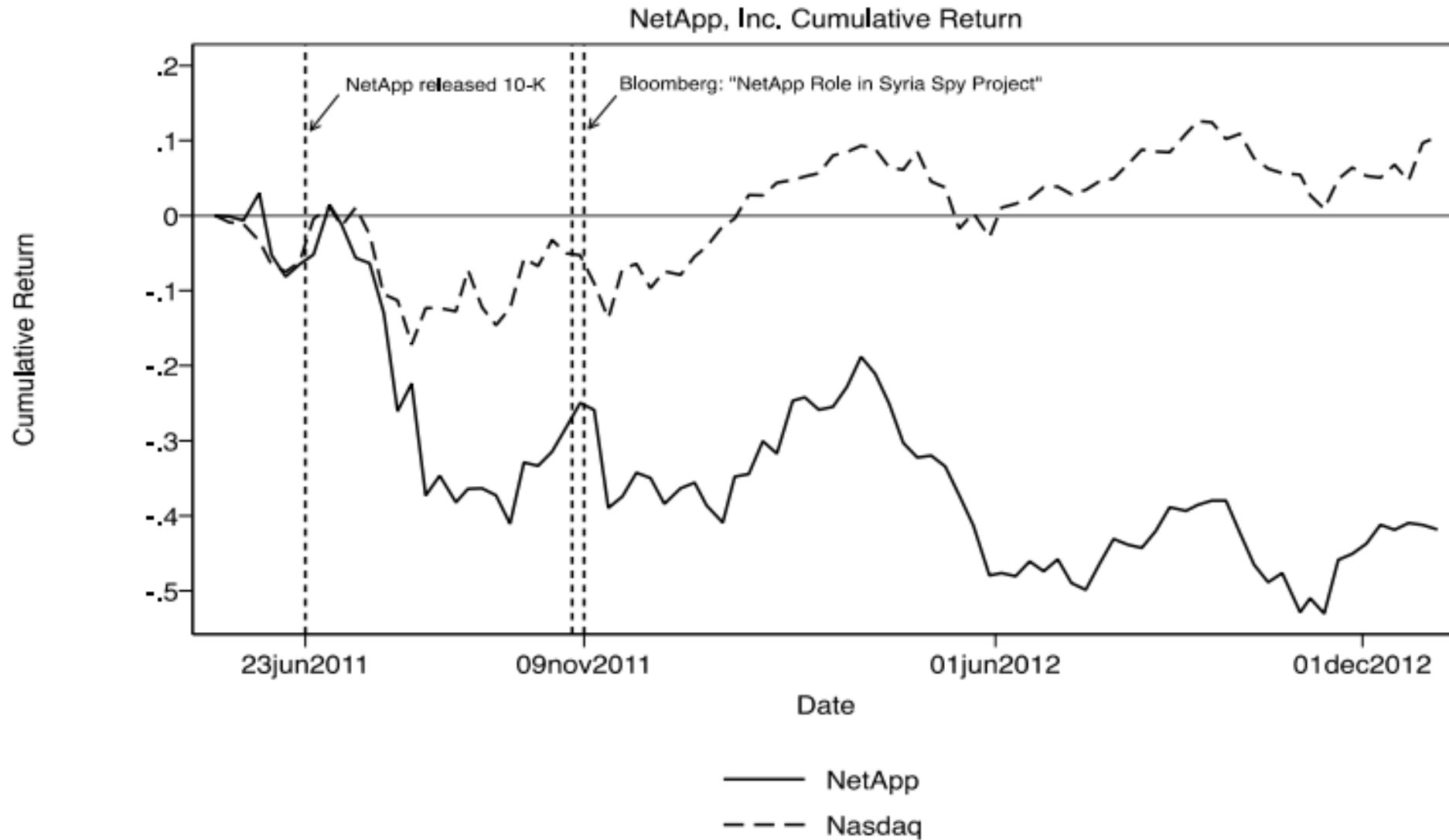
2010 (Old)

The U.S. government has contributed to our revenue growth and has become an important customer for us. Future revenue from the U.S. government is subject to shifts in government spending patterns. A decrease in government demand for our products could materially affect our revenues. In addition, our business could be adversely affected as a result of future examinations by the U.S. government.

2011 (New)

The U.S. government has contributed to our revenue growth and has become an important customer for us. Future revenues from the U.S. government are subject to shifts in government spending patterns. A decrease in government demand for our products could materially and adversely affect our revenues. In addition, our business could be adversely affected by claims that we or a channel partner have failed to comply with regulatory and contractual requirements applicable to sales to the U.S. government.

Textual similarity



Another example

Figure 1: Example Schweitzer-Mauduit International Similarity Score

This figure plots the similarity score of Schweitzer-Mauduit International annual reports (10-Ks) from 2001 to 2009.

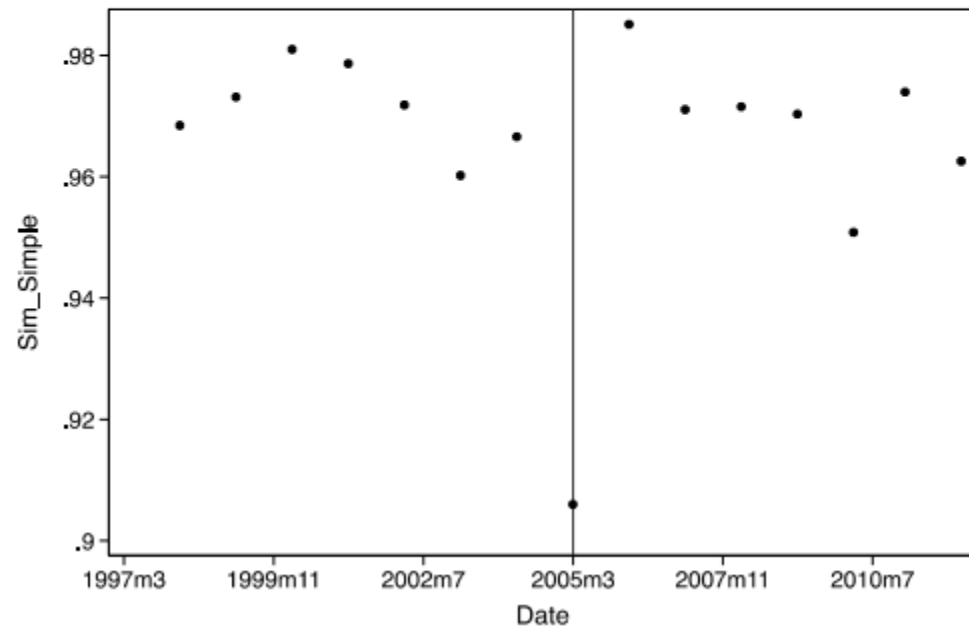
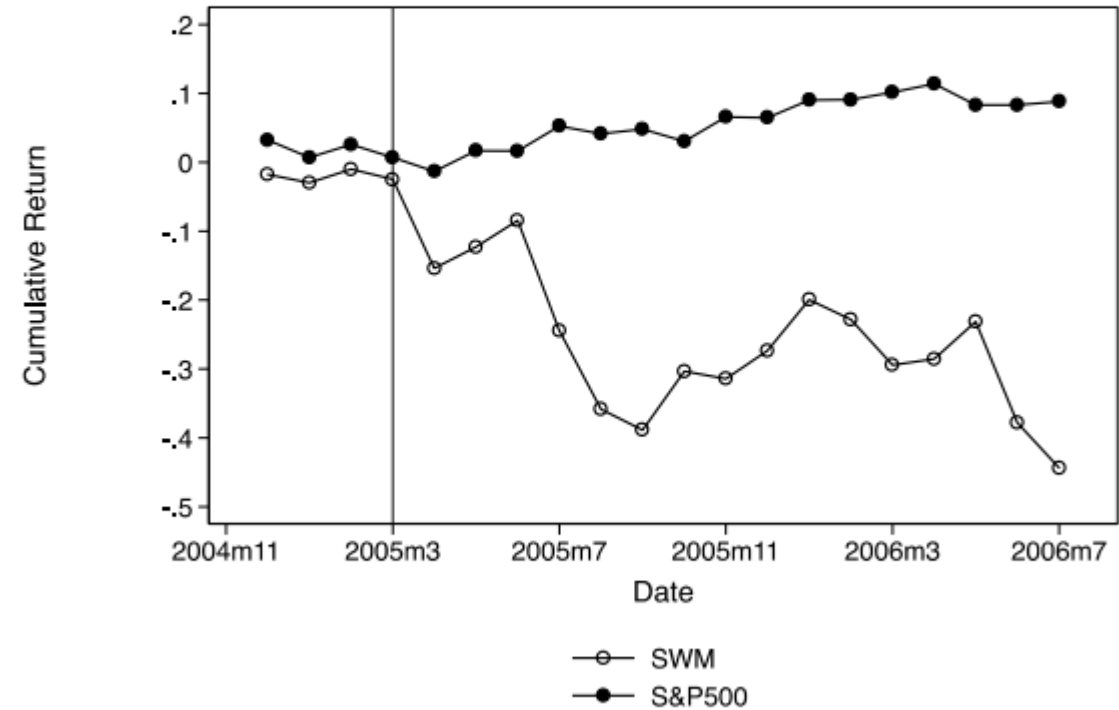


Figure 2: Schweitzer-Mauduit International Cumulative Return



Another example

This table shows the first few paragraphs that are taken from Item 7, "Management's Discussion and Analysis of Financial Condition and Results of Operations", for Schweitzer-Mauduit International's (NYSE:SWM) 2004 and 2005 10-K reports. The new discussion in the 2005 10-K is highlighted.

10-K 2005

Outlook

Consistent with recent historical trends, worldwide cigarette consumption is expected to increase at a rate of approximately one-half to one percent per year. The anticipated decline in the production of cigarettes in developed countries is expected to be more than offset by increased cigarette production in developing countries that currently represent approximately 70 percent of worldwide cigarette production. Age demographics and expected increases in disposable income are expected to support the increased consumption of cigarettes in developing countries. In addition, the litigation environment is different in most foreign countries compared with the United States, having less of an impact on the pricing of cigarettes, which, in turn, affects cigarette consumption. Cigarette production in the United States is expected to continue to decline as a result of a decline in domestic cigarette consumption caused by increased cigarette prices, health concerns and public perceptions. As well, cigarette consumption has declined in France and Germany following recent tax increases on cigarette sales in those countries.

We are experiencing weakness in our tobacco-related paper sales in western Europe caused by reduced cigarette consumption in several large European markets and new cigarette paper manufacturing capacity that was added in western Europe in mid-2004. This is expected to result in increased cigarette paper machine downtime in France in 2005.

In developing countries, there is a trend toward consumption of more sophisticated cigarettes, which utilize higher quality tobacco-related papers, such as those we produce, and reconstituted tobacco leaf. This trend toward more sophisticated cigarettes reflects increased governmental regulations concerning tar delivery levels and increased competition from multinational cigarette manufacturers.

Based on these trends, we expect worldwide demand for our products to continue to increase, with a shift from developed countries to developing countries. As a result, we are increasing some of our production capacity in developing countries such as Brazil, Indonesia and the Philippines.

The new RTL production line added at our Spay, France mill, which started up in the fourth quarter of 2003, is expected to continue to contribute positively to sales volumes and operating profit in 2005.

10-K 2004

Outlook

The markets for the Company's products are expected to remain relatively stable during 2004. Trends of improvement are expected to continue in tobacco-related paper sales in several key markets. Cigarette production in the United States continues to decline as a result of declines in domestic cigarette consumption and exports of cigarettes manufactured in the United States. The anticipated decline in the production of cigarettes in developed countries is expected to be more than offset by increased cigarette production in developing countries.

The new RTL production line added at the Company's Spay, France mill, which started up in the fourth quarter of 2003, is expected to be a major contributor to increased operating profit in 2004 compared with 2003. The new RTL production line is expected to achieve end of curve production rates by the end of the second quarter of 2004. The acquisition of a tobacco-related papers manufacturer in Indonesia that was completed in February 2004 is also expected to have a favorable impact on operating profit in 2004.

The Company did not have significant production or sale of banded or print banded cigarette papers during 2003. The Company continues to work with its customers in their development of papers for reduced ignition propensity cigarettes. In December 2003, the State of New York announced the adoption of final regulations for reduced ignition propensity cigarettes. The cigarette fire safety standard requires that all cigarettes sold in the State of New York as of June 28, 2004 have reduced ignition propensity properties. The regulations do contain a provision that allows wholesalers and retailers to transition their existing inventories. As a result of the new fire safety standards in the State of New York, the Company expects increased sales of reduced ignition propensity cigarette papers during 2004. These reduced ignition propensity papers sell for a higher price than the conventional cigarette papers they replace and are expected to have a positive impact on the Company's financial results. Since the State of New York only represents approximately ten percent of U.S. cigarette consumption and the regulations will only be in effect for one-half of 2004, the favorable impact on the Company's financial results is not expected to be significant in 2004.

Selling prices for the Company's tobacco-related products are expected to remain relatively stable during 2004. The recent weakening of the U.S. dollar versus the euro and certain other foreign currencies and higher wood pulp costs could enable the Company to implement selective selling price increases.

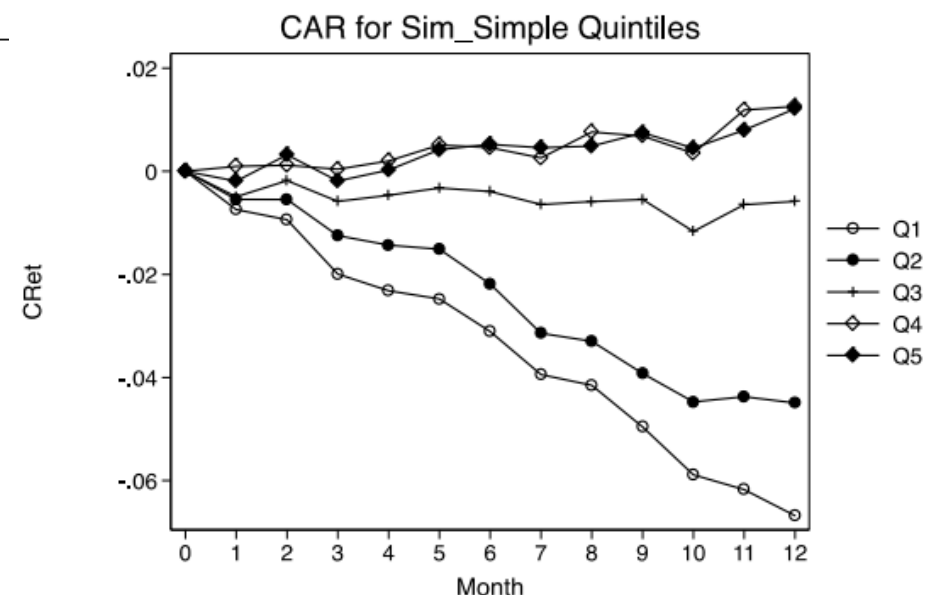
Strategy returns (~4.1% annualized)

Panel B: Value Weighted

Sim_Cosine	Q1	Q2	Q3	Q4	Q5	Q5 - Q1
Excess Return	0.0040 (1.2095)	0.0044 (1.3085)	0.0051 (1.6391)	0.0079** (2.5627)	0.0078** (2.3629)	0.0038*** (2.7547)
3-Factor Alpha	-0.0018** (-2.0280)	-0.0019** (-2.1017)	-0.0007 (-0.7910)	0.0018** (1.9748)	0.0019* (1.7411)	0.0037*** (2.7024)
5-Factor Alpha	-0.0013 (-1.4101)	-0.0021** (-2.2624)	-0.0009 (-1.0640)	0.0021** (2.3542)	0.0021* (1.9115)	0.0034** (2.3996)

Figure 3: Long term no reversals

This figure shows the average cumulative abnormal return for each quintile portfolio sorted based on firms' similarity score, for 1 month to 12 months after portfolio formation.



Which sections matter most?

Panel A: Equally Weighted

	Sim_Cosine			Sim_Jaccard		
	Excess Return	3-Factor Alpha	5-Factor Alpha	Excess Return	3-Factor Alpha	5-Factor Alpha
Management's Discussion and Analysis	0.0013 (1.5648)	0.0011* (1.6579)	0.0012* (1.6751)	0.0021** (2.5054)	0.0022*** (3.1451)	0.0020*** (2.8061)
Legal Proceedings	0.0036** (2.2428)	0.0037*** (3.0939)	0.0033*** (2.6989)	0.0028 (1.5729)	0.0030** (2.3602)	0.0025* (1.9341)
Quantitative and Qualitative Disclosures About Market Risk	0.0069*** (2.7465)	0.0068*** (2.6923)	0.0068*** (2.6481)	0.0020** (2.3738)	0.0021*** (2.9594)	0.0019*** (2.6049)
Risk Factors	0.0114 (1.6111)	0.0118 (1.6308)	0.0118 (1.6365)	0.0143** (2.1325)	0.0144** (2.4497)	0.0188*** (2.7601)
Other Information	0.0020 (1.0839)	0.0027 (1.4684)	0.0036* (1.9179)	0.0031* (1.7849)	0.0037** (2.1854)	0.0040** (2.2959)
	Sim_MinEdit			Sim_Simple		
	Excess Return	3-Factor Alpha	5-Factor Alpha	Excess Return	3-Factor Alpha	5-Factor Alpha
Management's Discussion and Analysis	0.0018* (1.9519)	0.0022*** (3.1616)	0.0019*** (2.6652)	0.0019*** (2.6673)	0.0019** (2.5405)	0.0017** (2.3253)
Legal Proceedings	0.0022 (1.2706)	0.0025** (2.3030)	0.0022* (1.9347)	0.0013 (0.8157)	0.0016 (1.4119)	0.0012 (1.1042)
Quantitative and Qualitative Disclosures About Market Risk	0.0016 (1.1822)	0.0023* (1.7374)	0.0022* (1.6712)	0.0013 (0.1581)	0.0011 (0.1319)	0.0007 (0.0801)
Risk Factors	0.0102 (1.1928)	0.0185*** (2.7728)	0.0138** (2.1663)	0.0125* (1.9310)	0.0154** (2.1914)	0.0177** (2.1156)
Other Information	0.0009 (0.5773)	0.0014 (0.9649)	0.0016 (1.0514)	0.0022 (1.2731)	0.0026** (2.3091)	0.0022* (1.9525)

summary

- Forming a long short portfolio that buys stocks with more similar portfolios... but which shorts stocks of firms with large changes in their reports
 - Earns significant excess returns
- Performance of strategy cannot be explained by typical risk factors
- Legal and Risk disclosures more informative about future stock return performance

Hoberg & Phillips

- One of the key challenges of quantitative analysis is defining an industry
 - After all,
- Hoberg and Phillips have a series of papers premised on extracting product descriptions from company filings
- Find the company description
 - Clean it of common words
 - Find all unique words that are product-related
 - Find how textually similar these product descriptions are

Hoberg & Phillips – product market fluidity + textual similarity

- Hoberg and Phillips have a series of papers premised on extracting product descriptions from company filings

Example product description

Business Organization

The Company manages its business primarily on a geographic basis. Accordingly, the Company determined its reportable operating segments, which are generally based on the nature and location of its customers, to be the Americas, Europe, Greater China, Japan, Rest of Asia Pacific and Retail. The Americas segment includes both North and South America. The Europe segment includes European countries, as well as India, the Middle East and Africa. The Greater China segment includes China, Hong Kong and Taiwan. The Rest of Asia Pacific segment includes Australia and Asian countries, other than those countries included in the Company's other operating segments. The results of the Company's geographic segments do not include the results of the Retail segment. Each operating segment provides similar hardware and software products and similar services. Further information regarding the Company's operating segments may be found in Part II, Item 7 of this Form 10-K under the subheading "Segment Operating Performance," and in Part II, Item 8 of this Form 10-K in the Notes to Consolidated Financial Statements in Note 11, "Segment Information and Geographic Data."

Products

iPhone

iPhone is the Company's line of smartphones that combines a phone, music player and internet device in one product, and is based on Apple's iOS Multi-Touch™ operating system. iPhone has an integrated photo and video camera and photo library app, and on qualifying devices, also includes Siri®, a voice activated intelligent assistant. iPhone works with the iTunes Store, the App Store and iBooks Store for purchasing, organizing and playing music, movies, TV shows, podcasts, books and apps. In addition to apps delivered with iOS, free downloads of iLife® and iWork® apps for iOS are available with all new iPhones. iPhone is compatible with both Mac and Windows personal computers and Apple's iCloud services which provide synchronization of mail, contacts, calendars, apps, music, photos, documents and more across users' devices. In September 2014, the Company introduced iPhone 6 and iPhone 6 Plus that feature larger 4.7-inch and 5.5-inch Retina® HD displays and support for Apple Pay.

iPad

iPad is the Company's line of multi-purpose tablets based on Apple's iOS Multi-Touch operating system, which includes iPad Air™ and iPad mini™. iPad has an integrated photo and video camera and photo library app, and on qualifying devices, also includes Siri. iPad works with the iTunes Store, the iBooks Store and the App Store for purchasing, organizing and playing music, movies, TV shows, podcasts, books and apps. In addition to apps delivered with iOS for qualifying devices, iLife and iWork apps for iOS are available as free downloads with all new iPads. iPad is compatible with both Mac and Windows personal computers and Apple's iCloud services. In October 2014, the Company introduced iPad Air 2 and iPad mini 3 that feature a Retina display, Touch ID™ and support for Apple Pay.

Mac

Mac is the Company's line of desktop and portable personal computers. Macs feature Intel microprocessors, the OS X operating system and include Mail, Safari® web browser, Messages, Calendar, Reminders, Contacts and the iLife apps. The Company's iWork apps are also available as free downloads with all new Macs. The Company's desktop computers include iMac®, Mac Pro® and Mac mini. The Company's portable computers include MacBook Pro®, MacBook Pro with Retina display and MacBook Air®. In October 2014, the Company introduced the 27-inch iMac with Retina 5K display with improved performance, power efficiency and visual quality; and updated the Mac mini.

iPod

The Company's iPod line of portable digital music and media players includes iPod touch, iPod nano® and iPod shuffle®. All iPods work with iTunes® to purchase and synchronize content. iPod touch, based on the Company's iOS Multi-Touch operating system, is a flash-memory-based iPod with an integrated photo and video camera and photo library app, and also includes Siri. iPod touch works with the iTunes Store, the App Store and the iBooks Store for purchasing and playing music, movies, TV shows, podcasts, books and apps. In addition to apps delivered with iOS, iLife and iWork apps for iOS are available as free downloads for all new iPod touch products. iPod touch is compatible with both Mac and Windows personal computers and Apple's iCloud services.

Hoberg & Phillips – product market fluidity + textual similarity

- Hoberg and Phillips have a series of papers premised on extracting product descriptions from company filings
- Find the company description
 - Clean it of common words
 - Find all unique words that are product-related
 - Find how textually similar these product descriptions are

Cross-momentum strategy?

- I was playing around with this a few weeks ago, but think about it like your product market peers
- Trade the following stocks long / short if they are high / low on ...

$$\text{➤ } \sum_i^{\text{peers}} r_{i,t-1}$$

Approach : predictive modeling

- Machine learning can be used to improve a regression
- Examples of these methodologies include LASSO, Ridge regression, elastic net, which is a combination of the two
- Basic idea is: we have 5000 variables,
 - With lots of variables, we can fit data really well
 - how do we pick something that will work out of sample?
- Just having data does not help – the data needs to be economically meaningful

Suggested reading

- For practice: Kaggle
- For academic reading
 - “Taming the Factor Zoo: A Test of New Factors” by Guanhao Feng (City U), Stefano Giglio at Yale and Dacheng Xiu at Chicago Booth **tries to run a horse race of all the factors we developed in the academic literature – in particular, gross profitability that we discussed earlier – do very well**
 - “Empirical Asset Pricing via Machine Learning” by Shihao Gu, Kelly and Xiu (Booth, Yale and Booth) **compares different methodologies, says Neural Nets are probably the best**
- For class: take Kurt Ye Luo’s class

OLS vs LASSO

- The basic idea of OLS is to run a simple regression:

- $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \epsilon$

- In order to minimize the sum of squared errors

- $\min_{\beta_0, \beta} \frac{1}{n} \sum_{i=1}^N (y_i - \beta_0 - X_i^T \beta)^2$

OLS vs LASSO

➤ OLS

$$➤ \min_{\beta_0, \beta} \frac{1}{n} \sum_{i=1}^N (y_i - \beta_0 - X_i^T \beta)^2$$

➤ Shrinkage method:

- Try to fit the regression using all available p predictors, but constraint the algorithm from choosing them all by a penalty

➤ Lasso minimizes the sum of squared errors but adds a penalty term

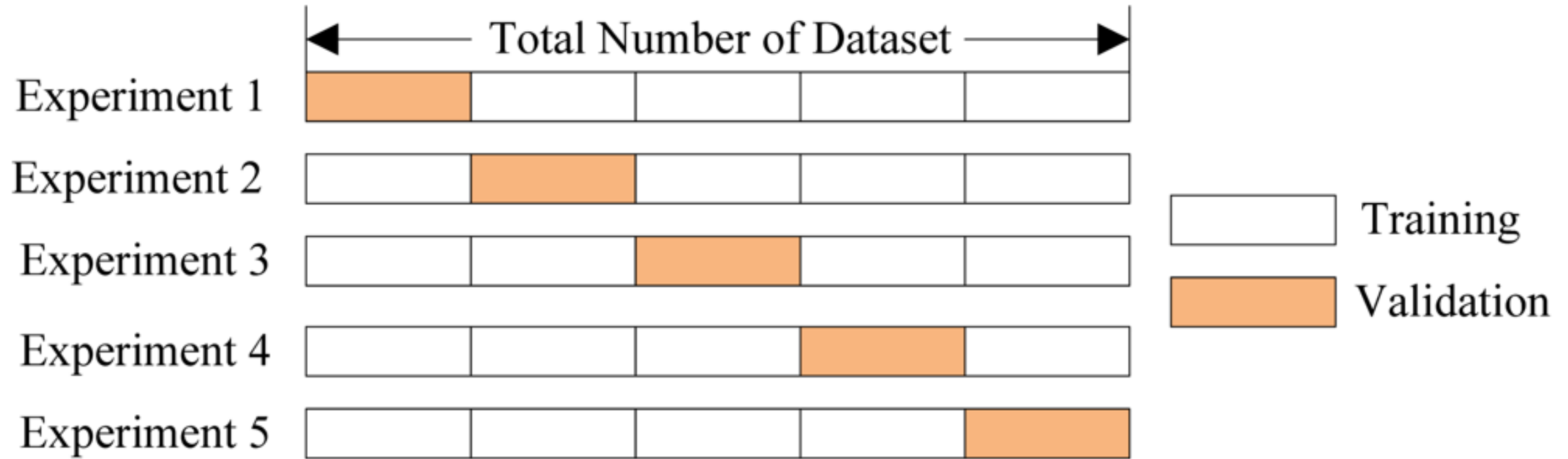
$$➤ \min_{\beta_0, \beta} \frac{1}{n} \sum_{i=1}^N (y_i - \beta_0 - X_i^T \beta)^2 + \underbrace{\lambda \sum_{j=1}^p |\beta_j|}_{\text{shrinkage term}}$$

- The absolute term means that we get penalized for each variable we add by the amount λ
- $\lambda = 0 \Leftrightarrow$ the ordinary least squares
- $n=1$ is lasso, $n=2$ is ridge,

How do we decide the penalty parameter?

- We try different values of λ and see what produces the best result
- The problem of overfitting:
 - We cut the data into K-folds
- K-fold cross-validation is one method of λ selection
- We leave some data out and test out-of-sample
 - We cut the data into K-folds
 - Run the data on the training set, test on the validation set, record the mean-squared error
 - The λ with the best result (mean-squared-error, or mean absolute error say) wins

K-fold cross validation visualized



Simple LASSO code

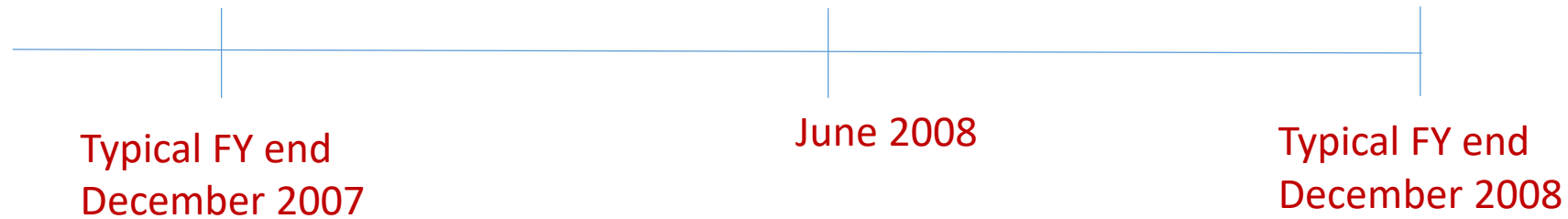
- Let's discuss some simple code

What is the value strategy?

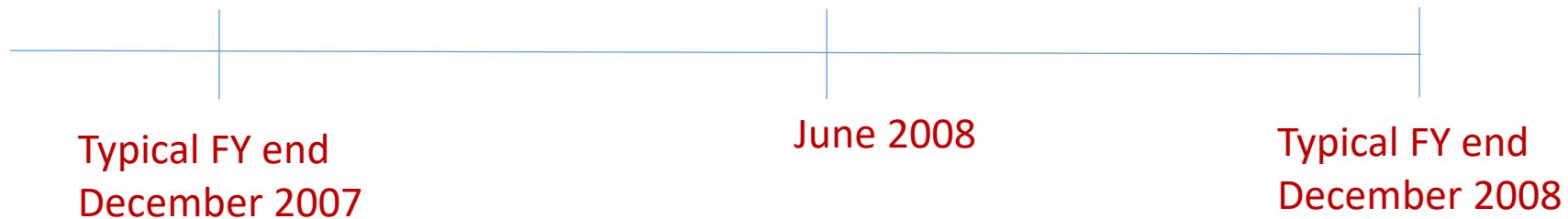
- We buy stocks that, based on the market value of assets relative to accounting value of assets, appear “cheap”
- While the value investor uses a “fundamental” or “intrinsic” value a quant essentially assumes what the market gives them
- There are many formulations of the value strategy
 - Price to earnings
 - Price to dividend
- Based on tests in which we’ve tried many measures, the best one appears to be book to market
 - $\frac{PRC * SHROUT}{BOOK\ VALUE} = \frac{MARKETCAP}{BOOKVALUE}$

Variation 1 – use recent market cap

Use book value from here
Use market cap from here

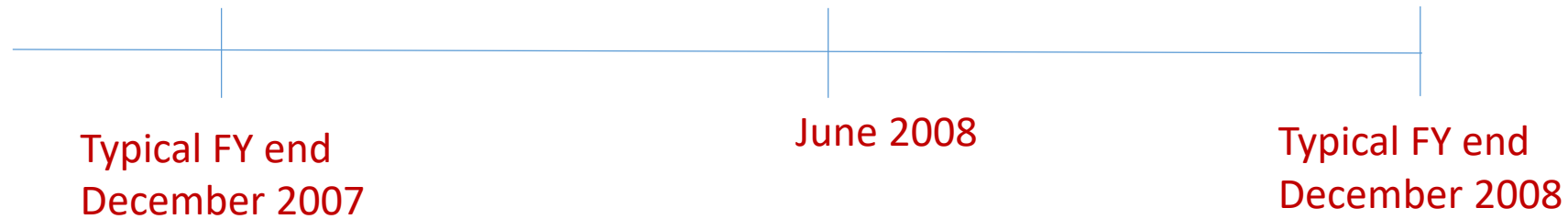


Use book value from here *Use market cap from here*



Variation 2 – rebalance more frequently

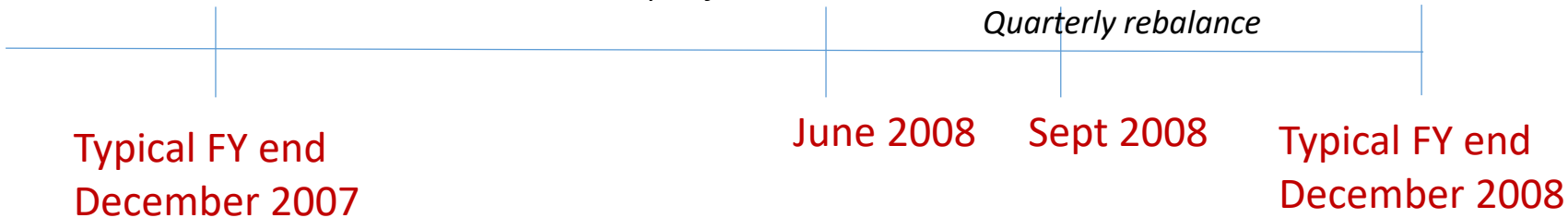
Use book value from here
Use market cap from here



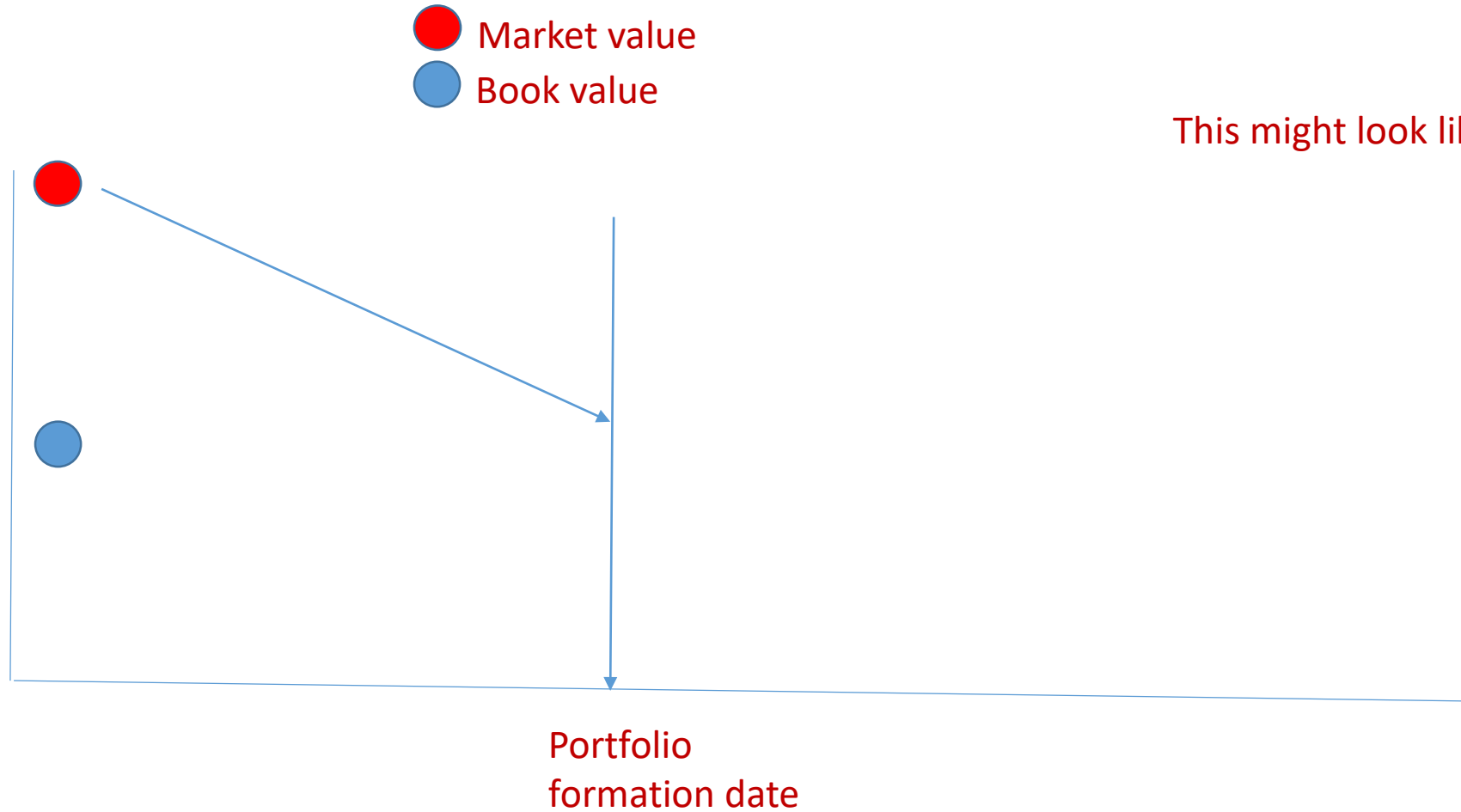
Use book value from here

*Use market cap
from here before
portfolio*

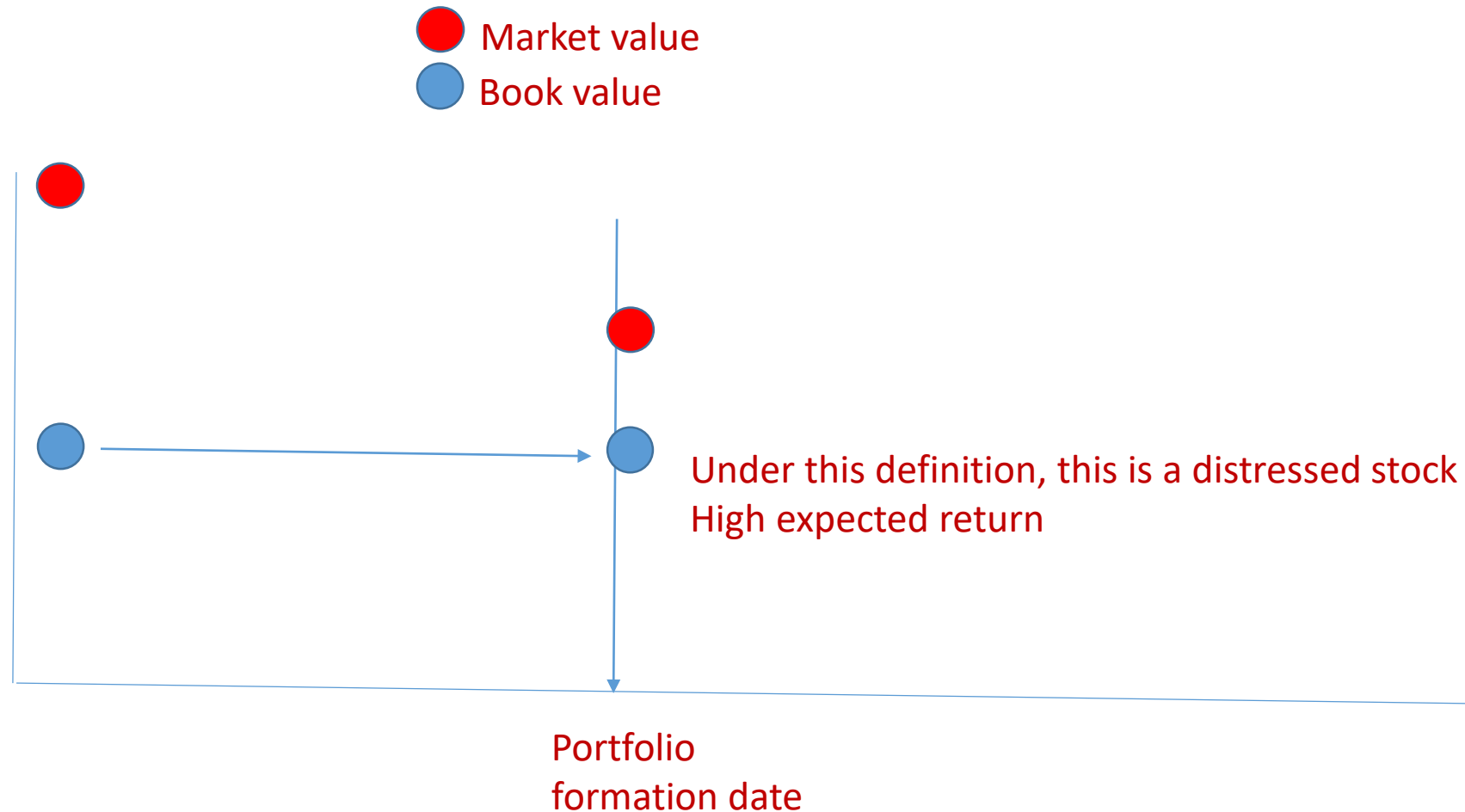
*Use market cap
from here before
Portfolio, hypothetical
Quarterly rebalance*



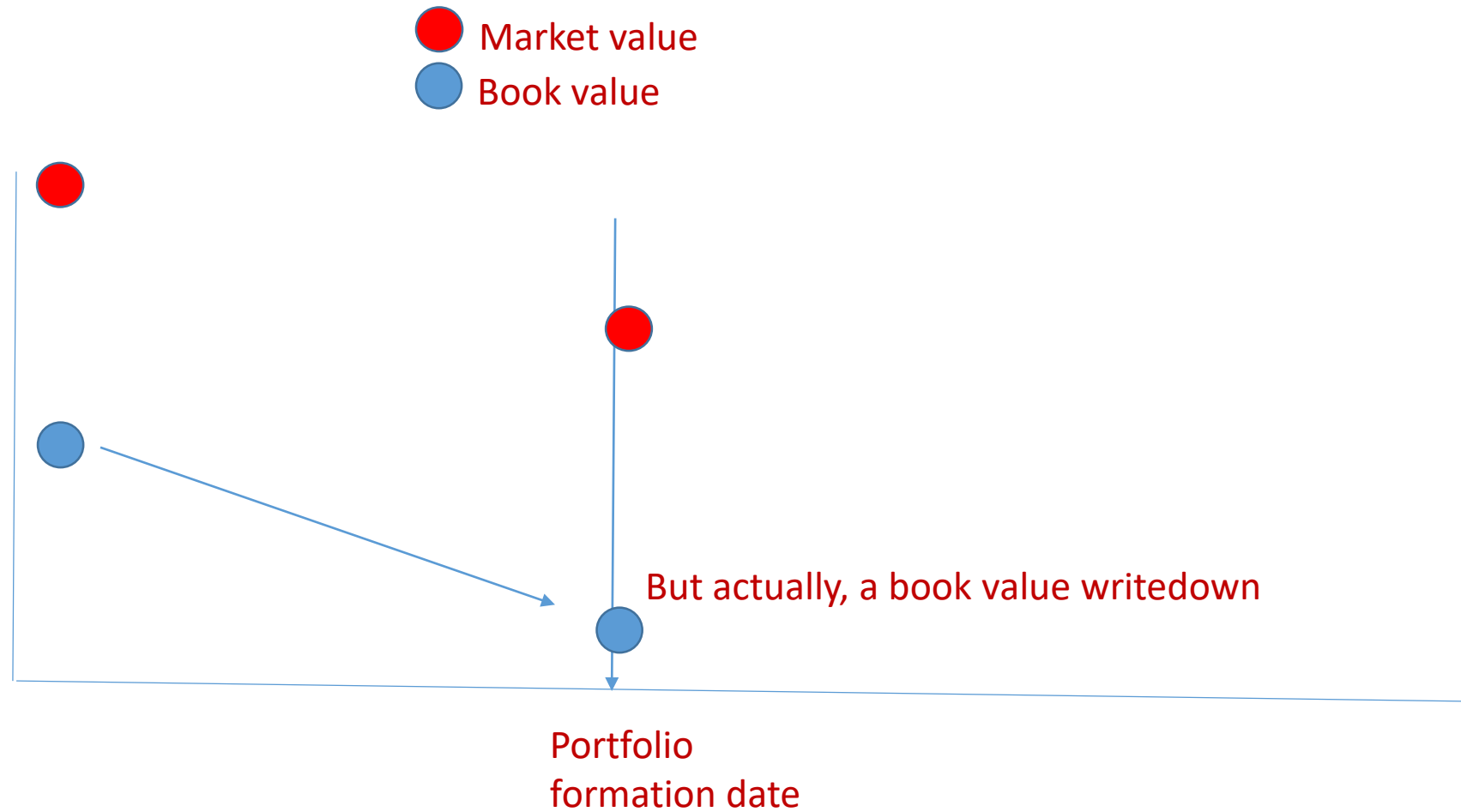
The value trap



If nothing changed about book value, then..



A book value writedown



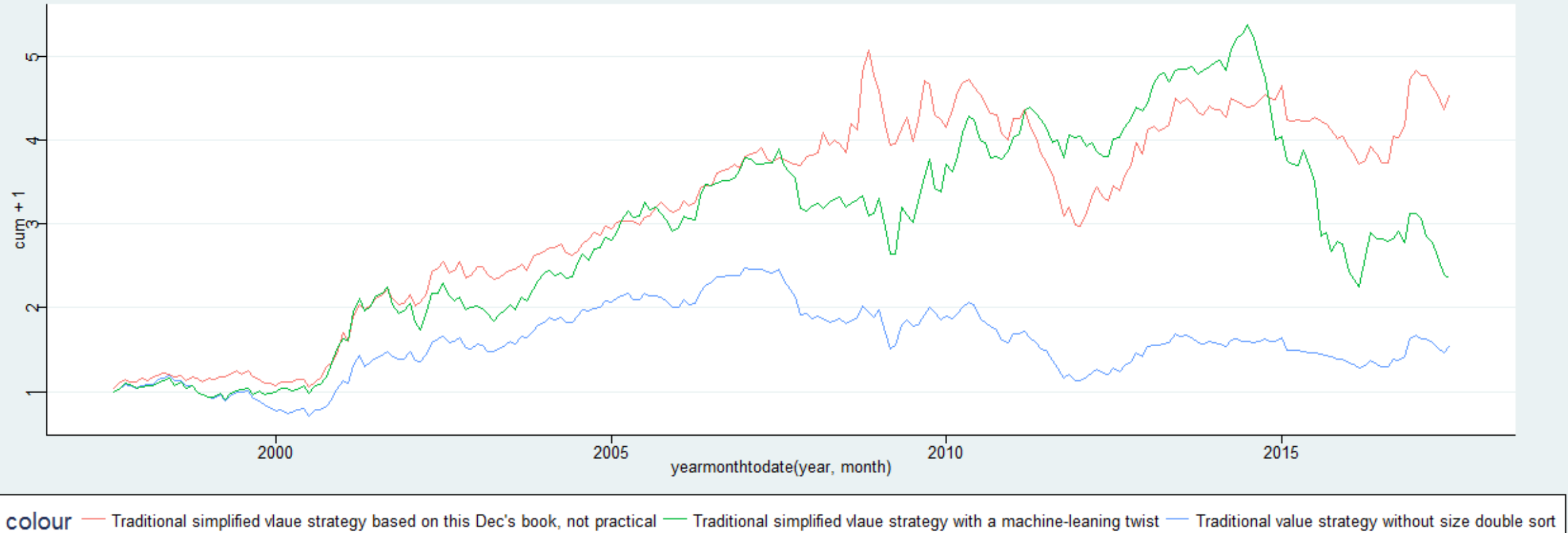
Nowcasting the value trap

- If I want to invest but don't know the December 1996 B/M data, using pre-1996 data, let's run the following regression:
- $\log(B_{t+1}) - \log(B_t) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$
- But instead of using OLS, use a 6-fold cross validation LASSO regression
 - This will pick the optimal penalty parameter
 - The optimal variables to include in the regression
- What $x_1 \dots x_n$ do I choose?
 - Use anything you think might be helpful – news data, industry factors, returns, etc. analyst forecasts. Ideally the best possible factors
 - But we let the model talk

Implementing it into trading

- Once you train the model, you might find some variables matter, some variables don't
- $\log(B_{t+1}) - \log(B_t) = \alpha + \beta_1 x_{1,t} + \beta_2 x_{2,t} \dots + \beta_n x_{n,t} + \epsilon$ this is your model using data from before 1996
- $\log(B_{1996}) - \log(B_{1995}) = \alpha + \beta_1 x_{1,1995} + \beta_2 x_{2,1995} \dots + \beta_n x_{n,1995} + \epsilon$
- Now, what you do is:
 - $\frac{B_{1996}^{predicted}}{MKT_{june,1996}}$

A student's answer from last yaer



Sparse Signals in the Cross-Section of Returns, Chinco, Clark-Jseoph and Mao Ye (2017)

- Economists spend a lot of time thinking about factors that predict stock returns over time, that are stable. But...
 - The economy is a complex system that adapts all of the time
- Question: what factors predict future returns?
 - If we let the data talk, what would emerge?
 - Can we let the data talk over time
- Applies LASSO at the 1 minute level to the entire cross-section of lagged predictors
 - Does the out-of-sample future R-squared improve relative to traditional economic factors?
- Identify factors that appear to work, which appear to be tied to economically meaningful events

Isn't this data mining?

- The authors say yes but..
 - But, the Bible does not say that all sources of return predictability should make intuitive sense.
- The takeaway is *not* that researchers shouldn't use inputs that make sense, but that short-lived economic phenomena drive asset returns
 - If we let the data talk, what would emerge?
 - Can we let the data talk over time?

What does this paper do?

- Sample period + sample
 - Past 30 minutes, pick 250 stocks at random next period to forecast
- Estimated using the `cv.glmnet` package in R
 - K=10 cross-validation

Estimation and Forecast Timing

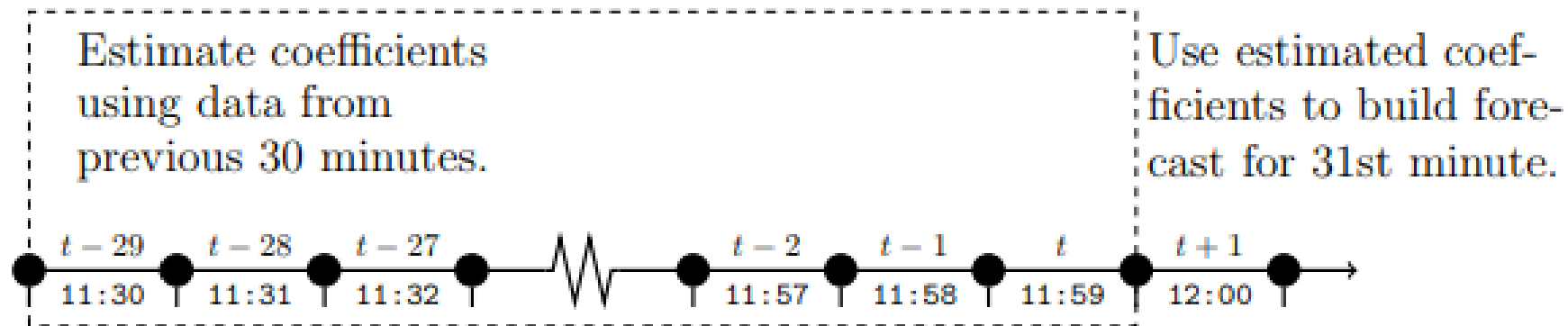


Figure 2: To make our 1-minute-ahead forecast for the n th stock's return in minute $(t + 1) = 12:00$, we first estimate a model using data from the previous 30 minutes, $\{11:30, \dots, 11:59\}$. Then, we apply the estimated coefficients to the most recent 3-minutes of data, $\{11:57, 11:58, 11:59\}$. We use $f_{11:59}^{\text{LASSO}}$ to denote this forecast for the n th stock's return in minute 12:00, because it only uses information up to minute 11:59.

Benchmark for the LASSO

- Does the LASSO work at all?

- We can just judge by investment performance, but we can also benchmark

- Benchmark against the AR(3) model

- $r_t = \alpha + \beta_1 r_{t-1} + \beta_2 r_{t-2} + \beta_3 r_{t-3} + \epsilon_T$

Trading strategy

- Roughly: Buy stocks that are above the average return for the day, short stocks that based on lasso are forecasted to be below the mean for the day
 - Weight by $1/\overline{s^{LASSO}}$ for each percentage point above the implied mean
 - s^{LASSO} is the standard deviation of the forecasts for the stock over the course of the day
 - In other words, constant volatility across the stocks
- Trading costs
 - Only trade when the forecasted return exceeds the bid-ask spread, and subtract the spread from the cost

Trading Frequency of Forecast-Implied Strategies

Aggregate [# /min]	LASSO	AR(3)
Trades	8.624	17.643
Buy Orders	4.306	8.887
Successful	2.600	6.538

By Characteristics [# /min]	>50%ile Mkt Cap	>50%ile Volume	>50%ile Volatility	<50%ile Spread
LASSO Trades	5.188	5.115	4.821	5.114
LASSO Buy Orders	2.592	2.555	2.407	2.556
LASSO Successful	1.578	1.533	1.467	1.550

Table V: Trading frequency of forecast-implied strategies. **(Sample)** Minute-level trades for randomly selected subset of 250 stocks on each trading day from January 2005 to December 2012 for which we compute 1-minute-ahead return forecasts using the LASSO. **(Aggregate)** First row reports the average number of trades per minute made by LASSO- and AR(3)-implied strategies. Second row reports the number of buy orders per minute. And, third row reports the number of trades per minute that made money after accounting for the bid-ask spread. **(By Characteristics)** Number of trades, buy orders, and successful trades per minute for the LASSO-implied strategy among large stocks, stocks with high trading volume, stocks with high return volatility, and liquid stocks. **(Reads)** “The AR(3)-implied strategy trades roughly twice as often as the LASSO-implied strategy. And, the LASSO-implied strategy is more active in large, liquid, frequently traded stocks.”

At these frequencies, even a small increase in R-squared significant

Out-of-Sample Fit, The LASSO

	Mean	95% CI
\bar{a}_n [%/m]	0.002 (0.002)	.
\bar{b}_n [%/m]	1.433 (0.017)	[1.399, 1.467]
\bar{R}_n^2 [%]	2.467 (0.027)	[2.414, 2.520]

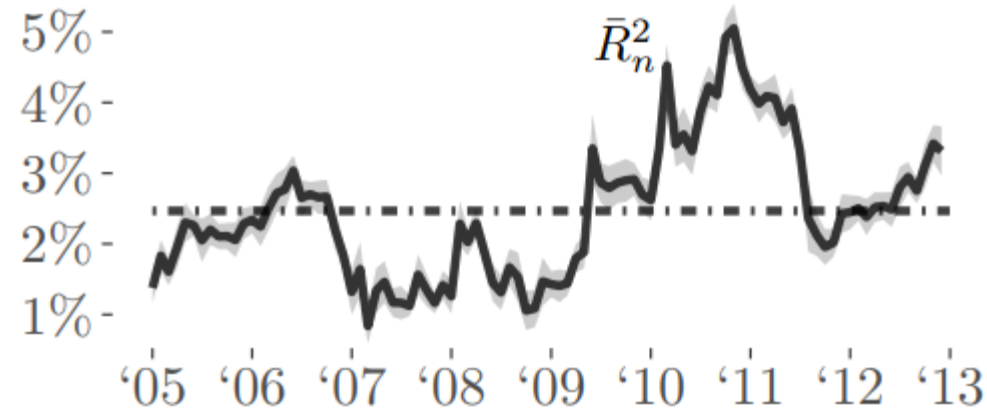


Table 1: *The LASSO’s out-of-sample fit. (Sample) Results of predictive regressions run at the stock-day level for 250 randomly chosen stocks on each trading day from January 2005 to December 2012. (Left) Population averages for regression coefficients and adjusted R^2 statistic. Numbers in parentheses are standard errors clustered by stock-day. 95% CI reports 95%-confidence intervals for population averages. (Right) Average adjusted R^2 statistic each month with the dashed line representing $\bar{R}_n^2 = 2.467\%$. Grey bands denote the 99.9% confidence interval computed using standard errors clustered by stock-day. (Reads) “On average, the LASSO’s 1-minute-ahead return forecast explains 2.467% of the variation in a stock’s returns on a given day. And, the LASSO explains at least 1% of the variation in returns during every month in our sample.”*

Increase in Out-of-Sample Fit, Main Results

	$\bar{R}_n^{2,\text{Bmk}}$ [%]	$\Delta \bar{R}_n^2$ [%pt]	p -val.
AR(3)	7.365 (0.076)	1.185 [1.162, 1.208] (0.012)	0.000
Market	0.311 (0.003)	2.469 [2.416, 2.522] (0.027)	0.000
AR(3), Market	5.553 (0.058)	1.424 [1.395, 1.453] (0.015)	0.000

Increase in Out-of-Sample Fit, Time Series

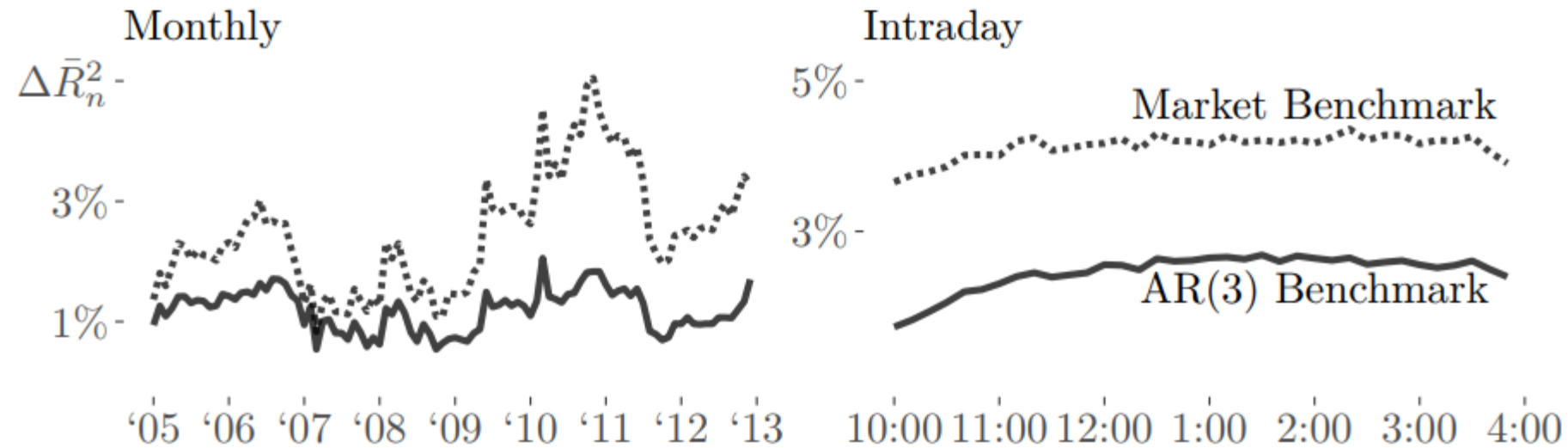
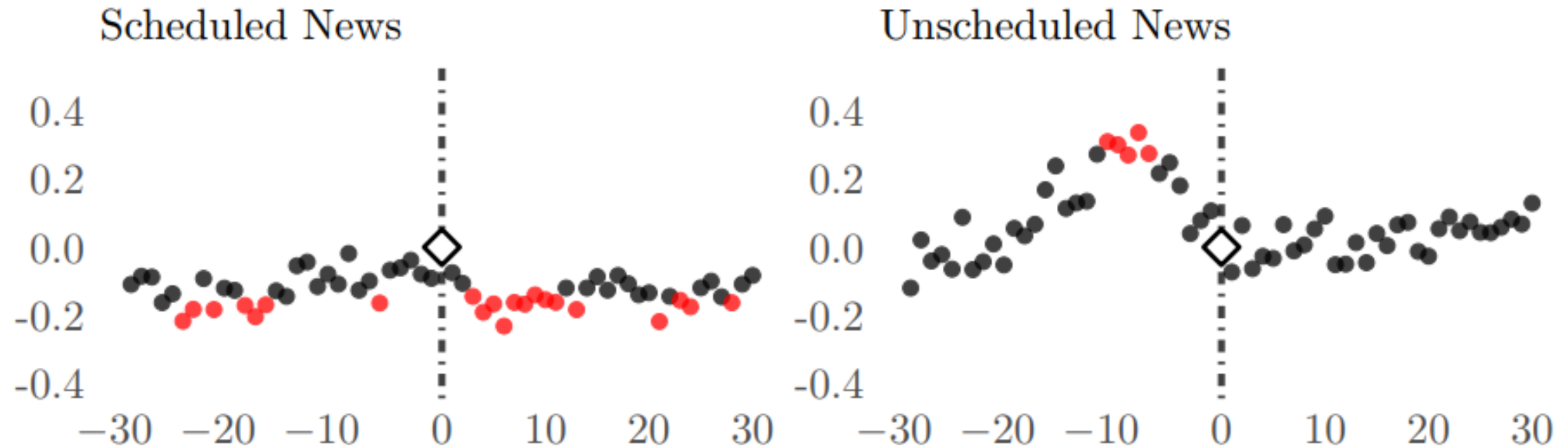


Figure 4: Time-series variation in the LASSO's increase in out-of-sample fit relative to the AR(3) and market benchmarks. Increase in out-of-sample fit is measured as the percentage point increase in adjusted R^2 . **(Sample)** Regression results for a randomly selected subset of 250 stocks on each trading day from January 2005 to December 2012. **(Monthly)** The LASSO's increase in out-of-sample fit by month. **(Intraday)** The LASSO's increase in out-of-sample fit by minute of the trading day. **(Reads)** "The LASSO increased in out-of-sample fit less during the financial crisis, but the LASSO's increase in out-of-sample fit is relatively stable over the course of the trading day."

The LASSO's Selection Rate Around News



clustered by year. (Scheduled News) News flashes about scheduled events. (Unscheduled News) News flashes about unscheduled events. (Sample) 61-minute window around each novel news flash about an NYSE-listed stock's revenues which occurred during normal trading hours from January 2005 to December 2012. (Reads) "When there is a scheduled news flash about the n 'th stock's revenues in minute t , the LASSO selects stock n ' as a predictor slightly more often when making its 1-minute-ahead return forecasts for minute t —that is, for $h = 0$. But, if the news flash is unscheduled, then the LASSO selects stock n ' as a predictor much more often when making its 1-minute-ahead return forecasts in the 10 minutes prior to minute t ."

Annualized Sharpe Ratios		
S&P 500	LASSO	AR(3)
0.123	1.791	−0.662

LASSO-Implied Strategy Abnormal Returns [%/yr]	α	Mkt	HmL	SmB	Mom
Market	2.709 (0.034)	0.004 (0.002)			
3-Factor Model	2.713 (0.034)	0.004 (0.002)	−0.004 (0.004)	0.000 (0.003)	
4-Factor Model	2.707 (0.034)	0.005 (0.002)	−0.004 (0.004)	0.003 (0.004)	0.003 (0.004)

Table 5: *Performance of forecast-implied strategies net of trading costs. (Sample) Each trading day from January 2005 to December 2012. (Sharpe Ratios) Annualized Sharpe ratios of forecast-implied trading strategies net of trading costs. First column reports results for strategy that invests \$1 in the S&P 500 at market open on January 3rd, 2005 and holds that position until market close on December 31st, 2012. The next column reports results for the LASSO’s forecast-implied trading strategy over the same time period. The third column reports analogous results for an AR(3)-implied strategy with the same initial investment. (Abnormal Returns) Net abnormal returns of the LASSO-implied strategy relative to the market, the Fama and French (1993) 3-factor model, and the Carhart (1997) 4-factor model. The size of the initial investment in the LASSO-implied strategy was chosen so that it has the same average excess return as the buy-and-hold S&P 500 strategy. First column reports annualized abnormal returns. Remaining columns report dimensionless slope coefficients associated with each factors. (Reads) “The LASSO-implied strategy generates positive excess returns net of the spread with an annualized Sharpe ratio of 1.791, and these excess returns are not explained by the strategy’s exposures to standard risk factors.”*

Group lasso

➤ Two step procedure

- More here: <http://hansheng.gsm.pku.edu.cn/pdf/2008/agLasso.pdf>
- <http://statweb.stanford.edu/~tibs/ftp/sparse-grlasso.pdf>
- Yuan & Lin (2007)

$$\min_{\beta \in \mathbb{R}^p} \left(\left\| \mathbf{y} - \sum_{\ell=1}^L \mathbf{X}_{\ell} \beta_{\ell} \right\|_2^2 + \lambda \sum_{\ell=1}^L \sqrt{p_{\ell}} \|\beta_{\ell}\|_2 \right), \quad (1)$$

➤ Each group has a different restriction

The adaptive lasso is a new trick – you *should* use it

- Adaptive lasso is a version of lasso with better statistical performance
 - It is not “biased”
- Instead of the same λ penalizing every coefficient equally, its weights vary for variables and “adapt”
 - It is equally efficient as the LASSO as it uses the same calculation method
- The adaptive lasso is implemented as follows:
 - It is equally efficient as the LASSO as it uses the same calculation method

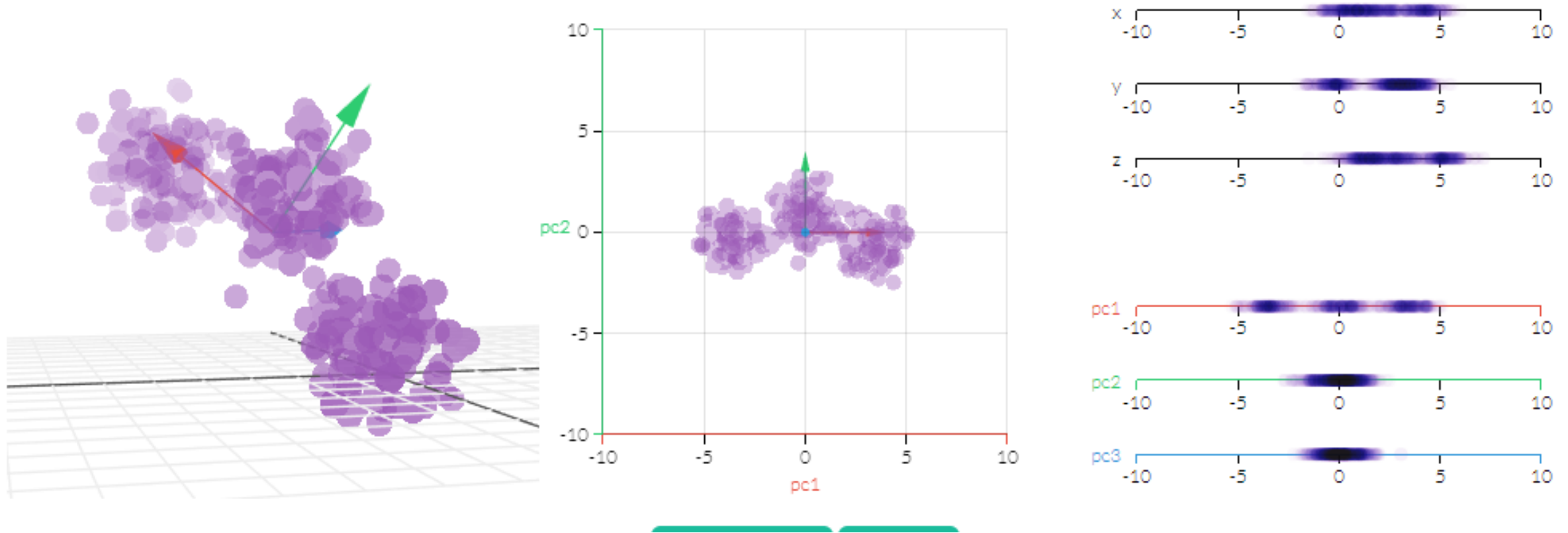
$$RSS(\beta) + \lambda \sum_{j=1}^p \hat{\omega}_j |\beta_j|$$

$$\hat{\omega}_j = \frac{1}{\left(|\hat{\beta}_j^{ini}|\right)^\gamma}$$

➤ Implementation details

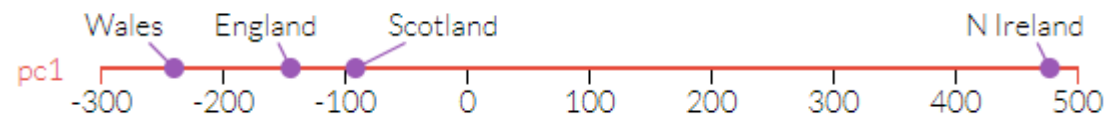
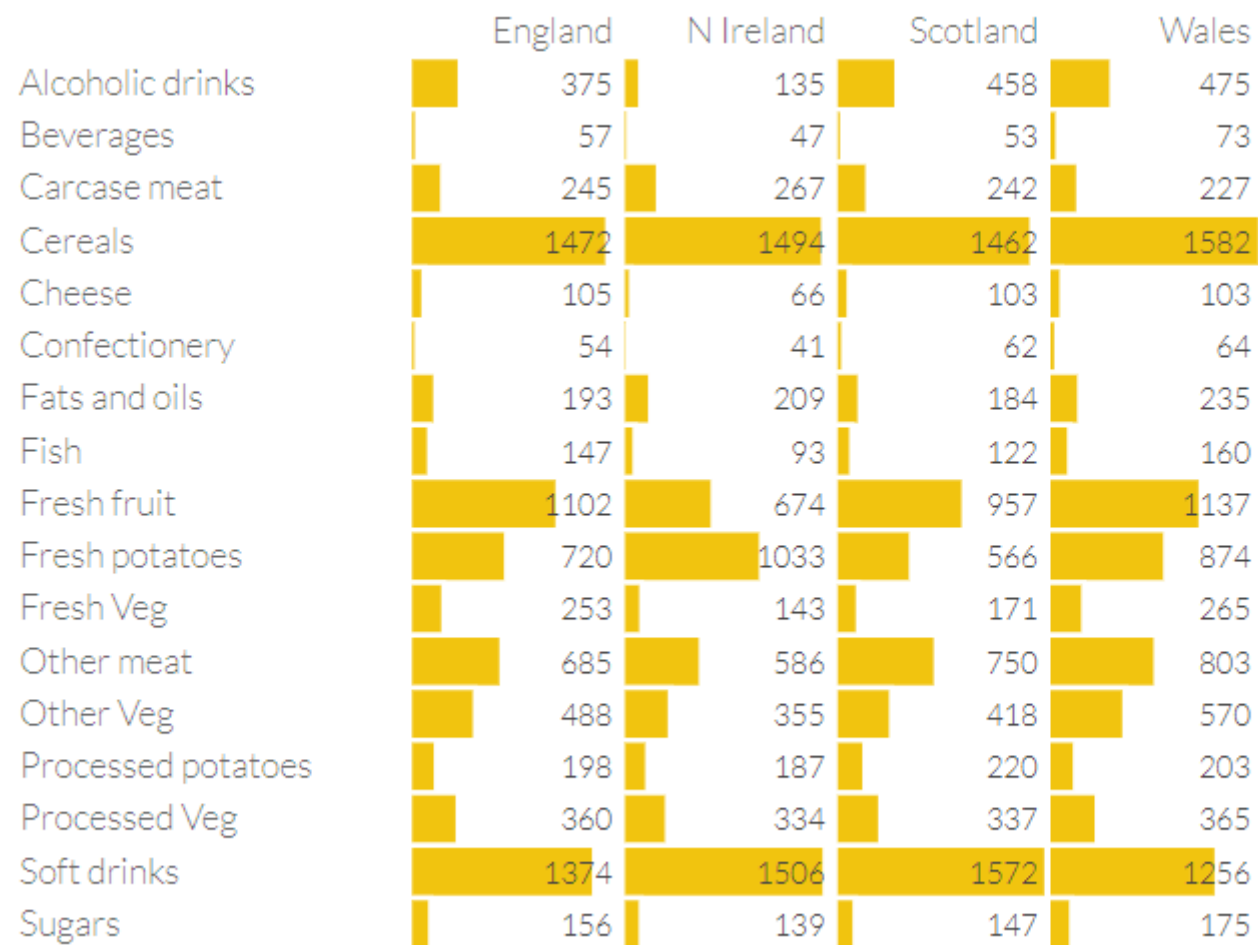
- <http://ricardoscr.github.io/how-to-adaptive-lasso.html>

PCA visualized: first map the data into a new vector



Graph: <http://setosa.io/ev/principal-component-analysis/>

Can do this for N variables



What does PCA do?

- Dimension reduces so you can combine data
- Tells you how many features are present in the data
 - How many risk factors do I have?
 - Helps me figure out what those risk factors are
- How is PCA calculated? (Not covered)

Algorithm 1 Principal Component Analysis

- 1: **procedure** PCA
 - 2: Compute dot product matrix: $\mathbf{X}^T \mathbf{X} = \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})^T (\mathbf{x}_i - \boldsymbol{\mu})$
 - 3: Eigenanalysis: $\mathbf{X}^T \mathbf{X} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T$
 - 4: Compute eigenvectors: $\mathbf{U} = \mathbf{X} \mathbf{V} \boldsymbol{\Lambda}^{-\frac{1}{2}}$
 - 5: Keep specific number of first components: $\mathbf{U}_d = [\mathbf{u}_1, \dots, \mathbf{u}_d]$
 - 6: Compute d features: $\mathbf{Y} = \mathbf{U}_d^T \mathbf{X}$
-

Bond PCA

- A very good worked out example is available here
 - <https://insightr.wordpress.com/2017/04/14/american-bond-yields-and-principal-component-analysis/>
- Suppose we have many US government bonds
 - At any given point in time, we may have “on the run” and “off the run” bonds
 - On the run: the recent issue
 - Off the run: is no longer the most recent issue. For example, a previously 5 year bond that has 4 years left
- Taken together, this is a yield curve
- Question: what are the risk factors?
 - Let's arrive at the conclusion the entire industry uses in 5 minutes with PCA

What are properties of a risk factor?

- A risk factor is a variable that when it moves, other stocks will move as well
 - For example, we saw in the consumption CAPM, if aggregate consumption moves, stocks will move
- In equities, this is what we saw:
 - In the regular CAPM, if the aggregate market moves, so to will on average most stocks
 - Thus the major risk factor in
- In equities, this is what we saw:

PCA example makes it easy for me

<https://insightr.wordpress.com/2017/04/14/american-bond-yields-and-principal-component-analysis/>

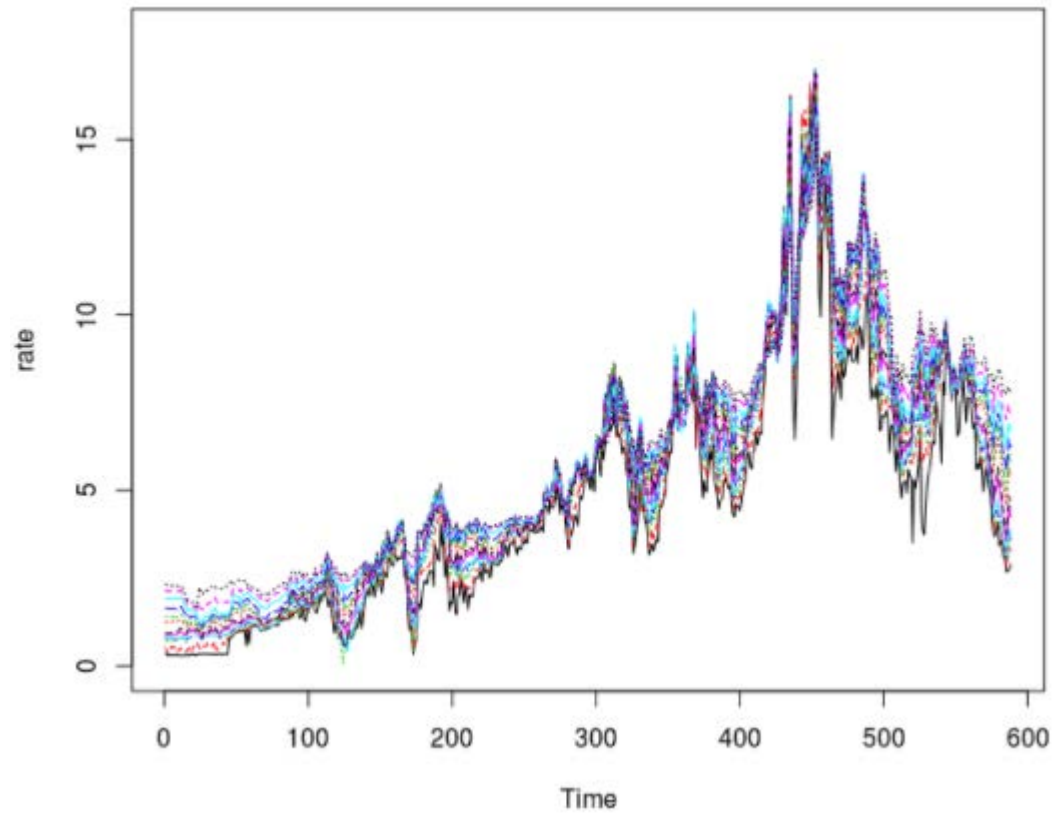
```
> url="https://www.dropbox.com/s/lg5q108k7cr4rmu/yields.csv?dl=1"
> t=tempfile()
> read_lines(url) %>% write_lines(t)
> read.csv(t, dec = ',', sep = ';', header = FALSE) -> yields
> colnames(yields) <- c(paste(c(1, 3, 6, 9, 12, 18), 'mth'), paste(c(2, 3, 4, 5, 7, 10, 15), 'yr'))
> data.table::setDT(yields)
> print(yields)
```

	1 mth	3 mth	6 mth	9 mth	12 mth	18 mth	2 yr	3 yr	4 yr	5 yr	7 yr	10 yr	15 yr
1:	0.34	0.51	0.70	0.79	0.84	0.90	0.93	1.24	1.40	1.65	1.93	2.16	2.33
2:	0.29	0.36	0.73	0.79	0.82	0.90	0.94	1.25	1.41	1.64	1.90	2.12	2.30
3:	0.32	0.42	0.78	0.80	0.81	0.88	0.92	1.25	1.42	1.65	1.91	2.11	2.27
4:	0.29	0.40	0.70	0.76	0.78	0.89	0.94	1.23	1.37	1.63	1.92	2.13	2.28
5:	0.30	0.55	0.74	0.78	0.79	0.89	0.94	1.24	1.39	1.64	1.92	2.12	2.28

584:	3.07	3.26	3.34	3.45	3.51	3.89	4.08	4.92	5.34	5.88	6.50	7.19	7.85
585:	2.67	2.81	2.97	3.08	3.14	3.55	3.76	4.58	4.99	5.58	6.26	7.02	7.75
586:	2.70	3.19	3.34	3.57	3.69	4.16	4.40	5.22	5.63	6.16	6.76	7.38	7.95
587:	2.75	3.35	3.58	3.83	3.95	4.54	4.83	5.64	6.04	6.45	6.91	7.43	7.94
588:	2.88	3.27	3.41	3.60	3.69	4.24	4.52	5.39	5.82	6.23	6.69	7.22	7.73

As you can tell, the bonds “co-move”

```
matplot(yields, type='l', ylim = c(0,18),ylab = 'rate', xlab = 'Time')
```



Calculate bond returns, then plot correlation

```
returns=colwise(diff)(yields)
> round(cor(returns)*100,2)
```

	1 mth	3 mth	6 mth	9 mth	12 mth	18 mth	2 yr	3 yr	4 yr	5 yr	7 yr	10 yr	15 yr
1 mth	100.00	78.74	72.92	69.30	65.62	63.12	60.37	54.00	48.98	47.58	43.92	39.31	30.86
3 mth	78.74	100.00	93.31	88.57	83.89	80.72	77.34	71.01	65.61	63.40	58.09	53.48	44.56
6 mth	72.92	93.31	100.00	96.76	92.61	90.50	87.52	82.24	77.17	74.94	69.22	64.90	55.78
9 mth	69.30	88.57	96.76	100.00	99.13	96.86	93.65	89.33	84.66	82.49	76.61	72.02	61.99
12 mth	65.62	83.89	92.61	99.13	100.00	97.67	94.42	90.72	86.39	84.29	78.48	73.87	63.65
18 mth	63.12	80.72	90.50	96.86	97.67	100.00	99.28	96.25	92.13	90.43	84.91	80.58	69.97
2 yr	60.37	77.34	87.52	93.65	94.42	99.28	100.00	97.38	93.47	92.03	86.79	82.74	72.16
3 yr	54.00	71.01	82.24	89.33	90.72	96.25	97.38	100.00	99.09	97.90	92.85	88.38	76.67
4 yr	48.98	65.61	77.17	84.66	86.39	92.13	93.47	99.09	100.00	98.97	94.15	89.53	77.43
5 yr	47.58	63.40	74.94	82.49	84.29	90.43	92.03	97.90	98.97	100.00	97.99	94.10	81.21
7 yr	43.92	58.09	69.22	76.61	78.48	84.91	86.79	92.85	94.15	97.99	100.00	97.21	83.61
10 yr	39.31	53.48	64.90	72.02	73.87	80.58	82.74	88.38	89.53	94.10	97.21	100.00	93.88
15 yr	30.86	44.56	55.78	61.99	63.65	69.97	72.16	76.67	77.43	81.21	83.61	93.88	100.00

```
>
```

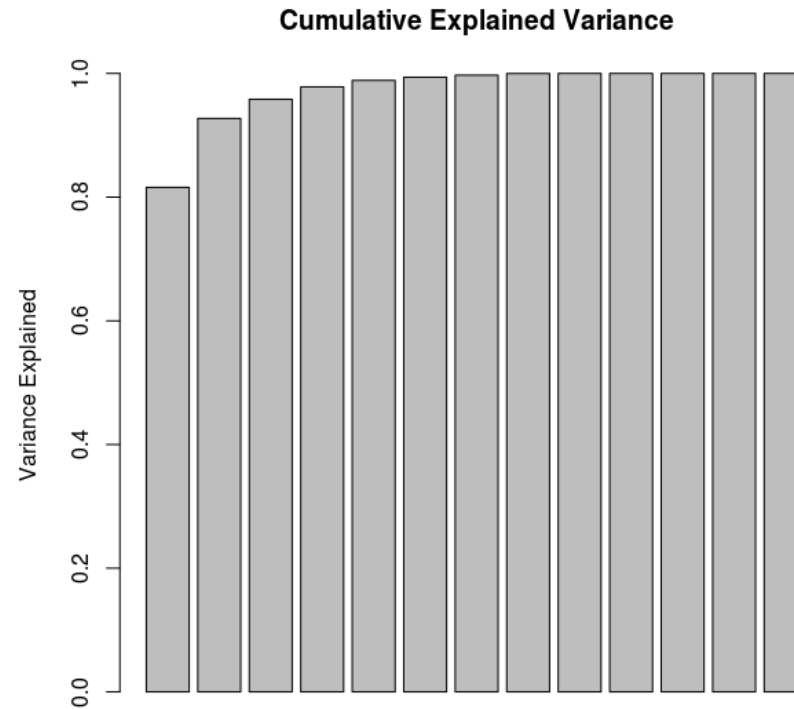
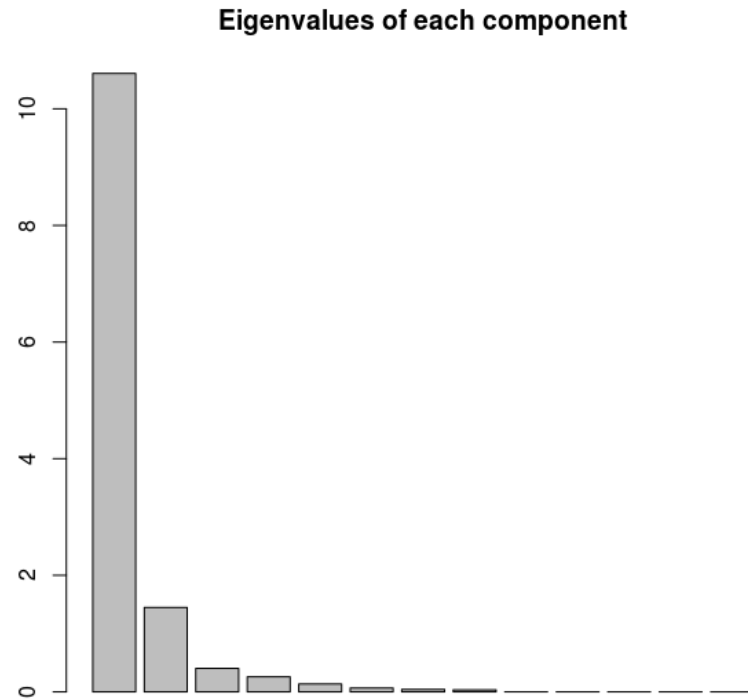
Very highly correlated, as you can see

Run the PCA command

```
> model = prcomp(returns, scale = TRUE, center = TRUE)
> summary(model)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	3.2568	1.2035	0.63521	0.50910	0.36849	0.25939	0.20759	0.18875	0.01382	0.01014	0.009669	0.008623	0.007291
Proportion of Variance	0.8159	0.1114	0.03104	0.01994	0.01044	0.00518	0.00332	0.00274	0.00001	0.00001	0.000010	0.000010	0.000000
Cumulative Proportion	0.8159	0.9273	0.95835	0.97828	0.98873	0.99390	0.99722	0.99996	0.99998	0.99998	0.999990	1.000000	1.000000



If you look closely at the loadings, you'll see they have a shape

```
> model$rotation[,1:3] %>% round(.,3)
```

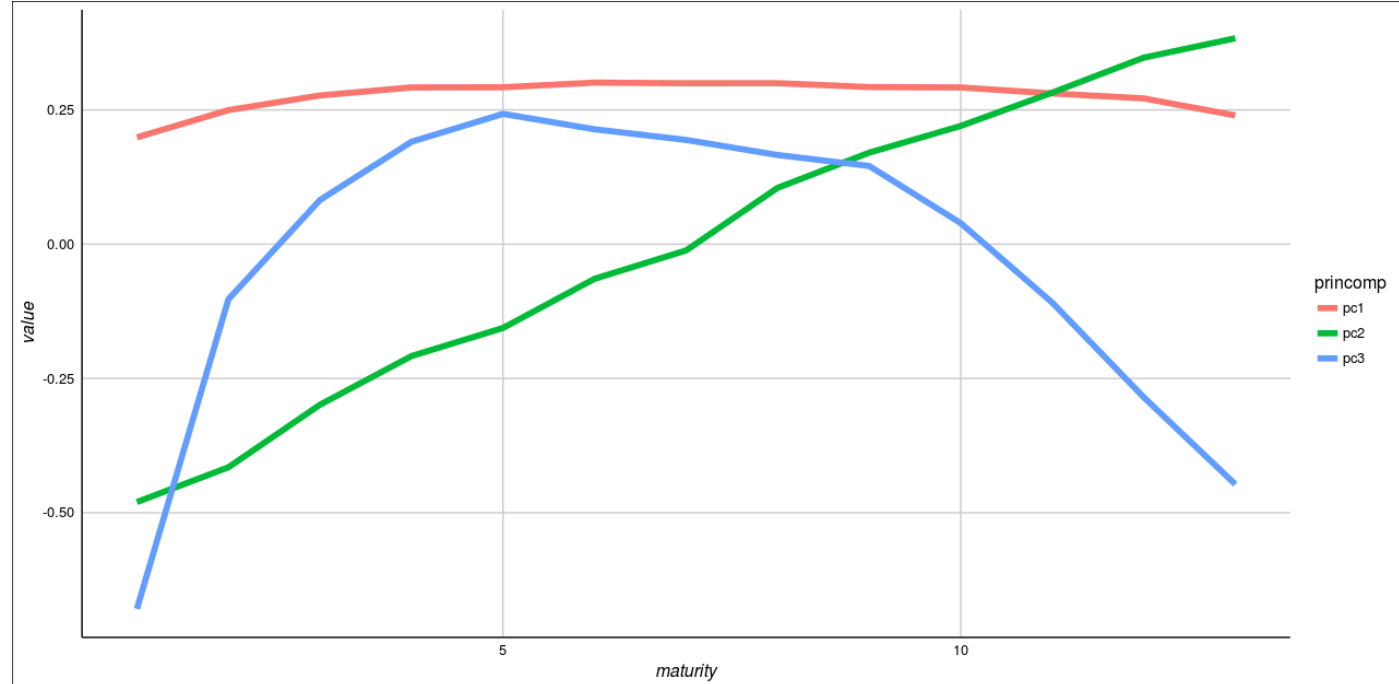
		PC1	PC2	PC3
1	mth	0.199	-0.480	-0.679
3	mth	0.250	-0.415	-0.102
6	mth	0.277	-0.299	0.082
9	mth	0.292	-0.208	0.191
12	mth	0.292	-0.156	0.243
18	mth	0.301	-0.064	0.214
2	yr	0.300	-0.012	0.194
3	yr	0.300	0.105	0.166
4	yr	0.293	0.170	0.146
5	yr	0.292	0.220	0.039
7	yr	0.281	0.282	-0.109
10	yr	0.271	0.347	-0.285
15	yr	0.240	0.383	-0.447

Let's plot against maturity

```
> data.table(names=model$rotation[, 'PC1'] %>% names,  
+            pc1=model$rotation[, 'PC1'],  
+            pc2=model$rotation[, 'PC2'],  
+            pc3=model$rotation[, 'PC3']) %>%  
+ mutate(maturity=1:length(pc1)) %>%  
+ melt.data.table(id.vars=c('names', 'maturity'), variable.name = 'princomp') -> temp  
>  
> print(temp)
```

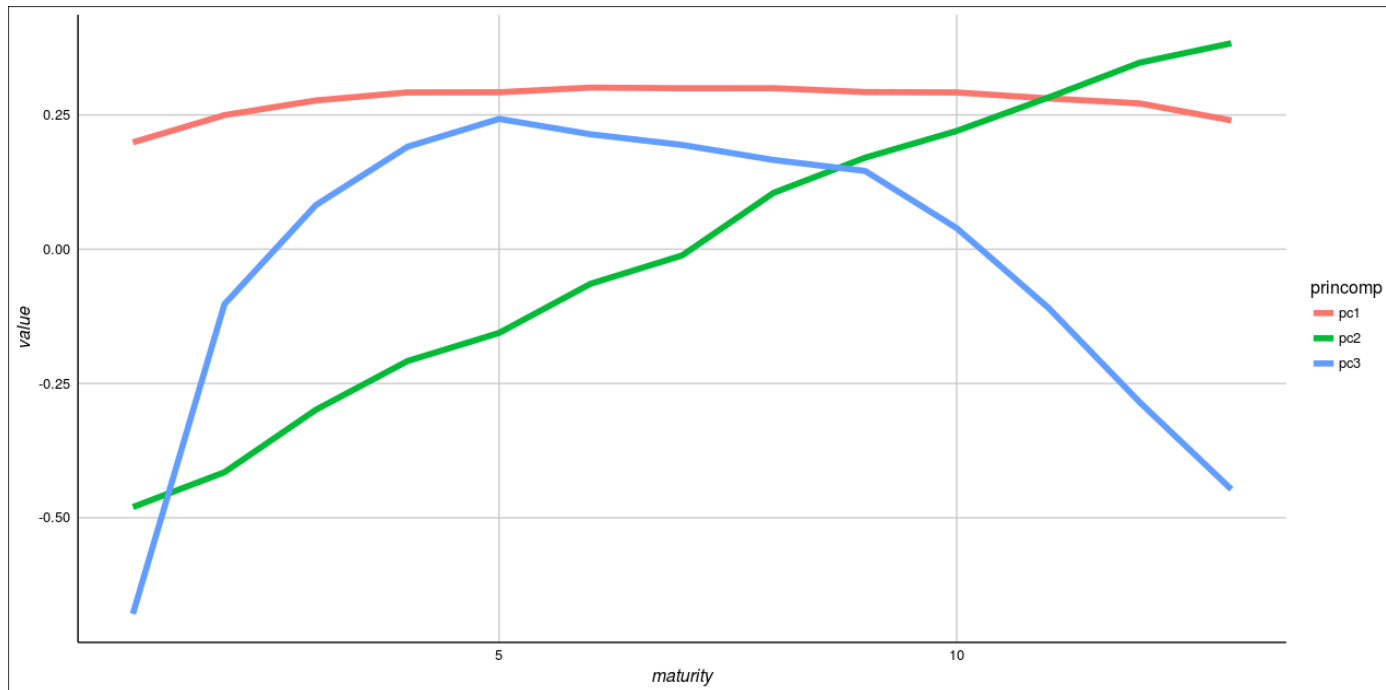
	names	maturity	princomp	value
1:	1 mth	1	pc1	0.19895706
2:	3 mth	2	pc1	0.24963095
3:	6 mth	3	pc1	0.27696182
4:	9 mth	4	pc1	0.29187502
5:	12 mth	5	pc1	0.29215588
6:	18 mth	6	pc1	0.30084398
7:	2 yr	7	pc1	0.29964646
8:	3 yr	8	pc1	0.29974334
9:	4 yr	9	pc1	0.29255957
10:	5 yr	10	pc1	0.29182716
11:	7 yr	11	pc1	0.28070195
12:	10 yr	12	pc1	0.27140155
13:	15 yr	13	pc1	0.23978907
14:	1 mth	1	pc2	-0.47998835
15:	3 mth	2	pc2	-0.41525866
16:	6 mth	3	pc2	-0.29883869
17:	9 mth	4	pc2	-0.20821349
18:	12 mth	5	pc2	-0.15587196
19:	18 mth	6	pc2	-0.06435274
20:	2 yr	7	pc2	-0.01157972
21:	3 yr	8	pc2	0.10503852
22:	4 yr	9	pc2	0.17045021
23:	5 yr	10	pc2	0.22019749
24:	7 yr	11	pc2	0.28226702
25:	10 yr	12	pc2	0.34737708
26:	15 yr	13	pc2	0.38343678
27:	1 mth	1	pc3	-0.67885949
28:	3 mth	2	pc3	-0.10240241
29:	6 mth	3	pc3	0.08231280
30:	9 mth	4	pc3	0.19078989
31:	12 mth	5	pc3	0.24262023
32:	18 mth	6	pc3	0.21389842
33:	2 yr	7	pc3	0.19413741
34:	3 yr	8	pc3	0.16602916
35:	4 yr	9	pc3	0.14561203
36:	5 yr	10	pc3	0.03921734
37:	7 yr	11	pc3	-0.10879610
38:	10 yr	12	pc3	-0.28498932
39:	15 yr	13	pc3	-0.44701103

```
names maturity princomp value  
> temp %>% ggplot(aes(x=maturity,y=value,group=princomp,color=princomp)) + geom_line(size=2) +  
+ ggthemes::theme_gdocs()
```

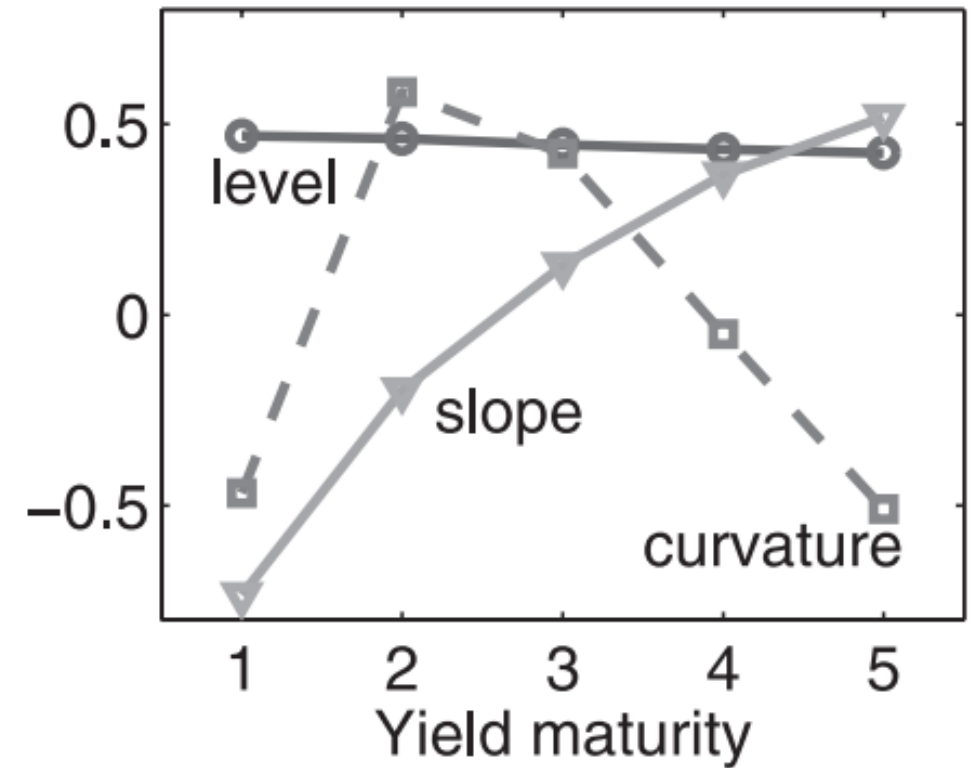


If you look closely at the loadings, you'll see they have a shape

Cochrane and Piazzesi (2005)
American Economic Review



B. Yield factors



Clark (2016) – The Level, Slope and Curve Factor Model for Stocks

- What explains the expected return on stocks? How many risk factors “should there be”?
- Every month, run this regression on the realized return:

$$XRet_{i,t+1} = \beta_0 + \beta_1 LogSize_{i,t} + \beta_2 LogB/M_{i,t} + \beta_3 Mom_{i,t} + \beta_4 zeroNS_{i,t} + \beta_5 NS_{i,t} + \beta_6 negACC_{i,t} + \beta_7 posACC_{i,t} + \beta_8 dA/A_{i,t} + \beta_9 posOP_{i,t} + \beta_{10} negOP + e_{i,t+1} \quad (3)$$

- Predict returns next month and sort stocks based on their imputed return
 - 25 return sorted portfolios
 - Similar to 25 maturity sorted portfolios

Create 25 expected return portfolios based on Fama French (2008)

$$XRet_{i,t+1} = \beta_0 + \beta_1 LogSize_{i,t} + \beta_2 LogB/M_{i,t} + \beta_3 Mom_{i,t} + \beta_4 zeroNS_{i,t} + \beta_5 NS_{i,t} + \beta_6 negACC_{i,t} + \beta_7 posACC_{i,t} + \beta_8 dA/A_{i,t} + \beta_9 posOP_{i,t} + \beta_{10} negOP + e_{i,t+1} \quad (3)$$

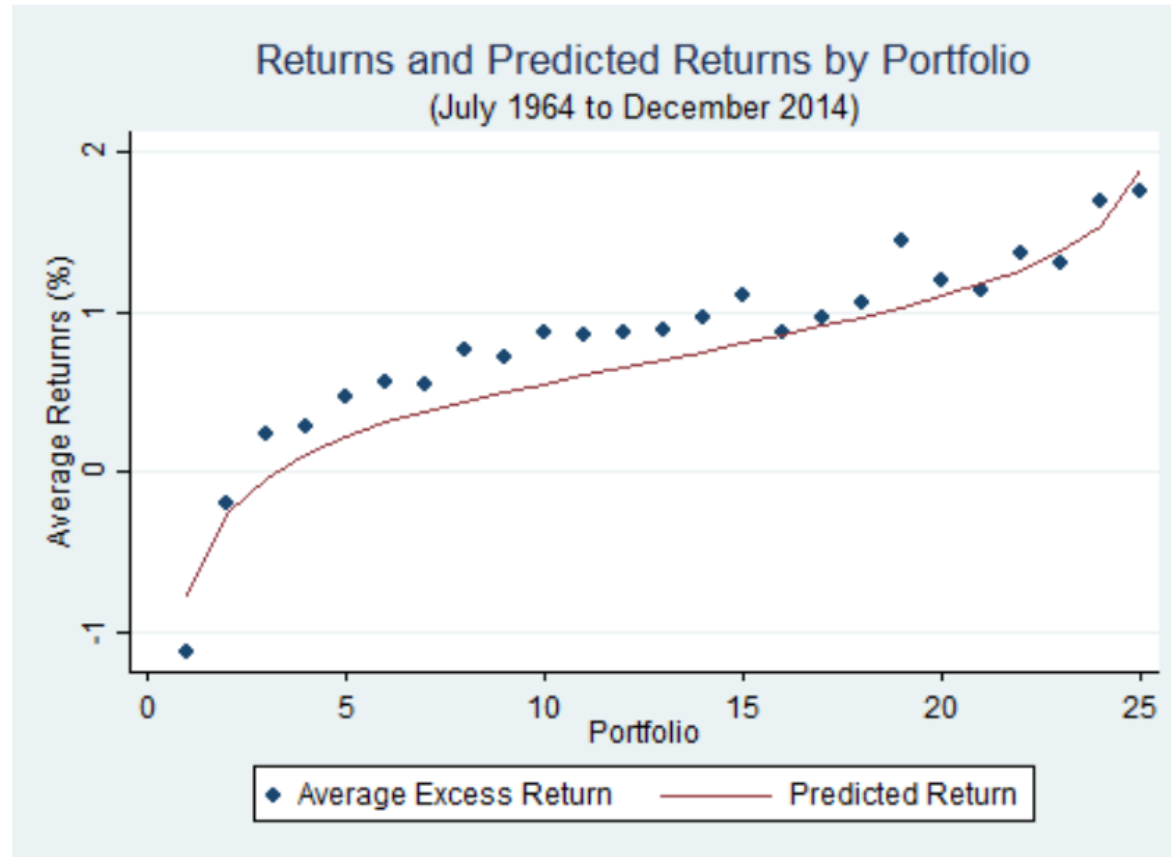


Figure 1: Twenty-Five Dissecting Anomaly Portfolios

Run the PCA and amazingly, you get level, slope and tent

Component	Eigenvalue	Variance Explained	Cumulative
Component 1	667.90	74.12%	74.12 %
Component 2	81.66	9.06%	83.18 %
Component 3	27.54	3.06%	86.24 %
Component 4	13.86	1.54%	87.78 %
Component 5	10.46	1.30%	89.08 %
Component 6	8.42	1.16%	90.24 %
Component 7	7.77	0.93%	91.18 %
Component 8	6.46	0.90%	92.08 %
Component 9	5.88	0.86%	92.94 %
Component 10	5.50	0.72%	93.66 %

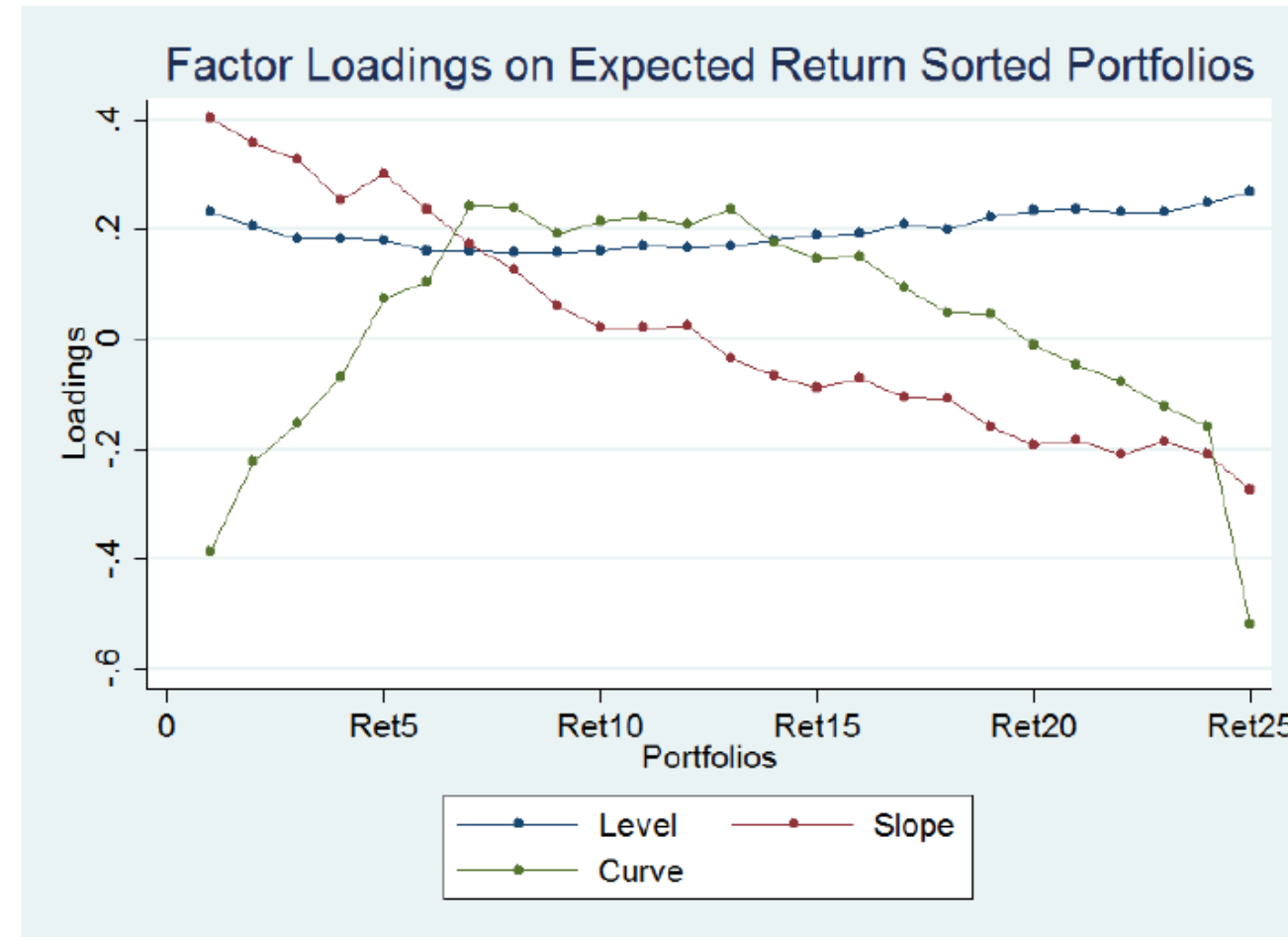


Figure 2: PCA Weights