Applications of Textual Analysis and Machine Learning in Asset Pricing

Benjamin Moritz

LMU Munich - Chair of Financial Econometrics

October 24, 2018

Disputation

Introduction

The central goal of financial asset pricing theory is to determine the price of an asset given its risks:

The risk-return relationship

Aggregate market over time[1]

Aggregate market over time
Merton (1973)

$$E_t[r_{t+1}] = \gamma \boldsymbol{E_t[Var_{t+1}]}$$

$E_t[r_{t+1}]$ = Expected excess return of the market portfolio
γ = Coefficient of relative risk aversion
$E_t[Var_{t+1}]$ = Expected variance of the market portfolio

[1] Joint work with Stefan Mittnik

Single securities in the cross-section[2]

Single securities in the cross-section
Fama and MacBeth (1973), Fama and French (1992)

$$E_t[r_{i,t+1}] = \alpha_i + \beta_i \boldsymbol{E_t[r_{t+1}]} + s_i E_t[SMB_{t+1}] + h_i E_t[HML_{t+1}] + ...$$

In general form:

$$E_t[r_{i,t+1}] = \boldsymbol{f_t}(E_t[r_{t+1}], E_t[SMB_{t+1}], E_t[HML_{t+1}], ...)$$

$E_t[r_{i,t+1}]$ = Expected excess return of company i
$E_t[SMB_{i,t+1}]$ = Expected return of small stocks minus big stocks
$E_t[HML_{i,t+1}]$ = Expected return of stocks with high B/M minus low B/M

[2]Joint work with Tom Zimmermann

AGGREGATE MARKET OVER TIME

Introduction

- In the very **long-run** the risk-return relation of the major asset classes is **positive and linear**.
- In the **short-run** the relation between stock market risk and return is **ambiguous**.

Risk-return relation in the short-run

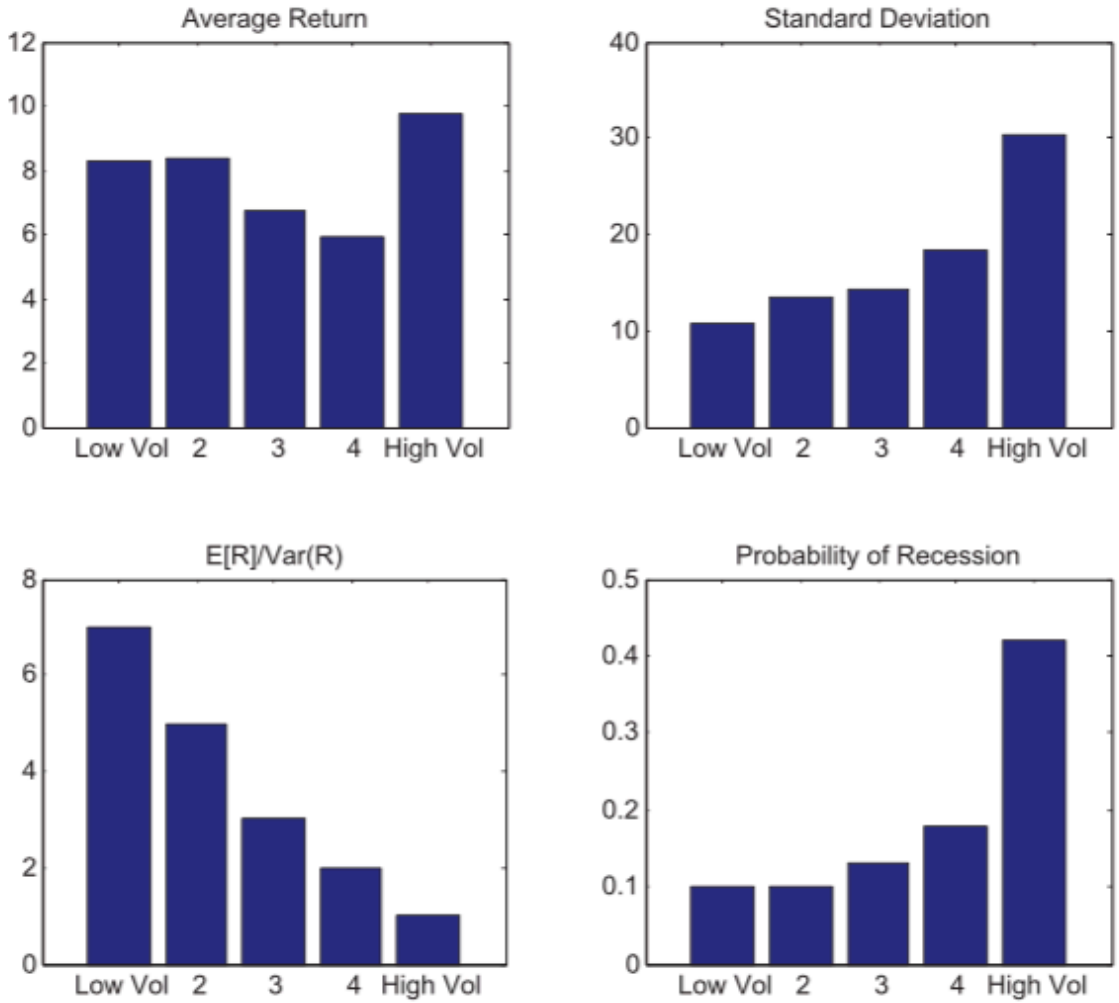


Figure: Source is Moreira and Muir (2017).

Risk-return relation in the short-run is ambiguous

$$E_t[r_{t+1}] = \gamma E_t[Var_{t+1}]$$

Examples

- **Negative relation:** Campbell (1987), Glosten et al. (1993) and Whitelaw (1994)
- **Nonlinear relation:** Ghysels et al. (2014)
- **Positive relation:** Bollerslev et al. (1988), Ghysels et al. (2005) and Lundblad (2007)

Two very important points in the theoretical continuous time model of Merton (1973) are,

1. the choice of horizon is open to the researcher
2. the market conditional variance is unobserved

Horizon

Only a small part of daily price changes can be explained by economic reasons or news.
Niederhoffer (1971), Roll (1988), Cutler et al. (1989), Fair (2002), Cornell (2013) and Campbell et al. (2018).

The level of prices is procyclical and the level of volatility is countercyclical.
Schwert (1989a), Schwert (1989b) and Panetta et al. (2006) for example.

Market conditional variance

$E_t[r_{t+1}] = \gamma E_t[Var_{t+1}]$

Decomposition: $E_t[Var_{t+1}] = E_t[Var_{t+1} - Var_t] + Var_t$

Most of the time, as in Moreira and Muir (2017), Var_t is used as a proxy for $E_t[Var_{t+1}]$.

We show that $E[Var_{t+1} - Var_t]$ has a negative relation with $E[r_{t+1}]$.

Regression a la French et al. (1987)

$$r_{mt} - r_{ft} = \alpha + \beta \hat{\sigma}^{p}_{mt} + \epsilon_t \tag{1}$$

$$r_{mt} - r_{ft} = \alpha + \beta \hat{\sigma}^{p}_{mt} + \gamma \sigma^{pu}_{mt} + \epsilon_t \tag{2}$$

where:
r_{mt} = return of the stock market
r_{ft} = return of the risk free rate
$\hat{\sigma}^{p}_{mt}$ = expected risk of the stock market
σ^{pu}_{mt} = unexpected risk of the stock market = $\sigma^{p}_{mt} - \hat{\sigma}^{p}_{mt}$
with $p = 1$ for standard deviation and $p = 2$ for volatility

Regression a la French et al. (1987)

Volatility measure	Equation 1		Equation 2		
	α	β	α	β	γ
	Sample A				
σ	0.0046	0.0015	0.0181	-0.0142	-0.0687
	(1.346)	(0.241)	(5.710)	(-2.594)	(-13.971)
	[1.265]	[0.257]	[5.152]	[-2.533]	[-9.470]
σ^2	0.0062	-0.0028	0.0095	-0.0003	-0.0230
	(3.850)	(-0.386)	(6.388)	(-0.052)	(-12.527)
	[1.779]	[-0.312]	[2.823]	[-0.040]	[-6.646]
	Sample B				
σ	0.0065	0.0001	0.0203	-0.0156	-0.0640
	(2.491)	(0.016)	(8.412)	(-3.943)	(-17.874)
	[2.262]	[0.017]	[7.289]	[-3.789]	[-12.573]
σ^2	0.0073	-0.0023	0.0107	-0.0011	-0.0255
	(5.893)	(-0.445)	(9.229)	(-0.241)	(-14.761)
	[2.595]	[-0.345]	[3.818]	[-0.178]	[-5.504]

This table shows the time-series regressions (WLS) for the two samples:
Sample A: February 1928 to December 1984 (T = 683)
Sample B: July 1926 to April 2018 (T = 1096)
t-Statistics are in parentheses below the coefficient estimates. The number in brackets are t-Statistics with standard errors based on the consistent heteroskdeasticity correction of White (1980).

Textual Analysis - Economist (1992)

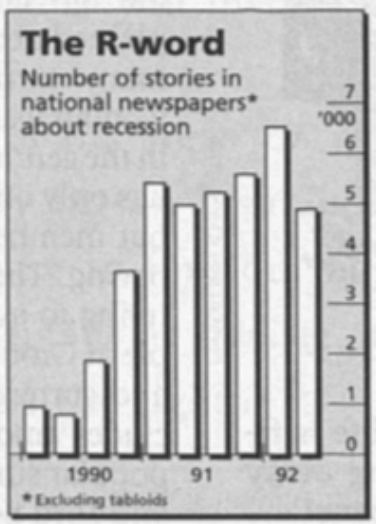
Counting the word "recession" on a quarterly frequency.

No news is good news

IS THE end of the recession in sight? Increasing mistrust of official statistics has encouraged economy-watchers to look for new indicators. *The Economist*'s tip is to watch the number of stories about recession in newspapers. FT-Profile, a computer data base, makes this easy. We asked it to count how many stories in Britain's quality newspapers had included the word recession in each quarter over the past three years.

The number of articles mentioning "recession" climbed from 776 in the second quarter of 1990, just before the economy started to shrink, to a peak of 6,524 in the first three months of this year. The good news is that the total has now started to fall. In May and June fewer recession stories were printed than a year ago, for the first time since the actual downturn began.

Alas, our indicator has a flaw. The computer counts all stories mentioning recession, including ones about countries other than Britain. Still, it may be no worse a guide to the real world than many fancier calculations, especially if you believe that press reports are partly responsible for spreading self-fulfilling gloom about the economy. So cheer up, better times may be ahead—provided papers continue to carry fewer articles (such as this one) on the R-word.



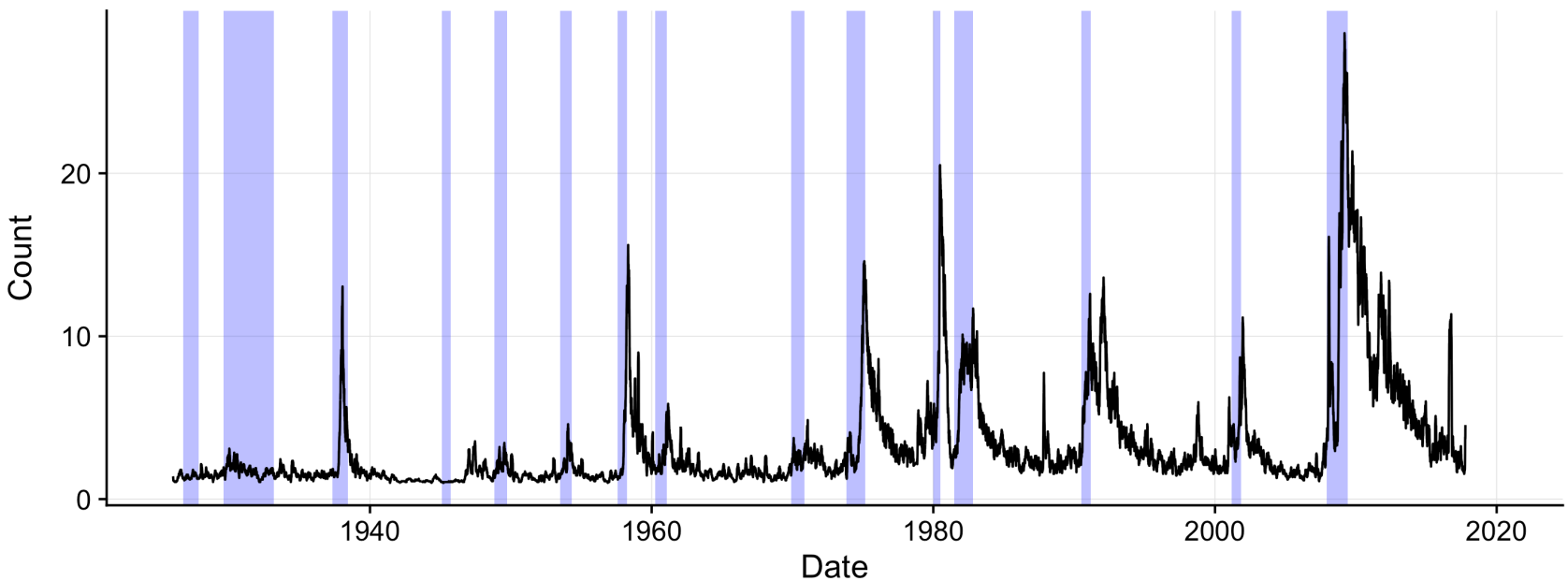
THE ECONOMIST JULY 11TH 1992

Textual Analysis - Methodology

We use methods from the field of **textual analysis** (or text mining, computational linguistics, natural language processing, information retrieval, ...). We proceed as follows:

1. Connection to New York Times using the NYT API
2. Search for the word ”recession” in all articles from 1851 until today (Total: 94427 Articles)
3. **Count the number of articles which include ”recession” on a daily frequency**

Counts



Textual Analysis - Methodology

We use methods from the field of **textual analysis** (or text mining, computational linguistics, natural language processing, information retrieval, ...). We proceed as follows:

1. Connection to New York Times using the NYT API
2. Search for the word "recession" in all articles from 1851 until today (Total: 94427 Articles)
3. **Count the number of articles which include "recession" on a daily frequency**
4. Scrap the full html-website and extract the text from 1981 until today (Total: 25334 Articles)
5. Parse the text into sentences
6. Calculate the sentiment using dictionaries
7. Calculate the total tone/sentiment of each article
8. **Average over all articles in one day to get the total tone/sentiment for each day**

The syuzhet dictionary

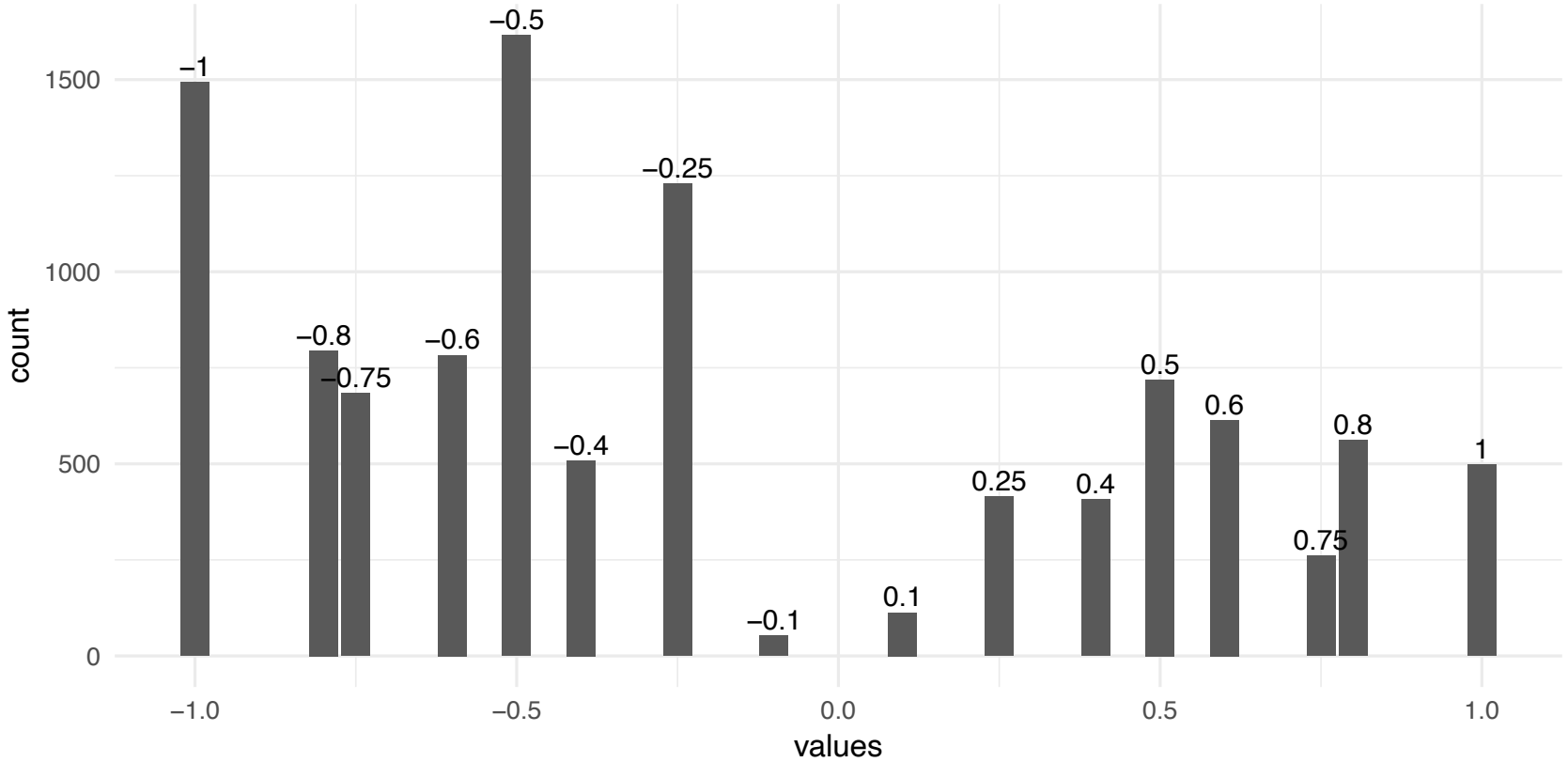


Figure: Histogram of the values of the words in the dictionary.

Total number of words: 10748
"Syuzhet" lexicon was developed in the Nebraska Literary Lab under the direction of Matthew L. Jockers.

Sentiment

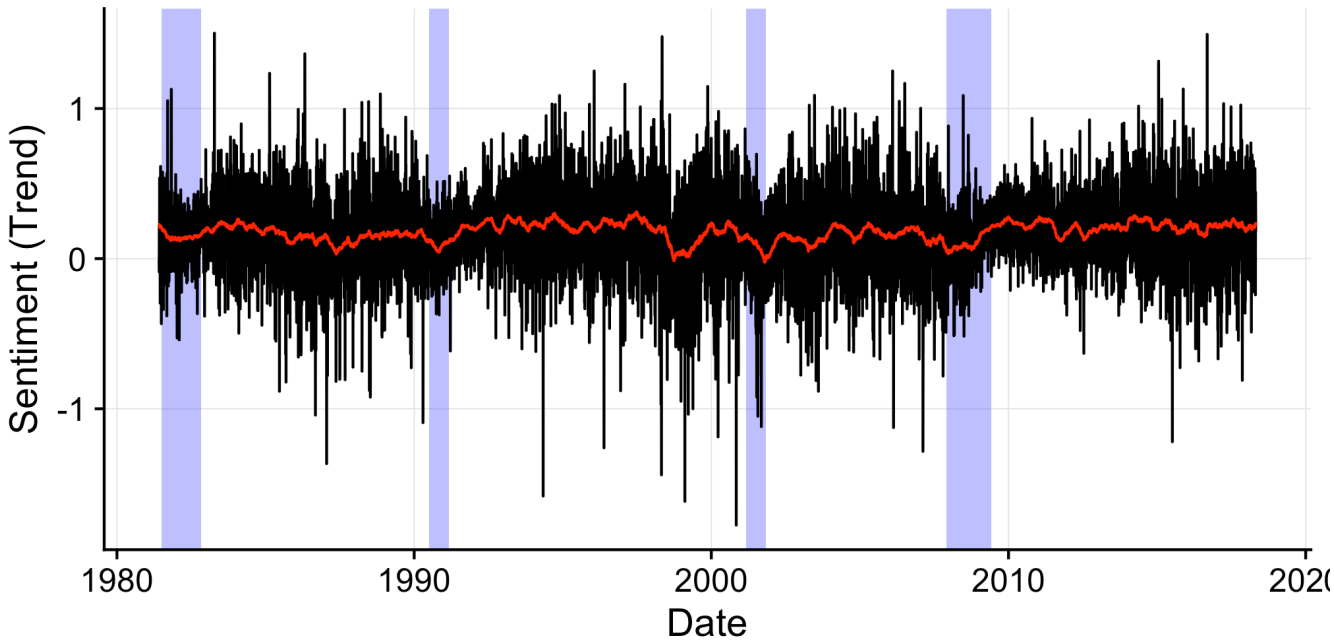
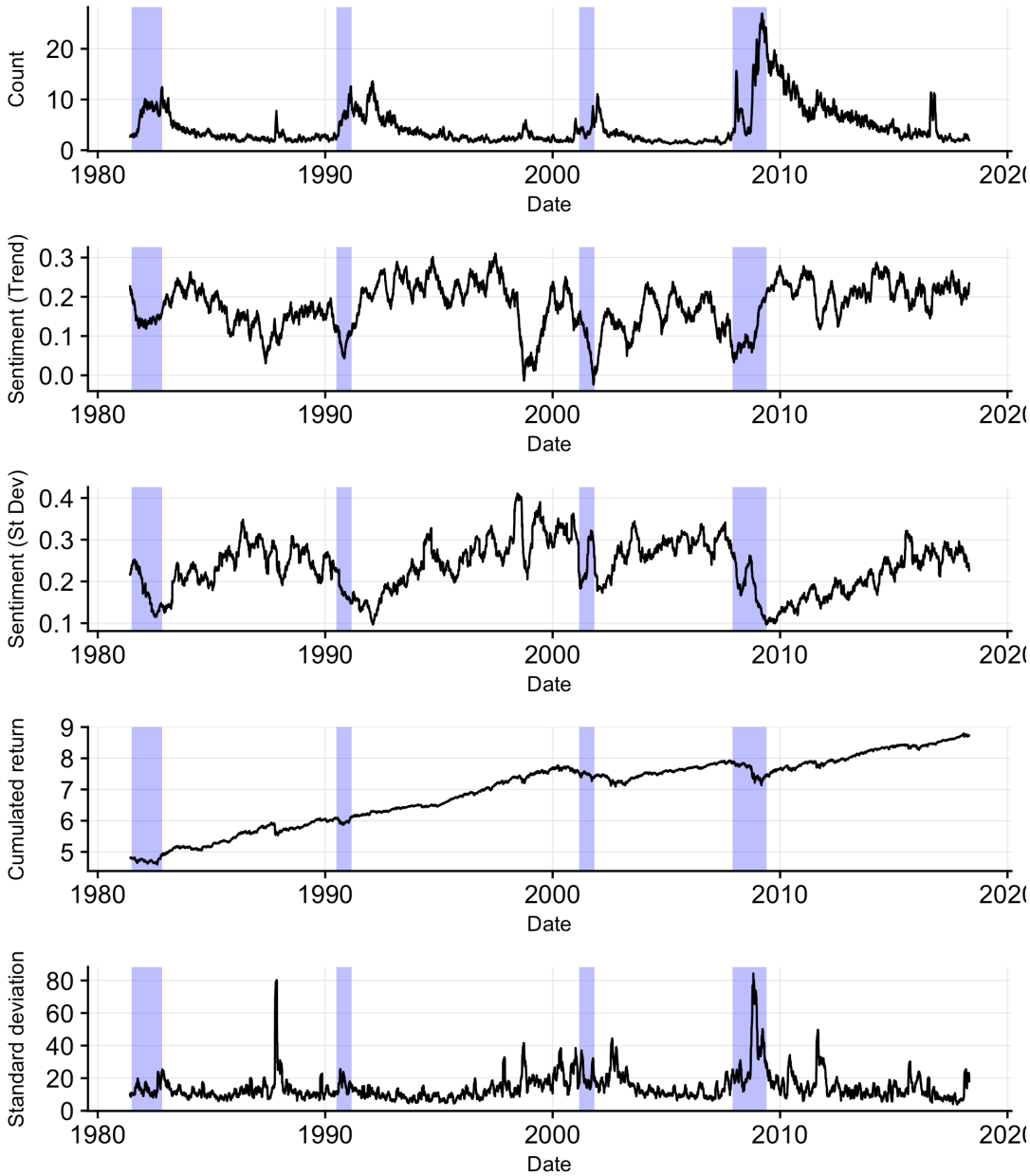


Figure: The figure shows the mean sentiment over all articles of one day. The red line is the 100 day moving average.

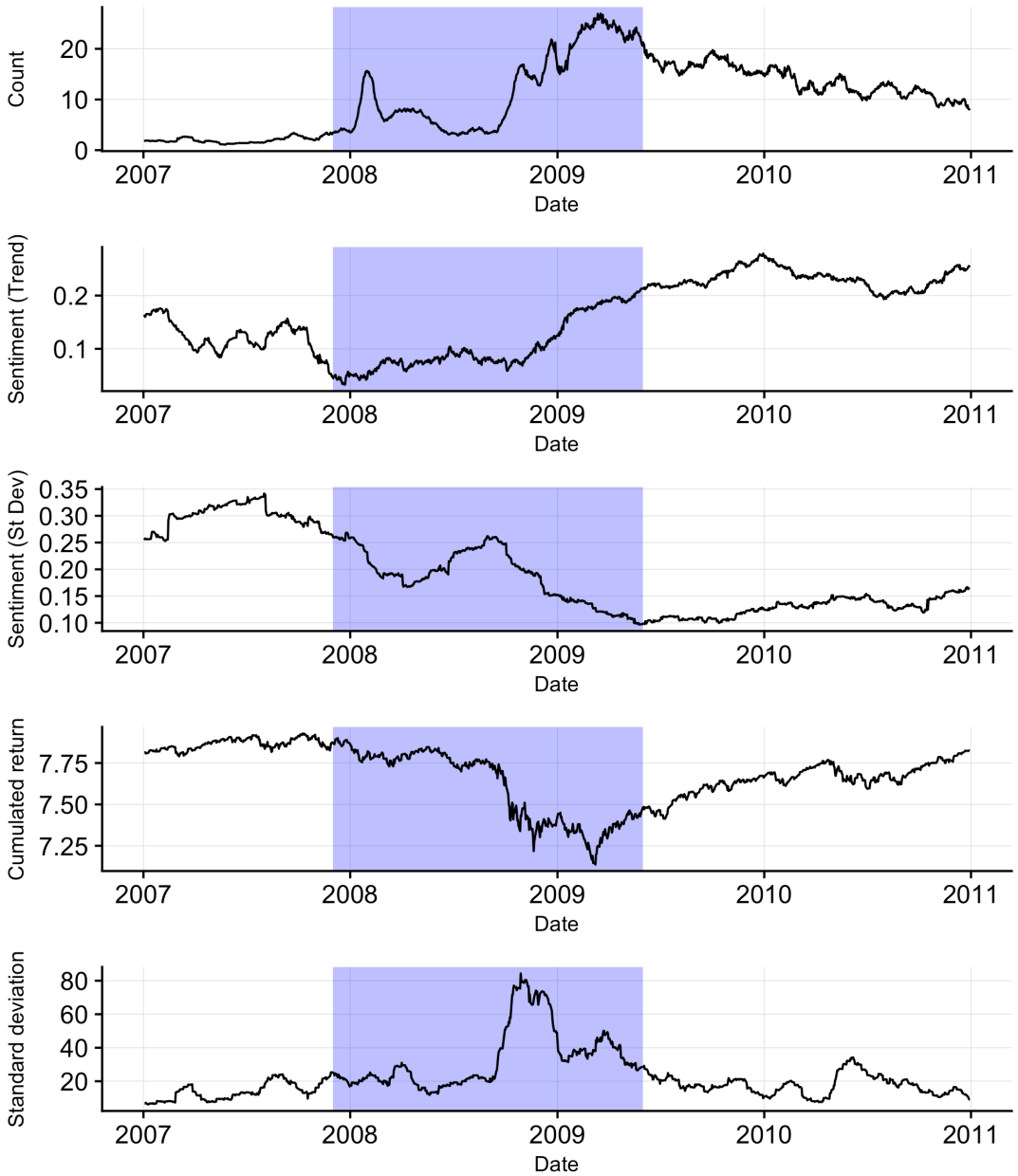
Sentiment: Trend and dispersion (All)



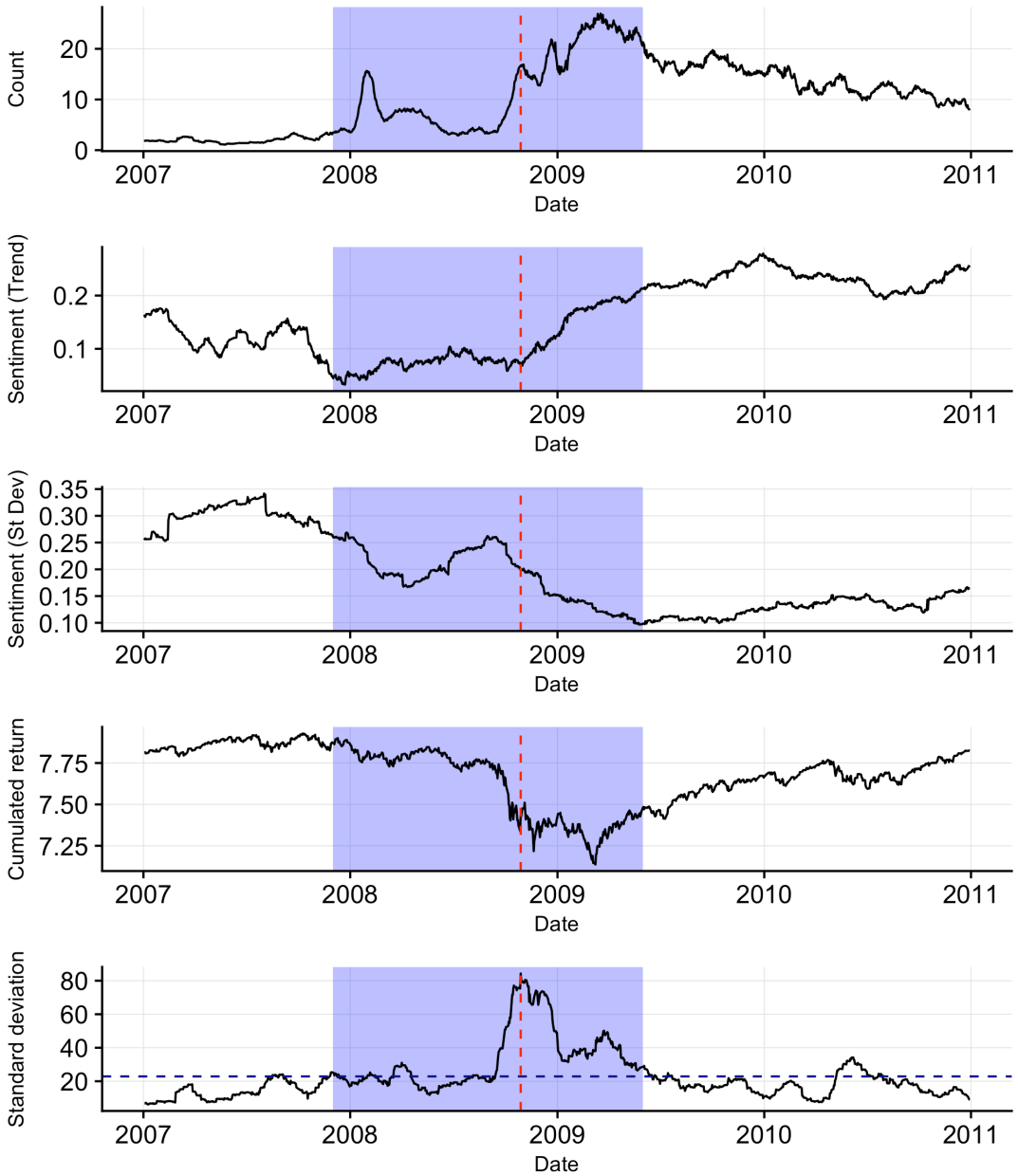
Observations

1. In the recessions of 1982, 1991, 2002, 2008 sentiment has its lowest point around the stock market low.
2. Sentiment trended slowly down ahead of 1987
3. In the recessionary environment of 1998/1999 (Asian crises) and 2011 (European crises) sentiment went down and volatility shot up while prices do not trended down.
4. The volatility of the sentiment goes down in every recession and bottoms out at the end of each recession.

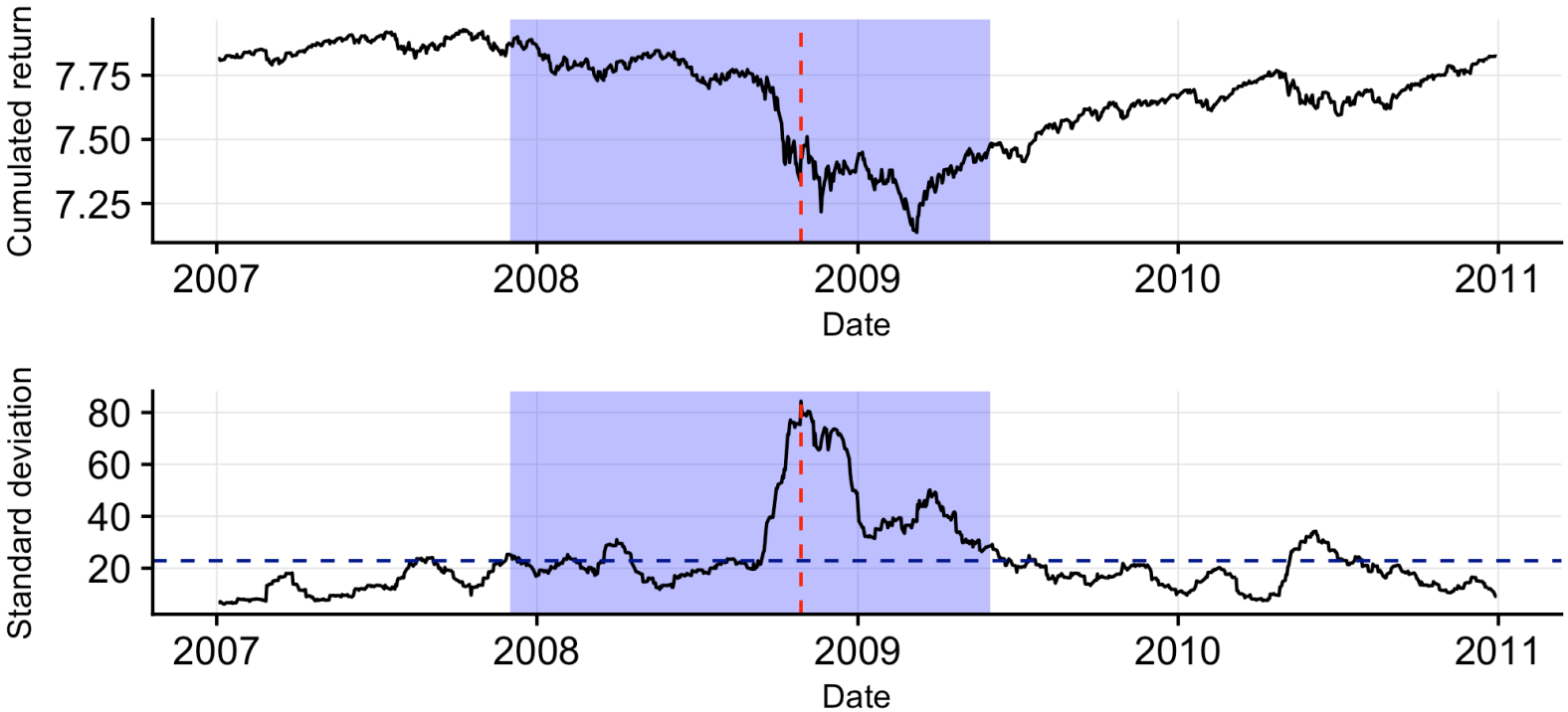
Sentiment: Trend and dispersion (2008)



Sentiment: Trend and dispersion (2008)



Sentiment: Trend and dispersion



	Mean standard deviation	Mean return
Above the blue line	38.77%	3.68%
Below the blue line	15.62%	4.03%
Before the red line	19.67%	-17.22%
After the red line	25.55%	21.70%

Conclusion

- **We define risk as the change of volatility and not the level.**
- The **short-term** relation between stock market risk and return is **negative**.
- We construct **a daily index** which measures the **sentiment of articles of the NYTimes which contain the word "recession".**
- **Trading on the level of volatility makes sense. But it is a strategy with it's typical advantages and drawbacks and it is not a puzzle** as Moreira and Muir (2017) write. Because the major risk for the investor is the change in volatility not the level of volatility.

SINGLE SECURITIES IN THE CROSS-SECTION

The cross-section of expected stock returns

- **A lot of data: More than 450 variables have been documented**
  See e.g. Hou et al. (2017)

The cross-section of expected stock returns

- **A lot of data: More than 450 variables have been documented**
  See e.g. Hou et al. (2017)
- **Higher complexity: Variable interactions and non-linearities. e.g. momentum returns are higher for stocks with high idiosyncratic volatility**
  Bandarchuk and Hilscher (2012)

The cross-section of expected stock returns

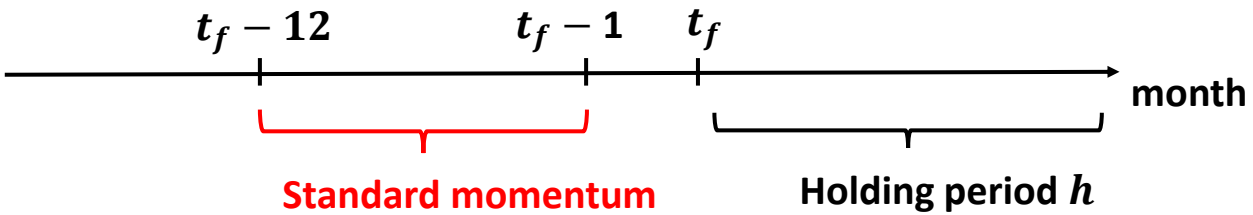
- **A lot of data: More than 450 variables have been documented**
  See e.g. Hou et al. (2017)
- **Higher complexity: Variable interactions and non-linearities. e.g. momentum returns are higher for stocks with high idiosyncratic volatility**
  Bandarchuk and Hilscher (2012)
- **Traditional methods have problems in evaluating this information jointly**
  See e.g. Fama and French (2015)

The cross-section of expected stock returns

- **A lot of data: More than 450 variables have been documented**
  See e.g. Hou et al. (2017)
- **Higher complexity: Variable interactions and non-linearities. e.g. momentum returns are higher for stocks with high idiosyncratic volatility**
  Bandarchuk and Hilscher (2012)
- **Traditional methods have problems in evaluating this information jointly**
  See e.g. Fama and French (2015)
- **As a result: Overfitting and data mining looms large**
  Harvey et al. (2016)

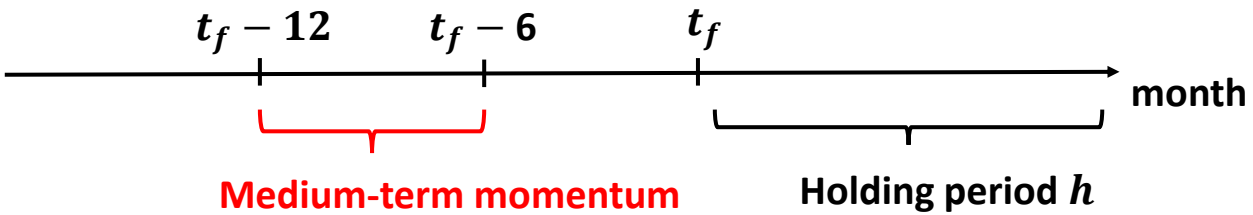
Example: Application using return-based variables

Jegadeesh and Titman (1993)



Example: Application using return-based variables

Novy-Marx (2012)



Example: Application using return-based variables

factors

- 12-2 Momentum: Jegadeesh and Titman (1993)
- 12-7 Momentum: Novy-Marx (2012)
- 6-2 Momentum: Goyal and Wahal (2015)
- 1 Reversal: Jegadeesh (1990), Lehmann (1990)
- 12 Saisonality: Heston and Sadka (2008)

conditionality in factors

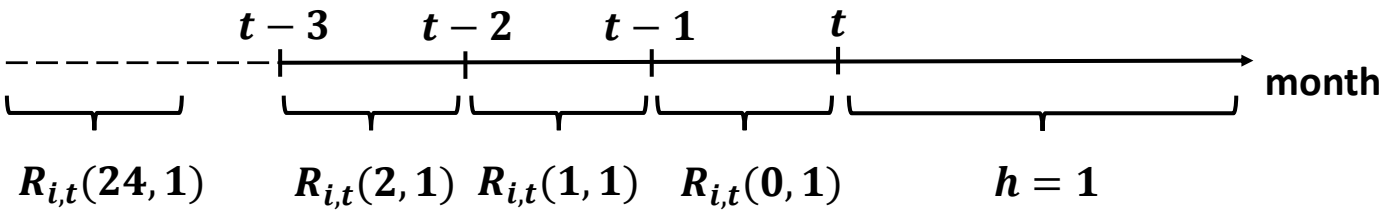
- High 12-2 Momentum in stocks with more extreme past returns: Bandarchuk and Hilscher (2012)

non-linearity in factors

- The winner portfolio is mostly responsible for 12-2 Momentum returns: Jegadeesh and Titman (1993)

Example: Application using return-based variables

Consider the case: $h = 1; l = 1; g = 0, \ldots, 24$:



Main results

We can show:

- which data is important when analyzed jointly
- interactions and non-linearity in the data
- behavior over time

Main results

We can show:

- which data is important when analyzed jointly
- interactions and non-linearity in the data
- behavior over time

... and that utilizing this information efficiently leads to better predictions:

	Mean	StDev	Mean/StDev	Method
12-2 Momentum	1.6	1.3	1.2	PS
12-7 Momentum	1.3	1.0	1.3	PS
6-2 Momentum	1.2	1.2	1.1	PS
1 Reversal	-1.8	1.1	-1.7	PS
12 Saisonality	0.9	0.8	1.2	PS
1, ... , 25	2.3	0.8	3.0	TBCPS

PS = Portfolio sorts, TBCPS = Tree-based conditional portfolio sorts

1968-2012

Investor problem

- General model of expected returns

$$E_t[r_{i,t+1}|R_{i,t}(0,1),\ldots,R_{i,t}(24,1)] = f_t(R_{i,t}(0,1),\ldots,R_{i,t}(24,1))$$

Investor problem: Modeling devices

$$E_t[r_{i,t+1}|R_{i,t}(0,1),\ldots,R_{i,t}(24,1)] = f_t(R_{i,t}(0,1),\ldots,R_{i,t}(24,1))$$

1. Fama-MacBeth regression

Investor problem: Modeling devices

$$E_t[r_{i,t+1}|R_{i,t}(0,1),\ldots,R_{i,t}(24,1)] = f_t(R_{i,t}(0,1),\ldots,R_{i,t}(24,1))$$

1. Fama-MacBeth regression
   - Use linear model

$$r_{i,t+1} = a + \sum_{g=0}^{24} \beta_g^t R_{i,t}(g,1) + \epsilon_{i,t}$$

Investor problem: Modeling devices

$$E_t[r_{i,t+1}|R_{i,t}(0,1),\ldots,R_{i,t}(24,1)] = f_t(R_{i,t}(0,1),\ldots,R_{i,t}(24,1))$$

1. Fama-MacBeth regression
   - Use linear model

$$r_{i,t+1} = a + \sum_{g=0}^{24} \beta_g^t R_{i,t}(g,1) + \epsilon_{i,t}$$

   - Make predictions based on coefficients estimated and averaged over the past m months

Investor problem: Modeling devices

$$E_t[r_{i,t+1}|R_{i,t}(0,1),\ldots,R_{i,t}(24,1)] = f_t(R_{i,t}(0,1),\ldots,R_{i,t}(24,1))$$

1. Fama-MacBeth regression

Investor problem: Modeling devices

$$E_t[r_{i,t+1}|R_{i,t}(0,1),\ldots,R_{i,t}(24,1)] = f_t(R_{i,t}(0,1),\ldots,R_{i,t}(24,1))$$

1. Fama-MacBeth regression
2. Tree-based conditional portfolio sort

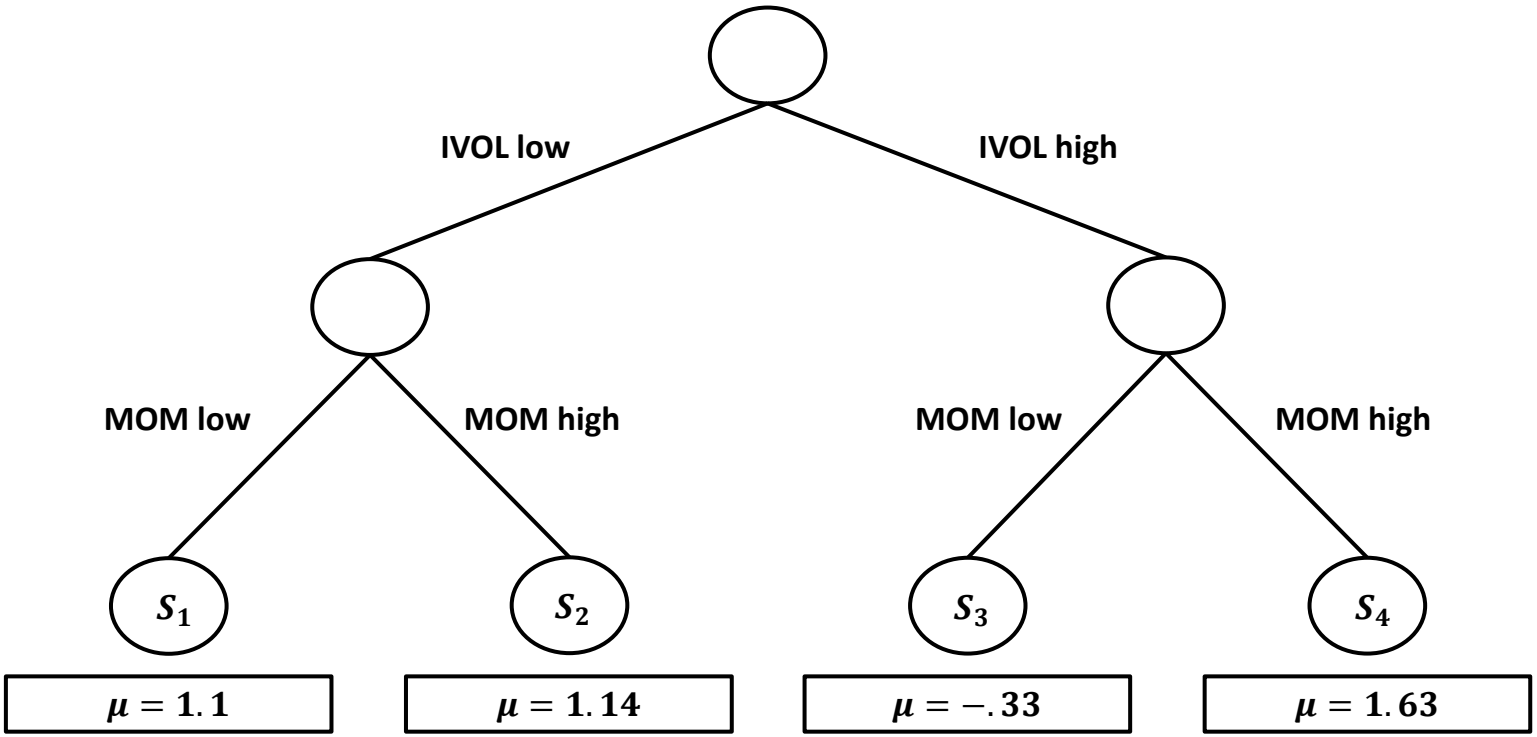
Investor problem: Modeling devices

$$E_t[r_{i,t+1}|R_{i,t}(0,1),\ldots,R_{i,t}(24,1)] = f_t(R_{i,t}(0,1),\ldots,R_{i,t}(24,1))$$

1. Fama-MacBeth regression
2. Tree-based conditional portfolio sort
   - Can we estimate the prediction equation more flexibly?
   - Adopt a machine learning approach

Conditional portfolio sort

From Bandarchuk and Hilscher (2012)

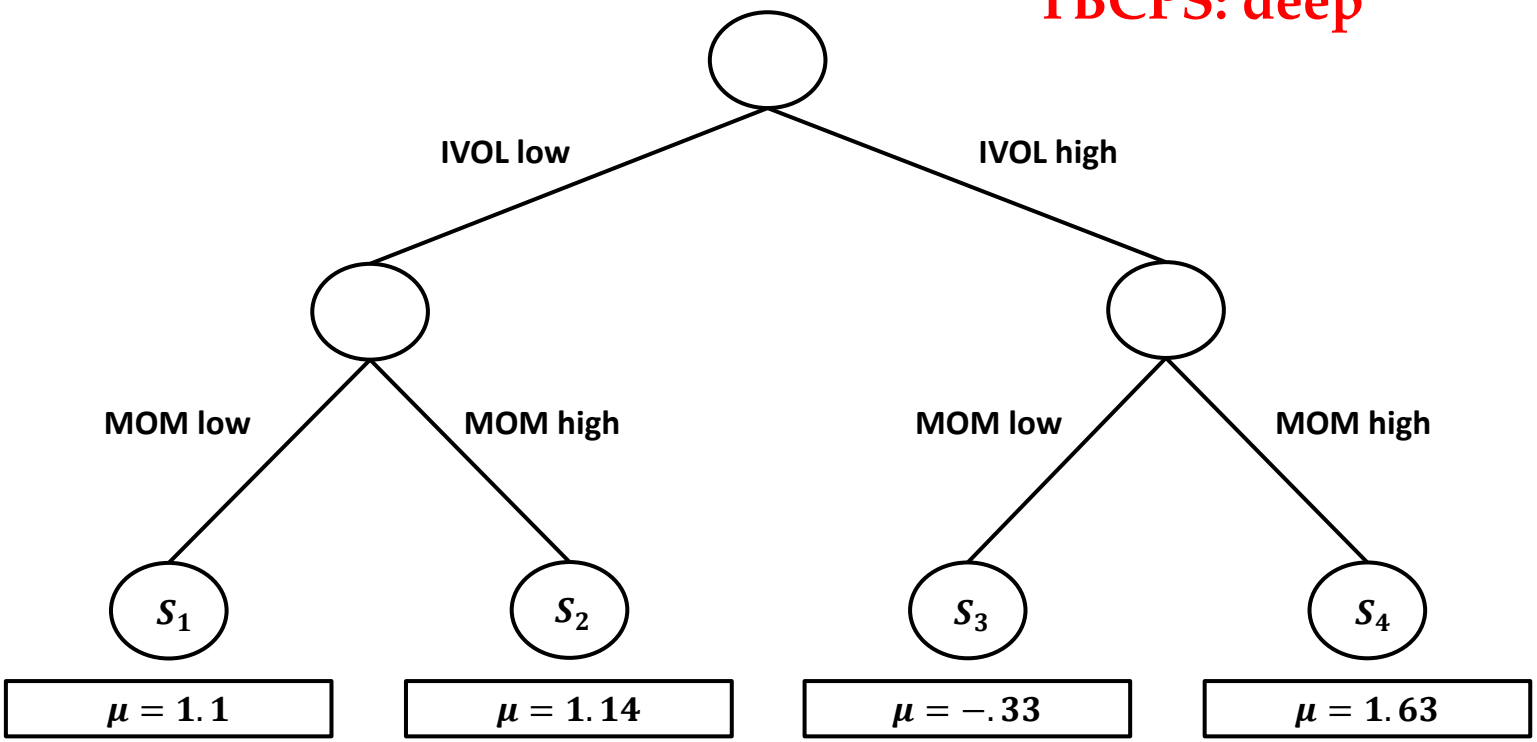


Tree-based conditional portfolio sort

From Bandarchuk and Hilscher (2012)

CPS: Stop after two levels

TBCPS: deep

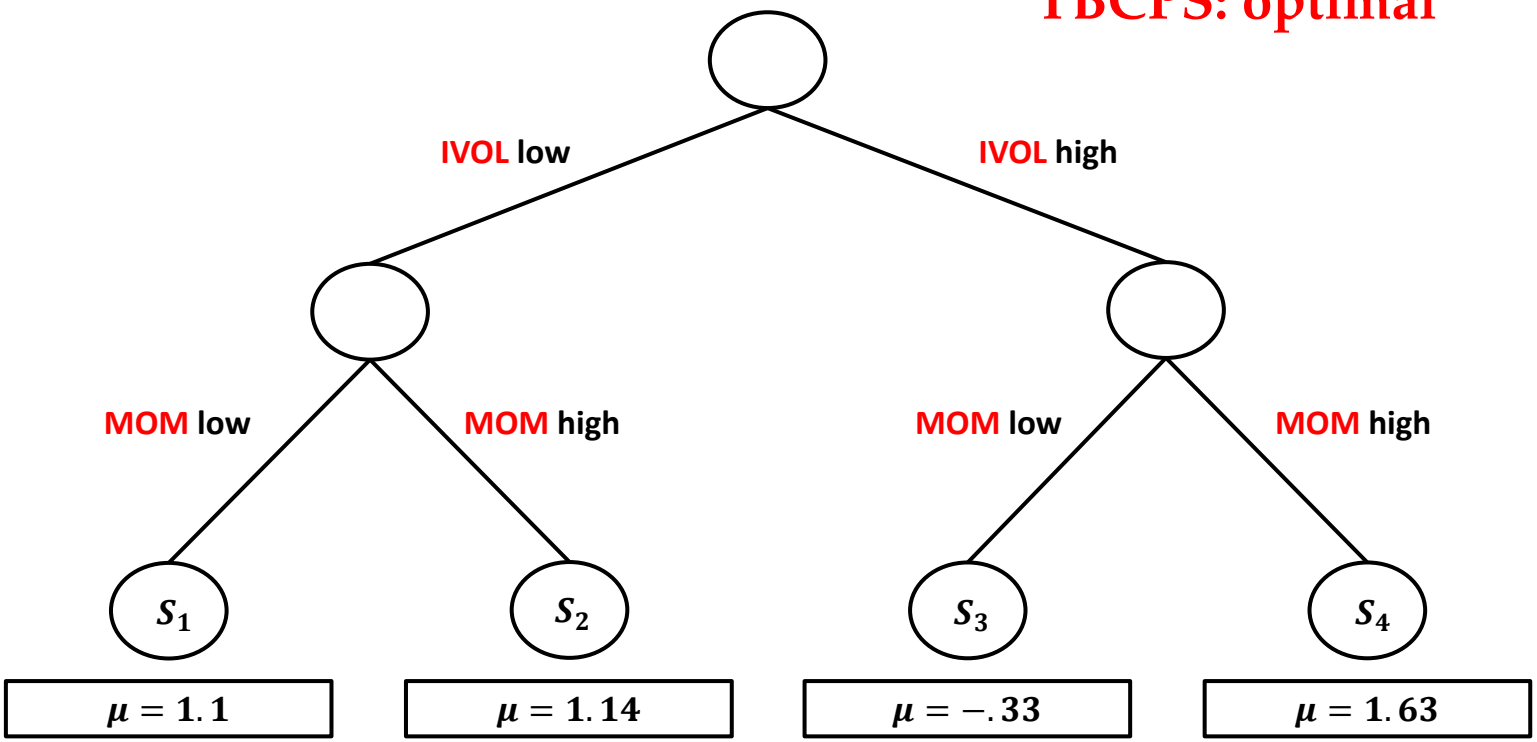


Tree-based conditional portfolio sort

From Bandarchuk and Hilscher (2012)

CPS: Sorting variables selected ex-ante

TBCPS: optimal

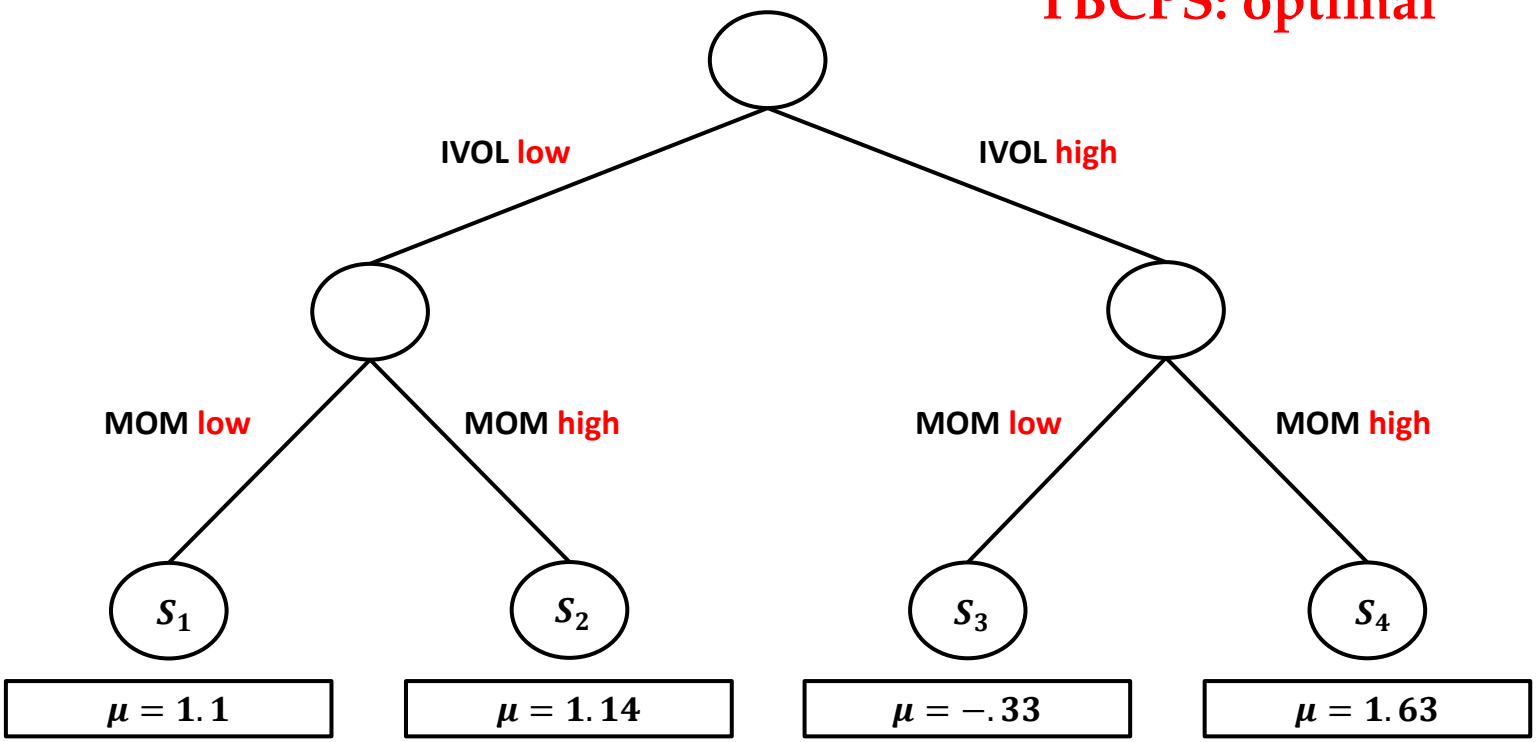


Tree-based conditional portfolio sort

From Bandarchuk and Hilscher (2012)

CPS: Sorting values selected ex-ante

TBCPS: optimal



Random forest

- But decision-trees turn out to be *weak learners*

Random forest

- But decision-trees turn out to be *weak learners*
- *Ensemble methods* combine many weak learners

Random forest

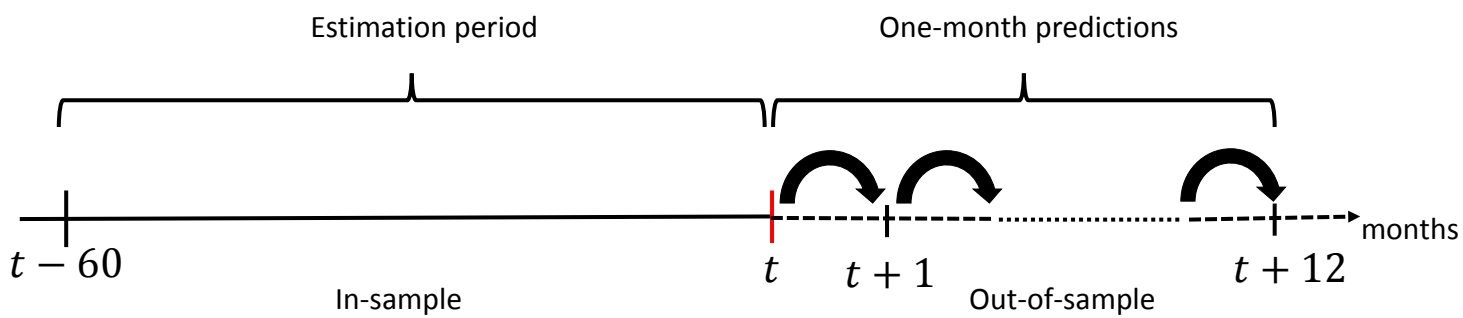
- But decision-trees turn out to be *weak learners*
- *Ensemble methods* combine many weak learners
- Idea - Majority voting

Random forest

- But decision-trees turn out to be *weak learners*
- *Ensemble methods* combine many weak learners
- Idea - Majority voting
- **Random forests** de-correlate individual decision-trees by bootstrapping different observations into the training samples of each tree and by randomly choosing a subset of the predictor variables at each node.

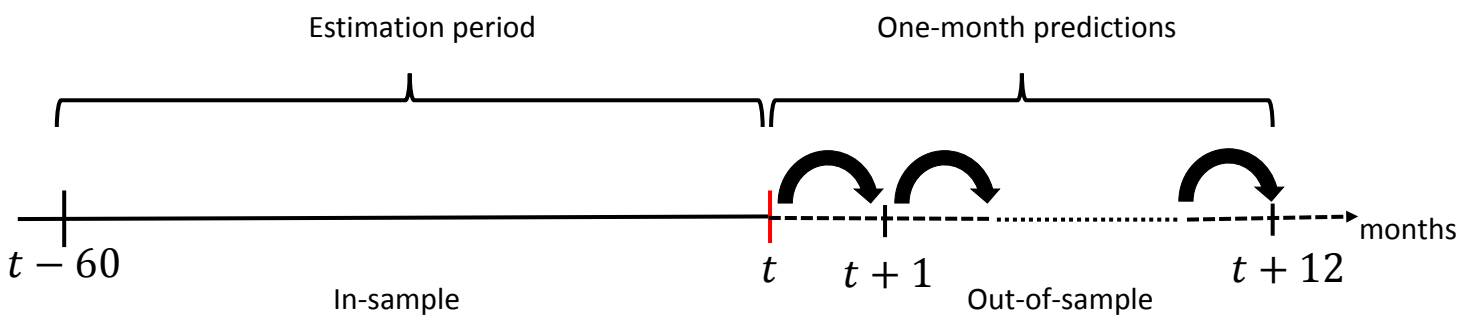
Out-of-sample testing procedure

1. At beginning of each year, estimate model with data over past 60 months



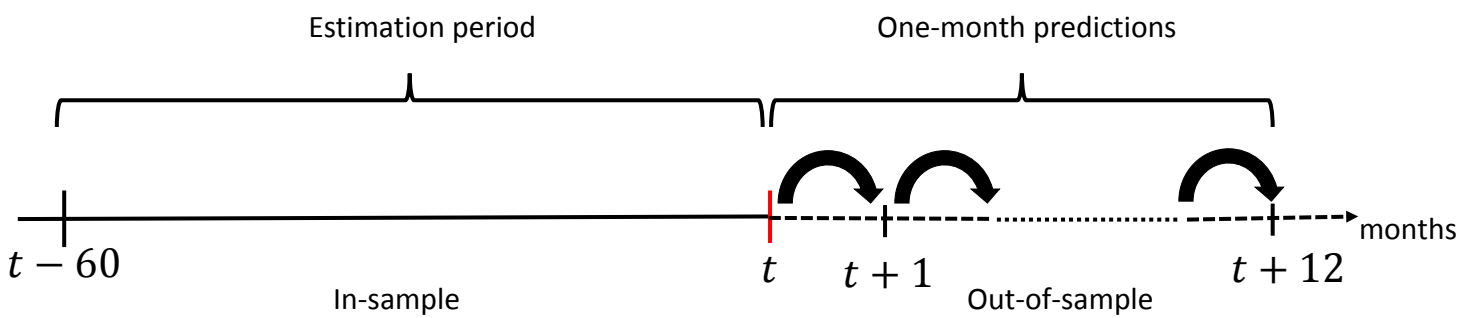
Out-of-sample testing procedure

1. At beginning of each year, estimate model with data over past 60 months
2. Fix model estimates over the next year



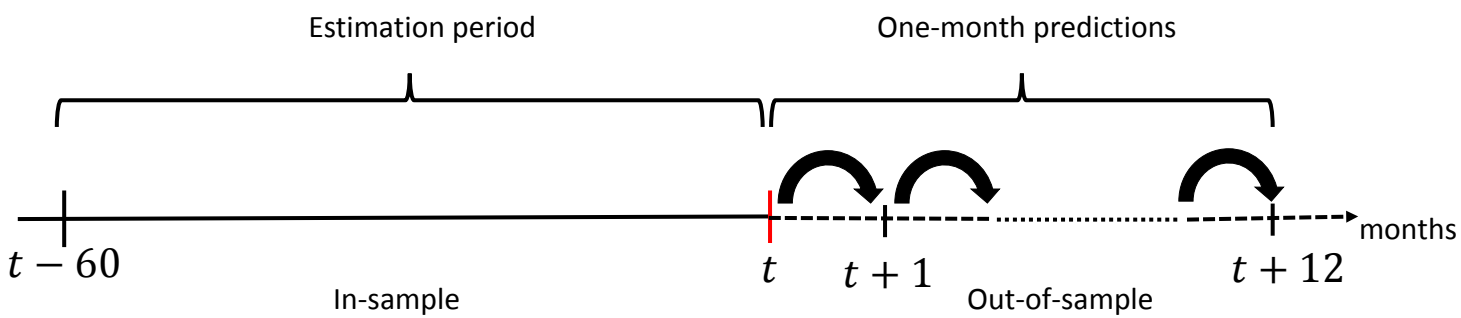
Out-of-sample testing procedure

1. At beginning of each year, estimate model with data over past 60 months
2. Fix model estimates over the next year
3. Predict next months' returns each month: buy highest decile of predictions, sell lowest decile of predictions



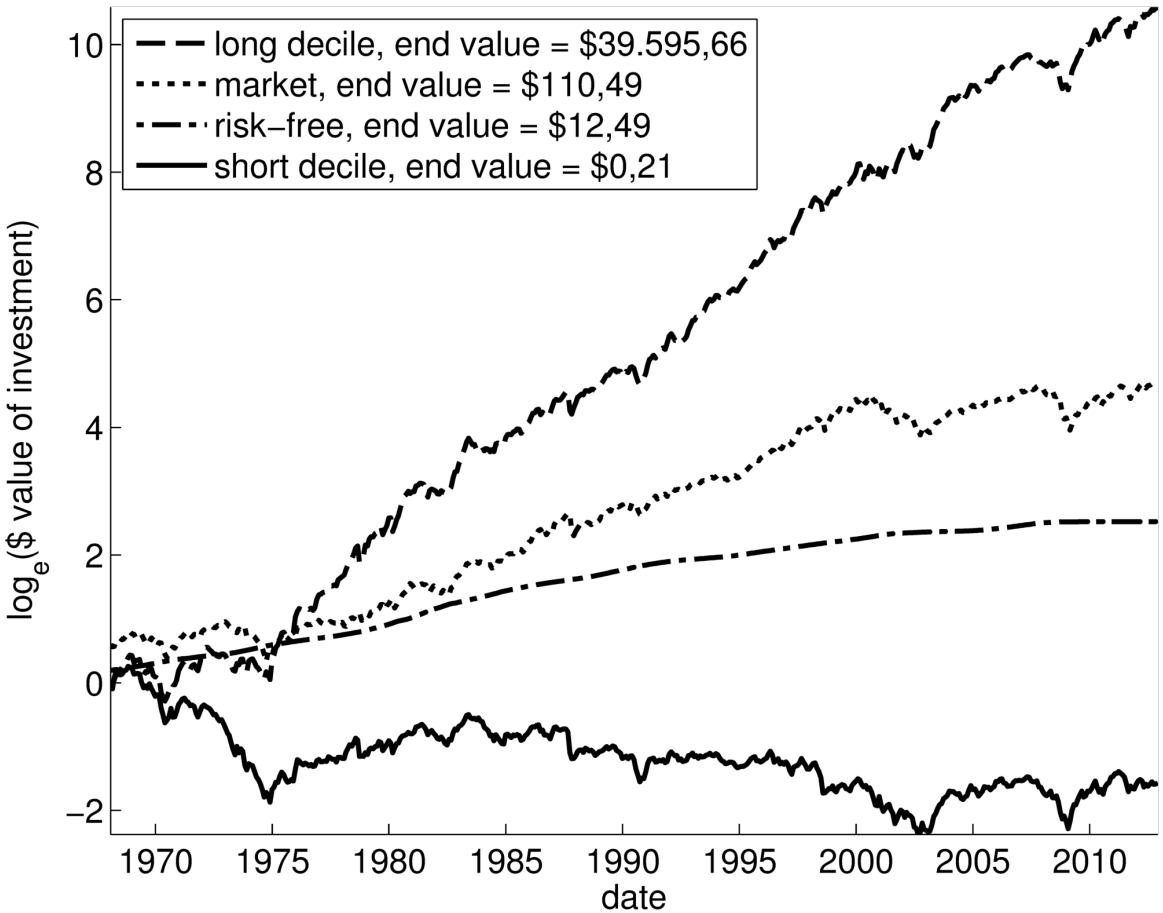
Out-of-sample testing procedure

1. At beginning of each year, estimate model with data over past 60 months
2. Fix model estimates over the next year
3. Predict next months' returns each month: buy highest decile of predictions, sell lowest decile of predictions
4. Re-calibrate model at beginning of next year



Earned profit from investing $1 in 1968

Return-based, non-overlapping, rolling



Variable importance

Sort by median rank over 45 years

	Median rank
R(0,1)	1
R(1,1)	4
R(2,1)	4
R(3,1)	6
R(11,1)	7
R(4,1)	8
R(5,1)	8
R(8,1)	11
R(10,1)	11
R(9,1)	12

R(g,1) = one-month return g months ago

Variable importance

Two well-known past returns rank highly on the list

	Median rank
R(0,1)	1
R(1,1)	4
R(2,1)	4
R(3,1)	6
R(11,1)	7
R(4,1)	8
R(5,1)	8
R(8,1)	11
R(10,1)	11
R(9,1)	12

R(g,1) = one-month return g months ago

Variable importance

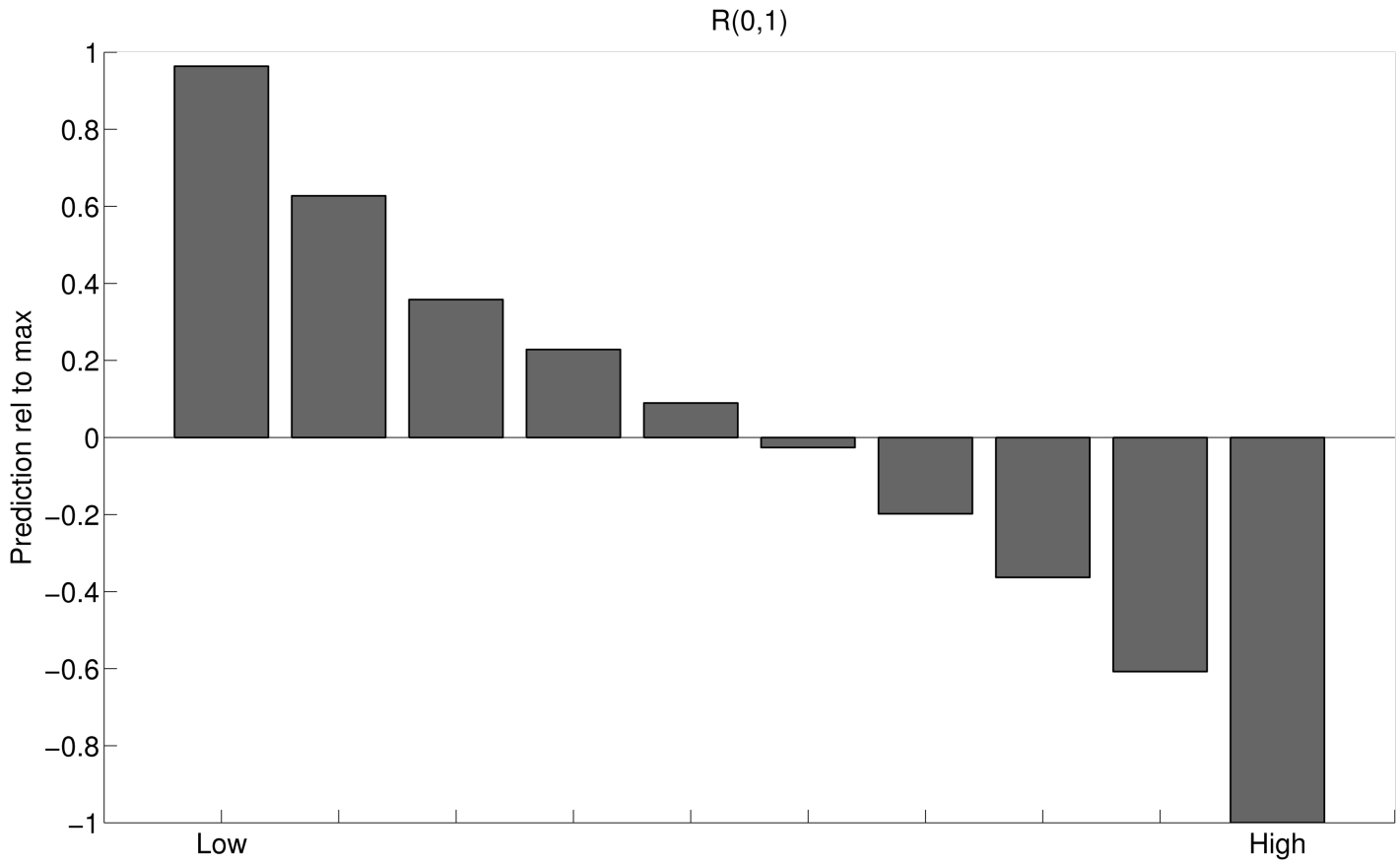
Mostly recent returns show up

	Median rank
R(0,1)	1
R(1,1)	4
R(2,1)	4
R(3,1)	6
R(11,1)	7
R(4,1)	8
R(5,1)	8
R(8,1)	11
R(10,1)	11
R(9,1)	12

R(g,1) = one-month return g months ago

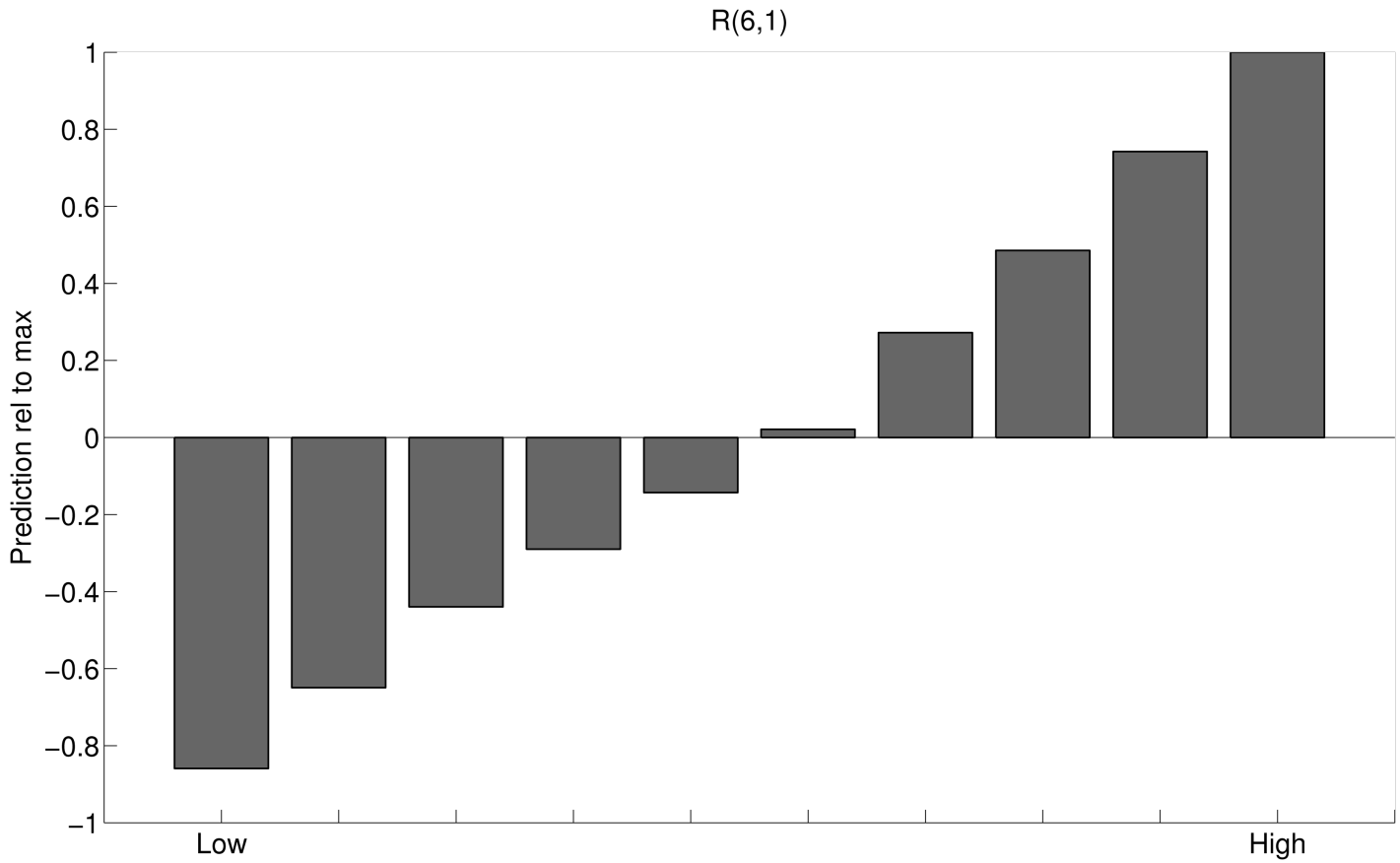
Partial derivatives

Find short-term reversal for R(0,1)



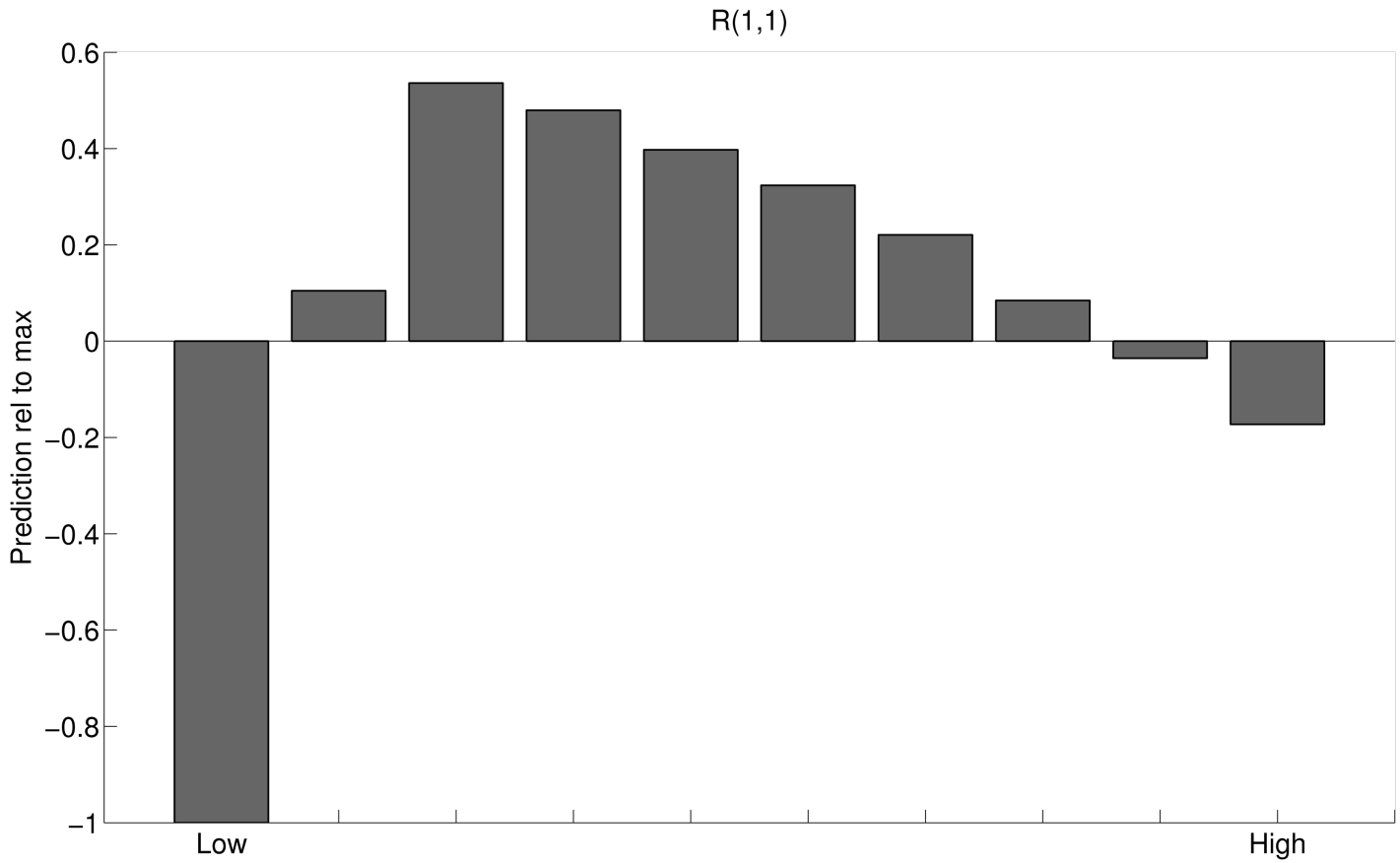
Partial derivatives

Find momentum for R(6,1)



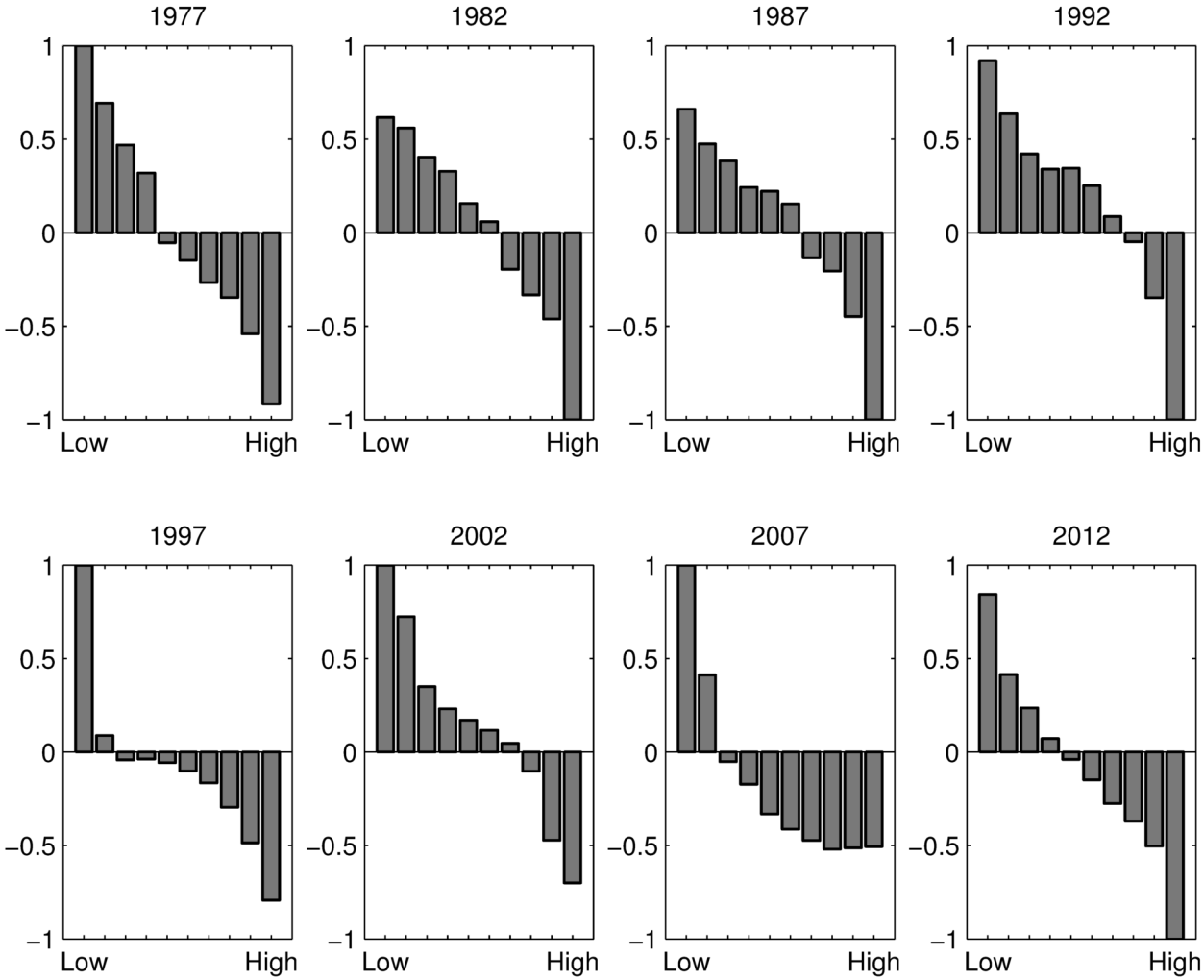
Partial derivatives

Find non-linear relationship for R(1,1)



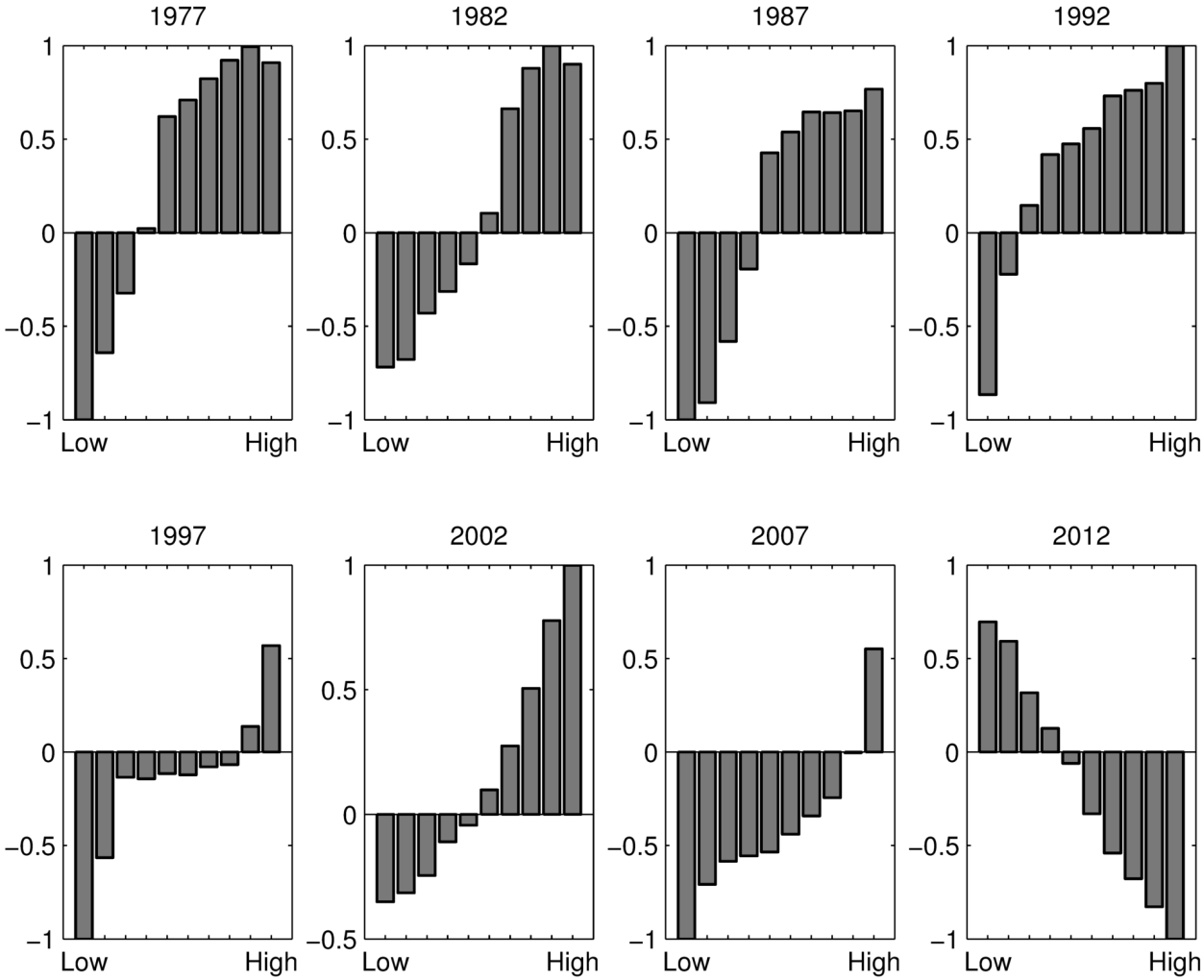
Partial derivatives over time

Stable for e.g. R(0,1)

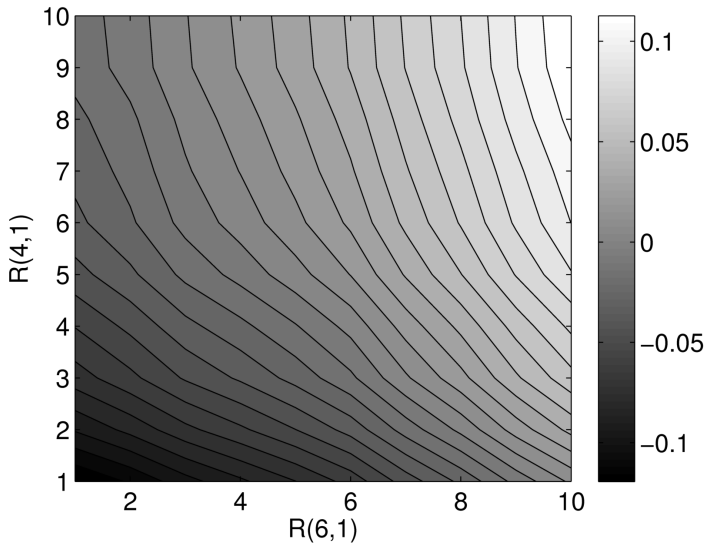


Partial derivatives over time

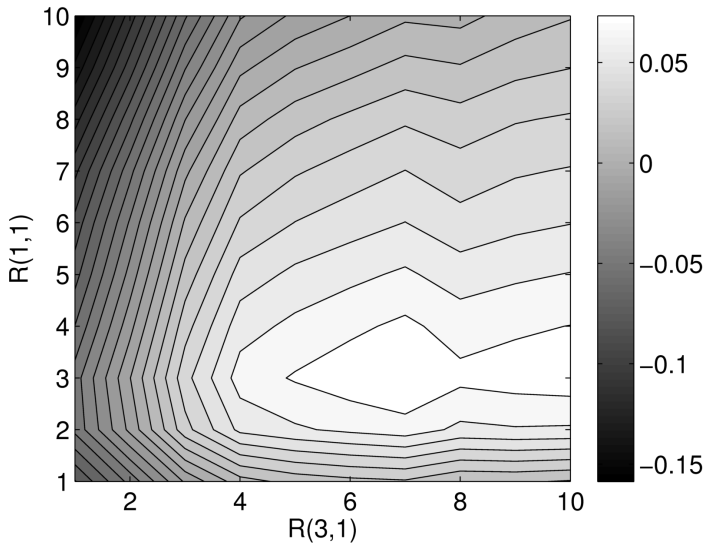
Interesting reversal of R(6,1) recently (momentum crash)



Double partial derivatives



(a) Increasing in both



(b) Interesting interaction

Conclusion

- The number of factors in the cross-section has been grown exponentially.
- Traditional methods used on cross-sectional asset-pricing are older than 40 years. A new methodology is needed.
- We import ideas from the machine learing literature (random forest) to extend the conditional portfolio sort to tree-based conditional portfolio sort and evaluate through variable importance and ”partial derivatives”

Final conclusion

- The technical innovations had two important consequences:
  1. Much easier to use advanced statistical methods
  2. New data sources are available due to newly collected data or simply due to the digitalization of existing data.
- Results:
  - Risk-return relation of the aggregate market: We use new (digitized) data from the New York Times. Dictionaries are used to process the ten thousands of texts to extraxt the sentiment.
  - Risk-return relation of single securities: The state-of-the-art method is from 1973 and not longer suitable. The random forest, the handle the high-dimensional problem much better.
- Outlook:
  - One feature of this work is the use of default values in each setting. It is the strength of this results, that nothing has been tuned. The tuning of the parameters is still open.
  - Aggregate market: 1. other search words like economic expansion and bubble; 2. topic modelling algorithm; 3. other finance-specific lexicons; 4. other newspaper sources
  - Single securities: 1. other statistical methods; 2. stock market of other countries; 3. Extending the variable set

THANK YOU.

APPENDIX TS

Appendix - Aggregate market

- Theory: Risk-return in the long-run ▸ Details
- Empirics: Moreira/Muir Analysis ▸ Details
- Theory: DDM ▸ Details
- Empirics: Data description ▸ Details
- Theory: Literature Textual Analysis ▸ Details
- Theory: Syuzhet dictionary (Jockers) ▸ Details
- Theory: Dictionaries in comparison ▸ Details
- Theory: Dictionaries histogram ▸ Details
- Empirics: Subperiods ▸ Details
- Empirics: R packages ▸ Details

Risk-return relation in the long-run

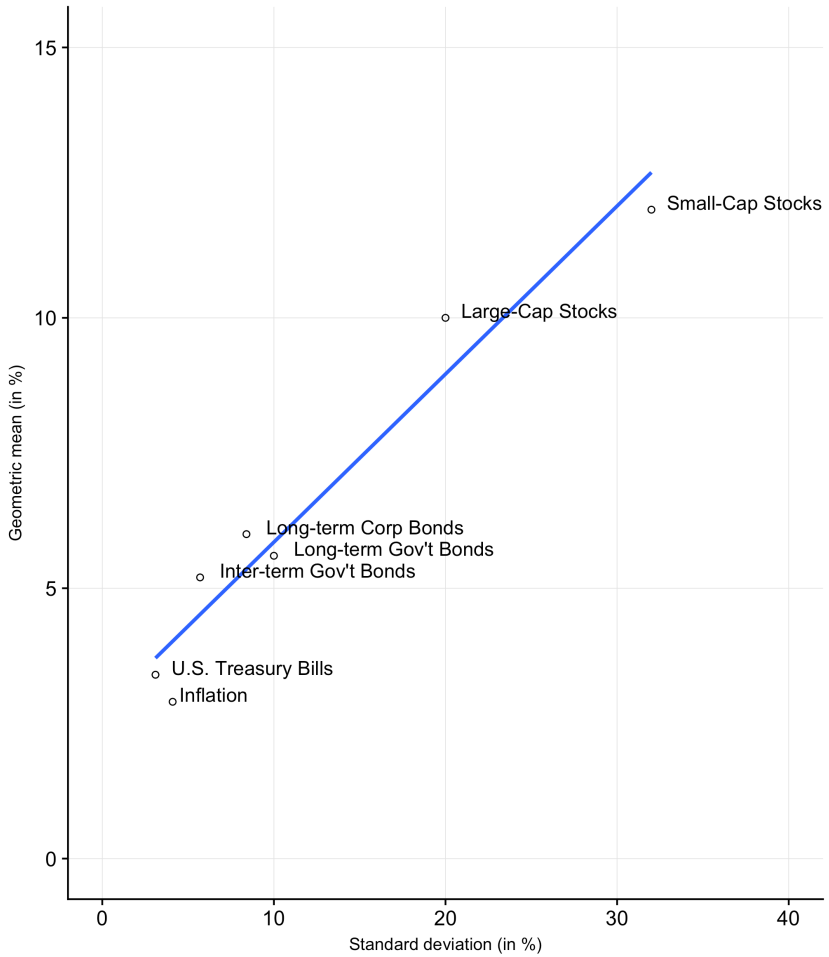
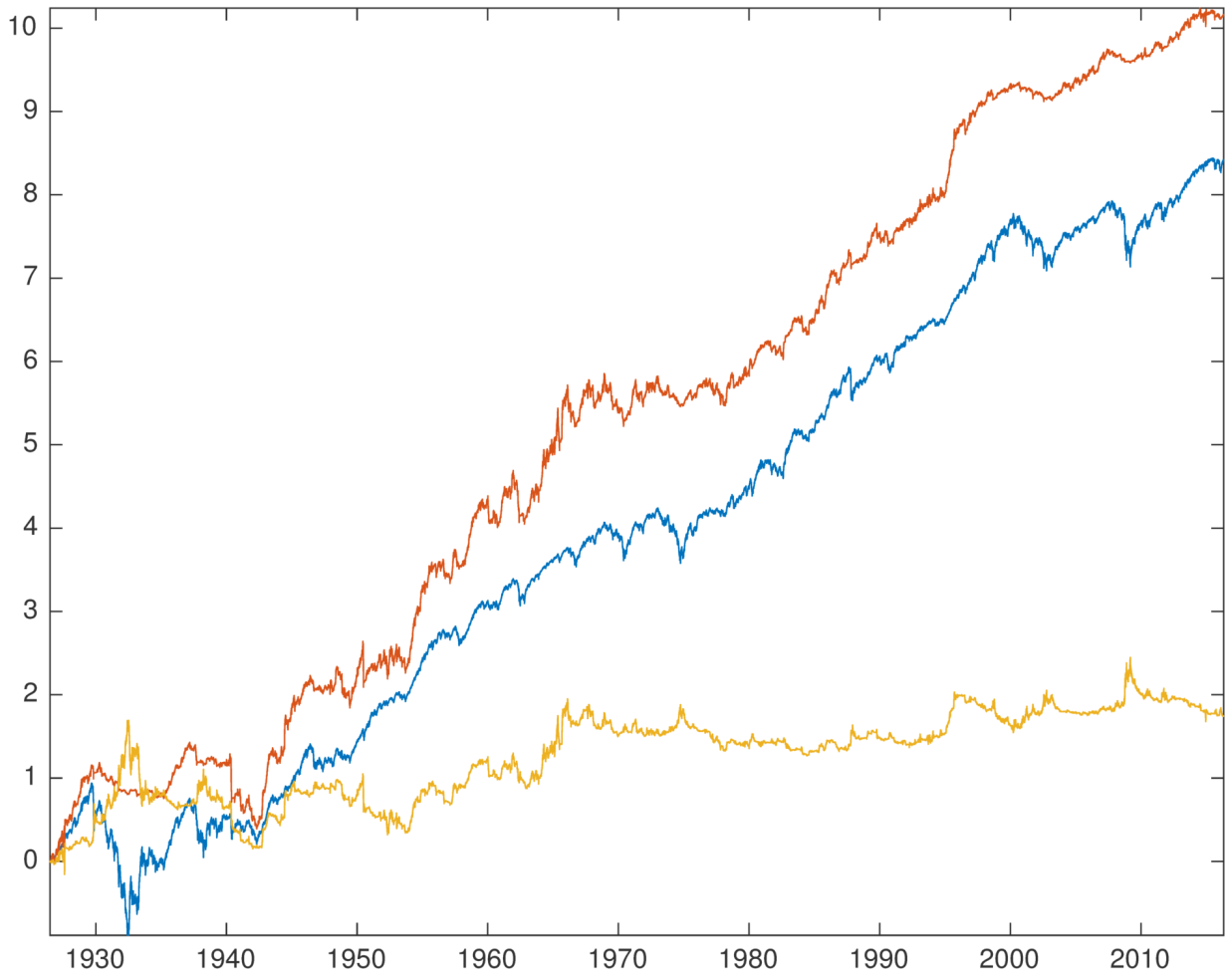


Figure: From 1926 to 2015. Source is Ibbotson et al. (2016).

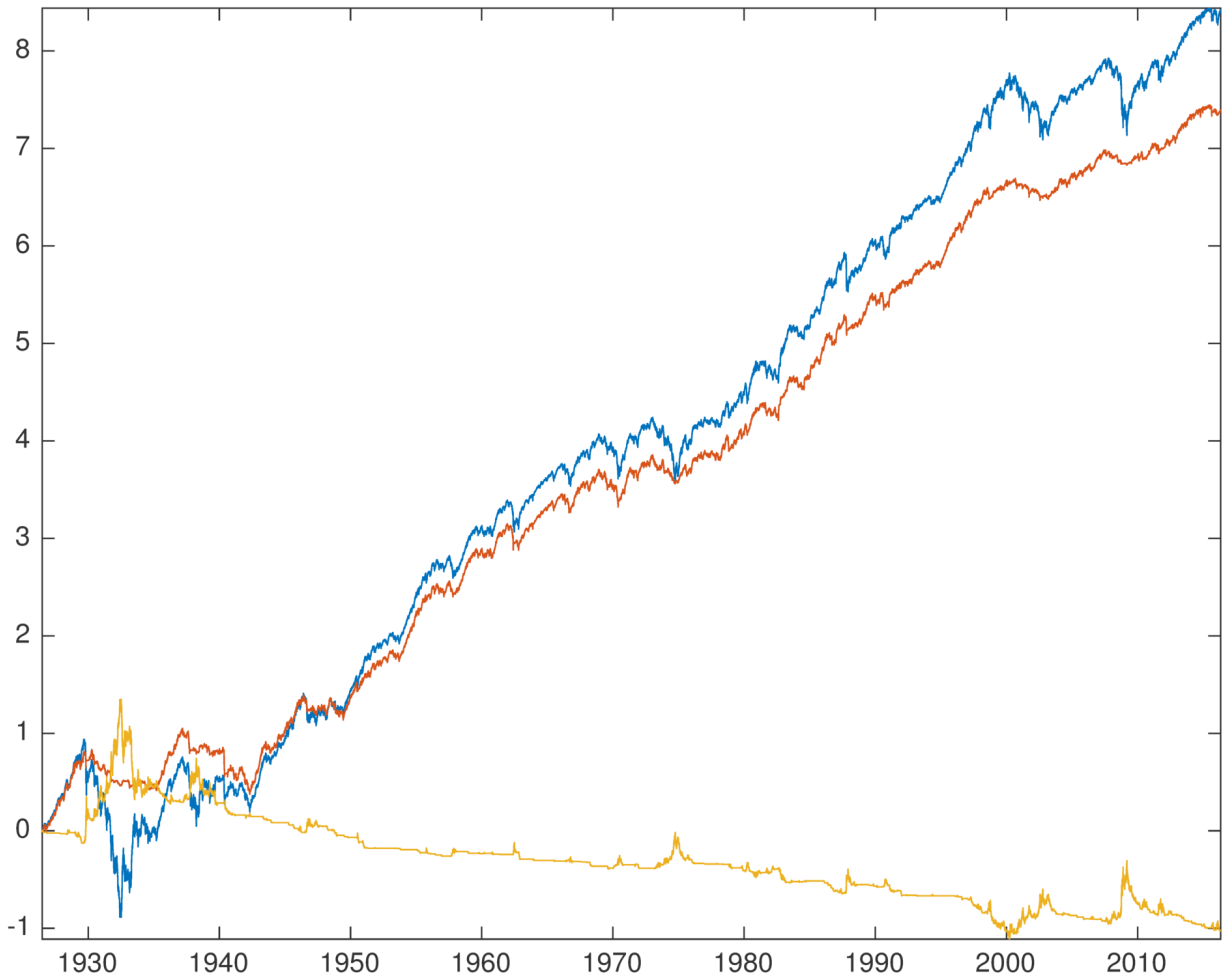
Back

Moreira / Muir



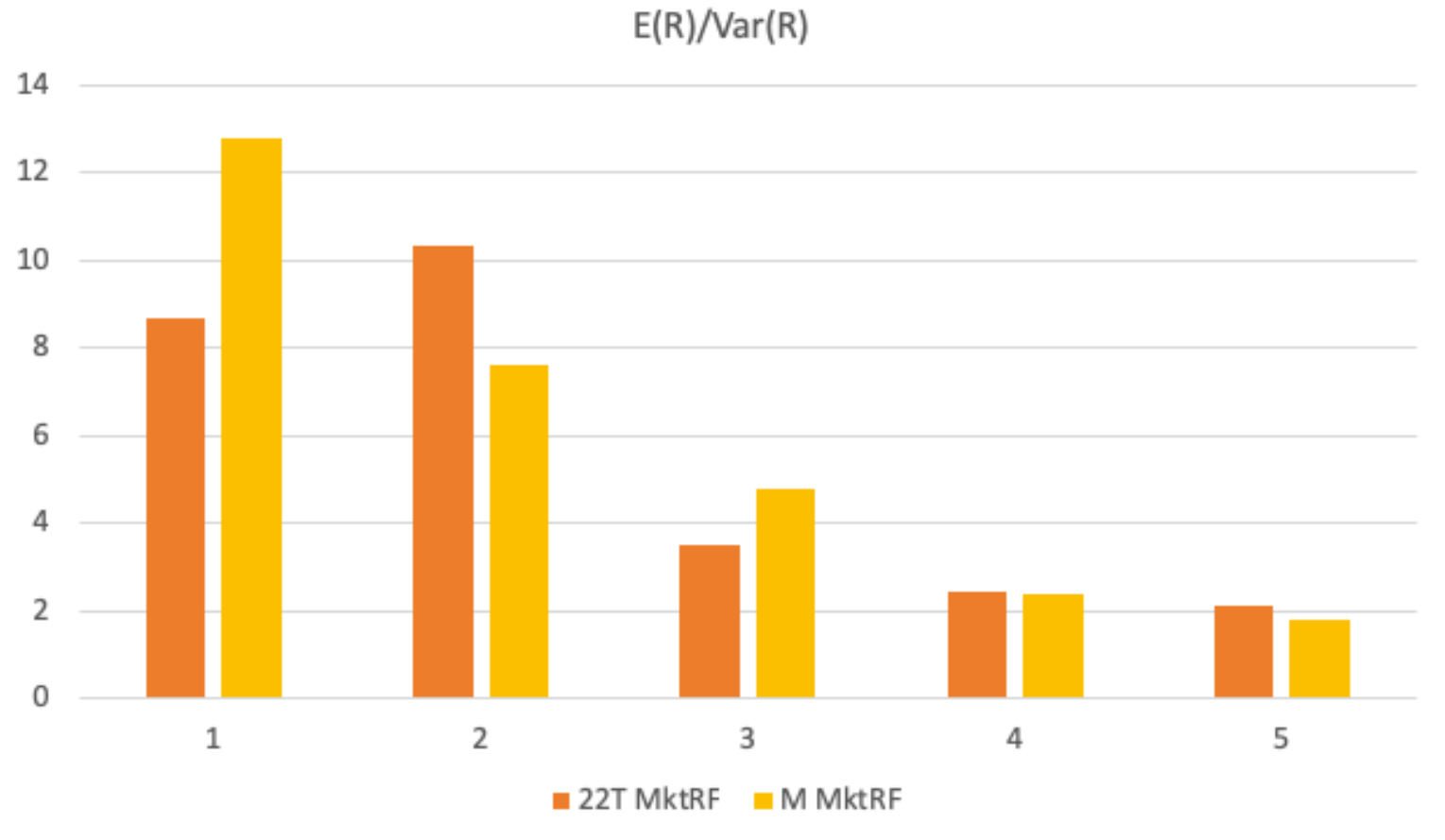
Back

Moreira / Muir (Leverage = 1)



Back

Moreira / Muir



Back

What moves stock market prices?

Dividend-Discount-Model (DDM)

$$E_{t-1}[P_t] = E_{t-1}[\sum_{k=1}^{\infty} \frac{D_{t+k}}{[1+R_{t+k}]^k}]$$

Schwert (1989)

▸ Back

Stock market data

- Stock market data: Kenneth R. French *http* : *//mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html*.
- It is the value-weighted return of all CRSP firms incorporated in the US and listed on the NYSE, AMEX, or NASDAQ that have a CRSP share code of 10 or 11 at the beginning of month t, good shares and price data at the beginning of t, and good return data for t.

Back

Textual Analysis in finance - Surveys

- Das "Text and Context. Language Analytics in Finance" (2014)
- Kearney, Liu "Textual sentiment in finance. A survey of methods and models", International Review of Financial Analysis (2014)
- Tetlock "Information Transmission in Finance", Annual Review of Financial Economics (2014)
- Loughran, McDonald "Textual Analysis in Accounting and Finance. A Survey", Journal of Accounting Research (2016)

Textual Analysis in finance and economics - A selection

- Antweiler and Frank (2004): internet stock message boards help predict volatilities on a horizon of 15 minutes.
- Tetlock (2007): the number of negative words in the Abreast of the Market column of the WSJ predicts stock returns at the daily frequency.
- Da et al. (2014): an index constructed out of user's google internet searches can forecast asset prices at a daily horizon.
- Doms and Morin (2004): build a recession word-index and analyze how the media affects consumers sentiment regarding the economy.
- Lawrence et al. (2017): determine the relevance of academic papers in economics and finance with textual analysis.
- Baker et al. (2016): create indexes of policy-related economic uncertainty based on newspaper coverage frequency.
- Manela and Moreira (2017) :construct a text-based measure of uncertainty.

▸ Back

Syuzhet dictionary - Matthew Jockers

Academic appointments

- 2018 - ... Professor. Department of English, University of Nebraska, Lincoln, NE.
- 2001 - 2012 Lecturer + Consulting Assistant Professor + Academic Technology Specialist, Department of English, Stanford University, Stanford, CA.

Industry and non-profit appointments

- 2017 - ... Co-founder, Archer Jockers, LLC.
- 2014 - 2015 Principal Research Scientist and Software Development Engineer. Apple Computer Inc. Cupertino, CA.

Books

- Archer, Jockers ”The Bestseller Code” (2016)
- Jockers ”Text Analysis with R for Students of Literature” (2014)
- Jockers ”Macroanalysis” (2013)

Source:
https://provost.wsu.edu/documents/2018/04/m-jockers-cv.pdf/

Back

Sentiment - Different dictionaries

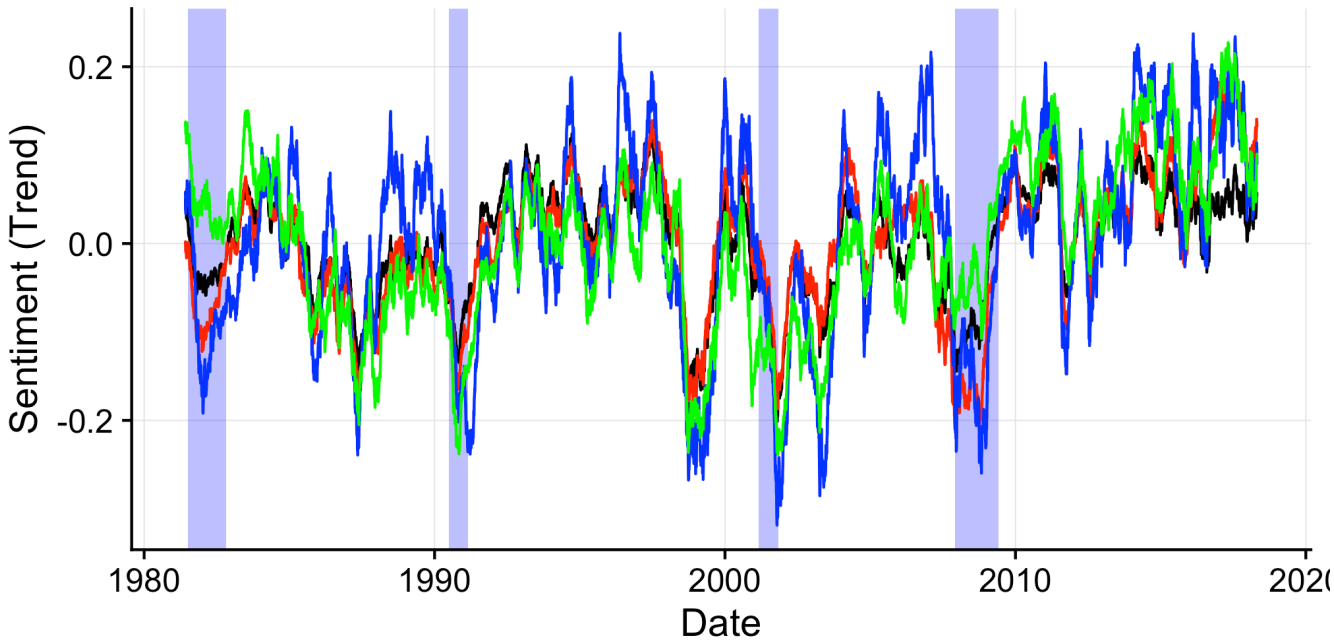


Figure: The figure shows the mean sentiment over all articles of one day (100 day moving average) for the four dictionaries. Black = Syuzhet, Red = Bing, Blue = Afinn, Green = NRC

Back

Sentiment - bing

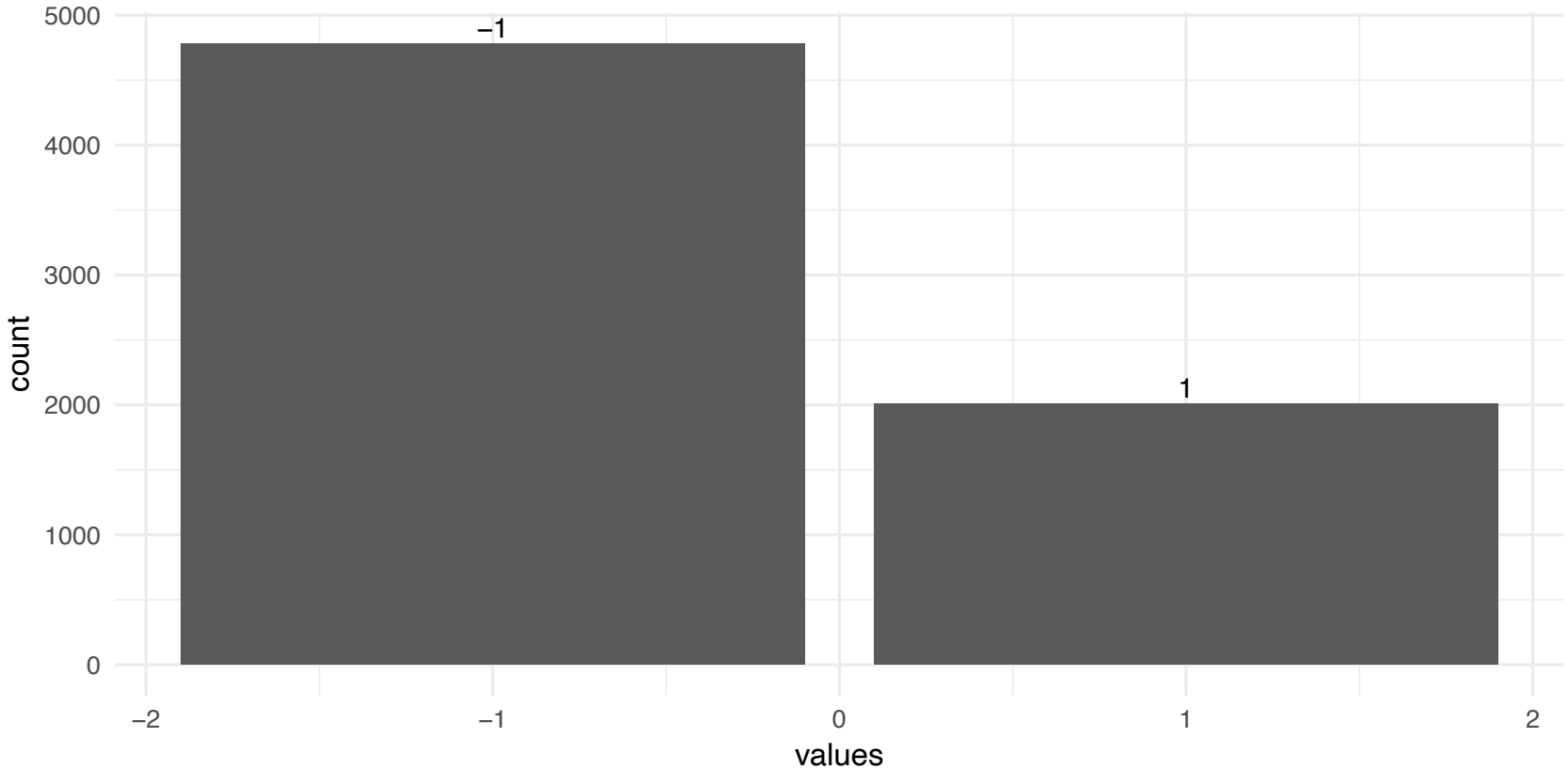


Figure: Histogram of the values of the words in the dictionary.

Total number of words: 6789
The ”bing” lexicon was develoepd by Minqing Hu and Bing Liu as the OPINION LEXICON.

Sentiment - afinn

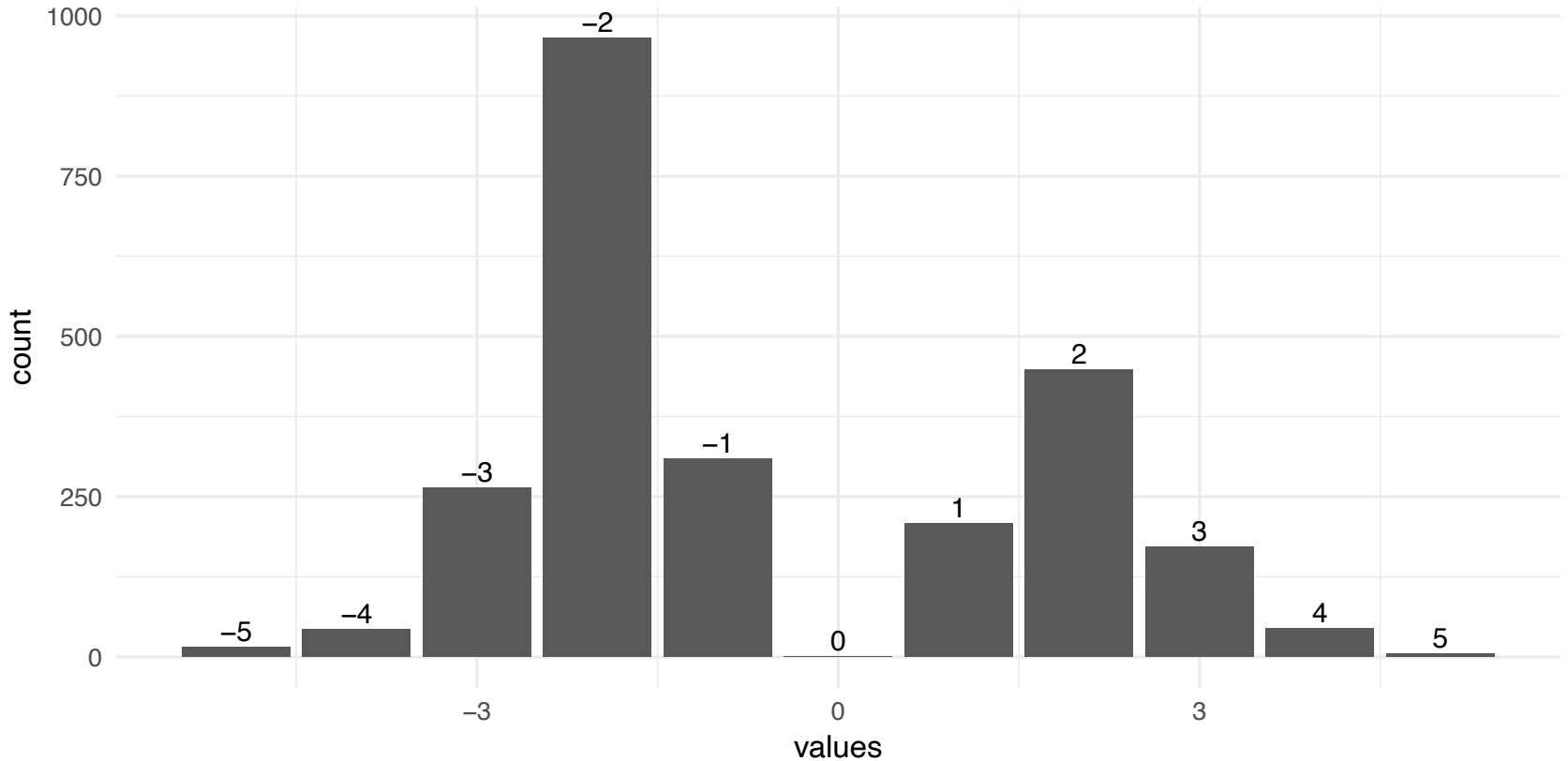


Figure: Histogram of the values of the words in the dictionary.

Total number of words: 2477
The ”afinn” lexicon was develoepd by Finn Arup Nielsen as the AFINN WORD DATABASE.

Sentiment - nrc

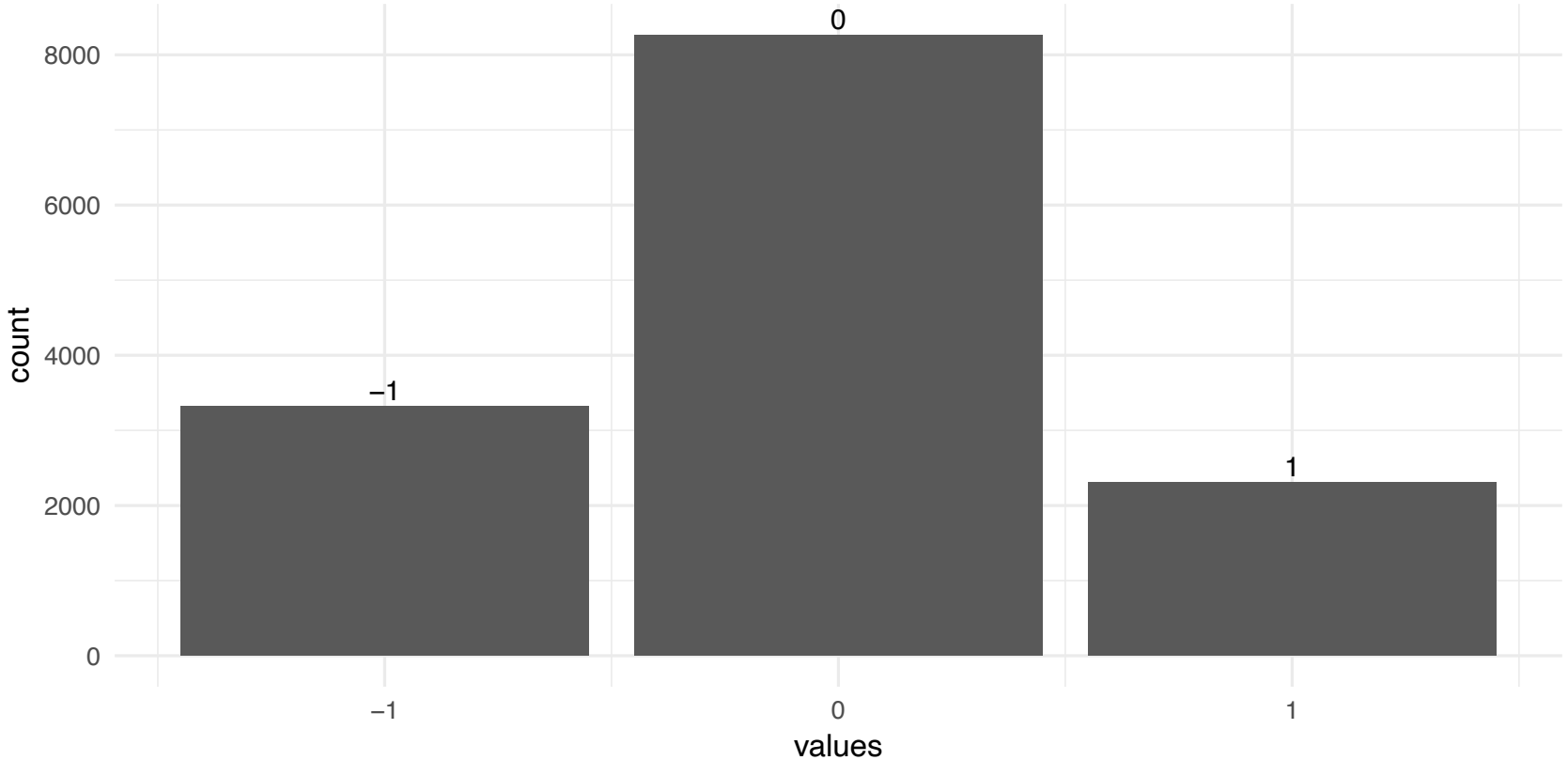


Figure: Histogram of the values of the words in the dictionary.

Total number of words: 5636
The ”nrc” lexicon was developed by Mohammad, Saif M. and Turney, Peter D. as the NRC EMOTION LEXICON.

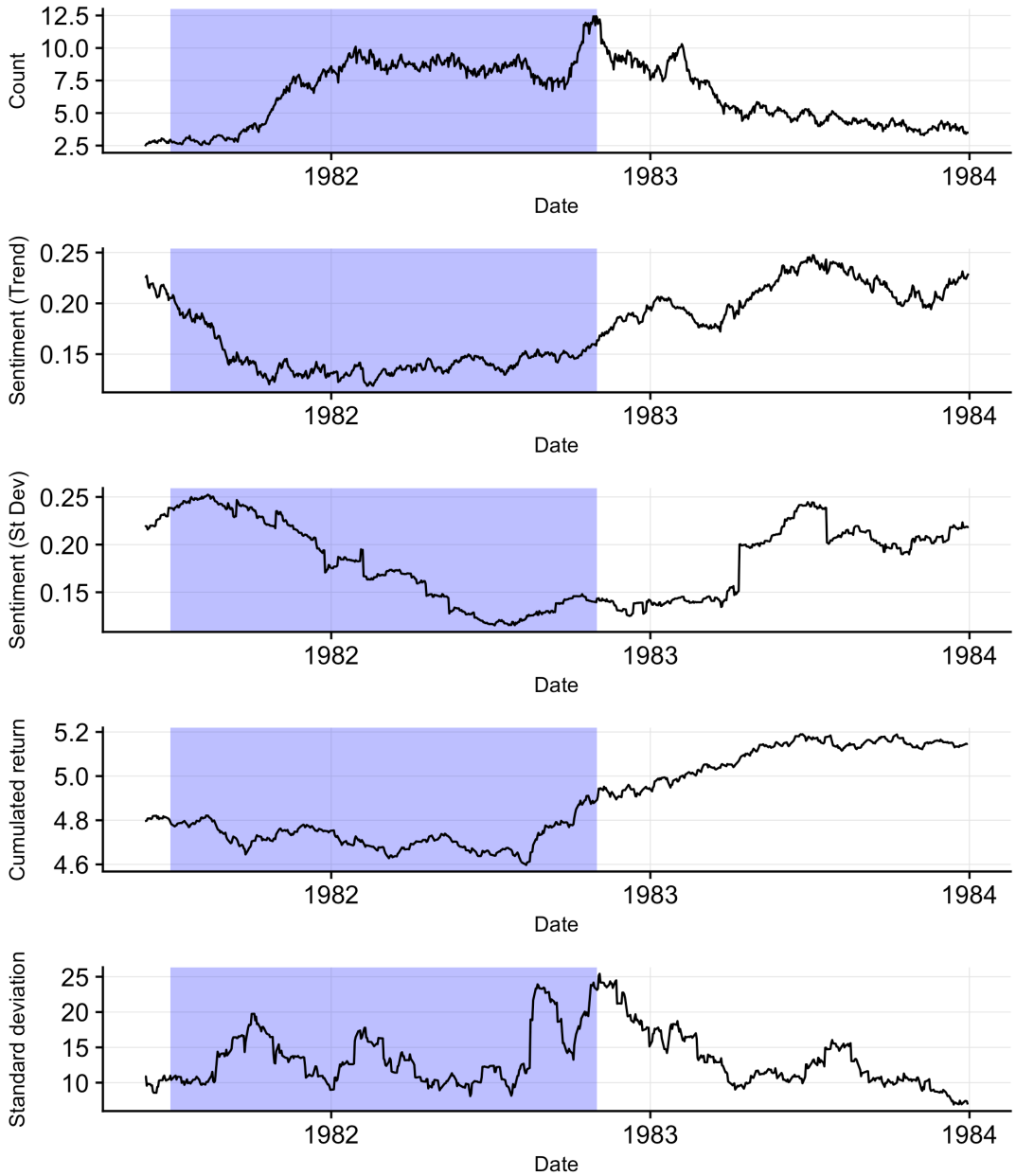
Sentiment - References

- "Syuzhet" lexicon was developed in the Nebraska Literary Lab under the direction of Matthew L. Jockers. The dictionary is created from a collection of 165,000 human coded terms taken from corpus of contemporany novels.
- Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.
- Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA.
- Finn Arup Nielsen. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs", Proceedings of the ESWC2011 Workshop on Making Sense of Microposts:Big things come in small packages 718 in CEUR Workshop Proceedings : 93-98. 2011 May.
- Saif Mohammad and Peter Turney. "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon." In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, June 2010, LA, California.

Back

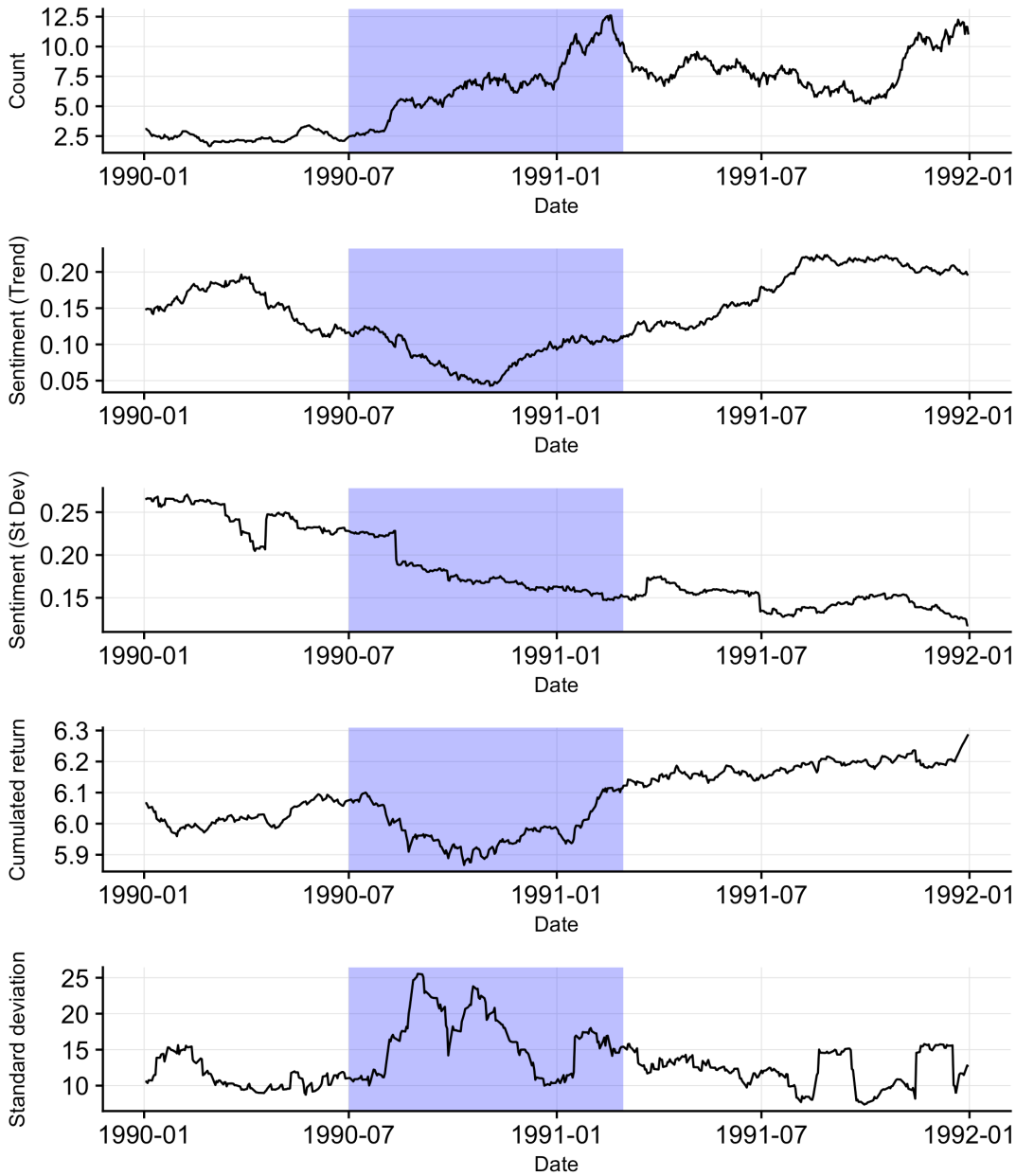
Recession 1980

Back



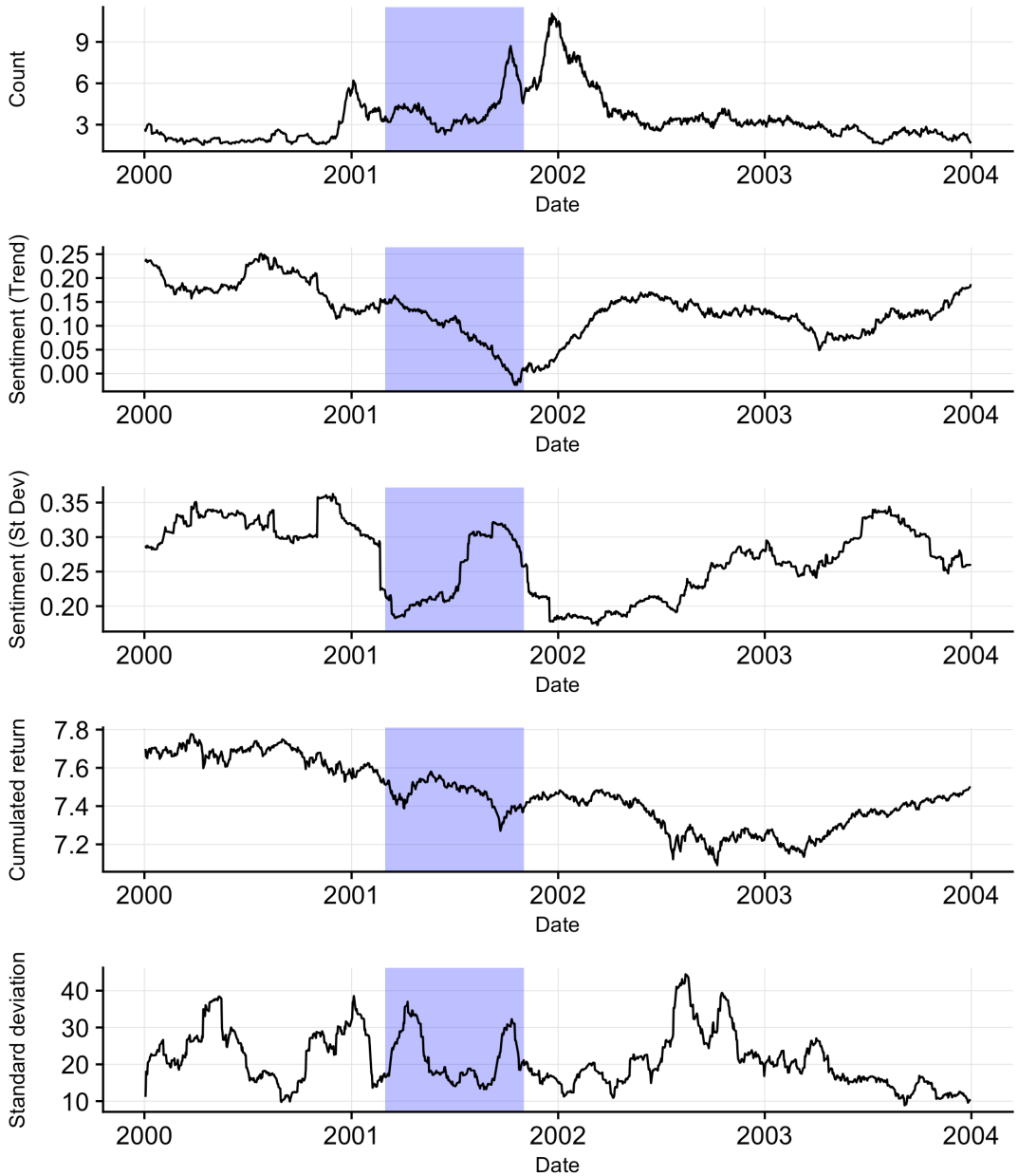
Recession 1990

Back



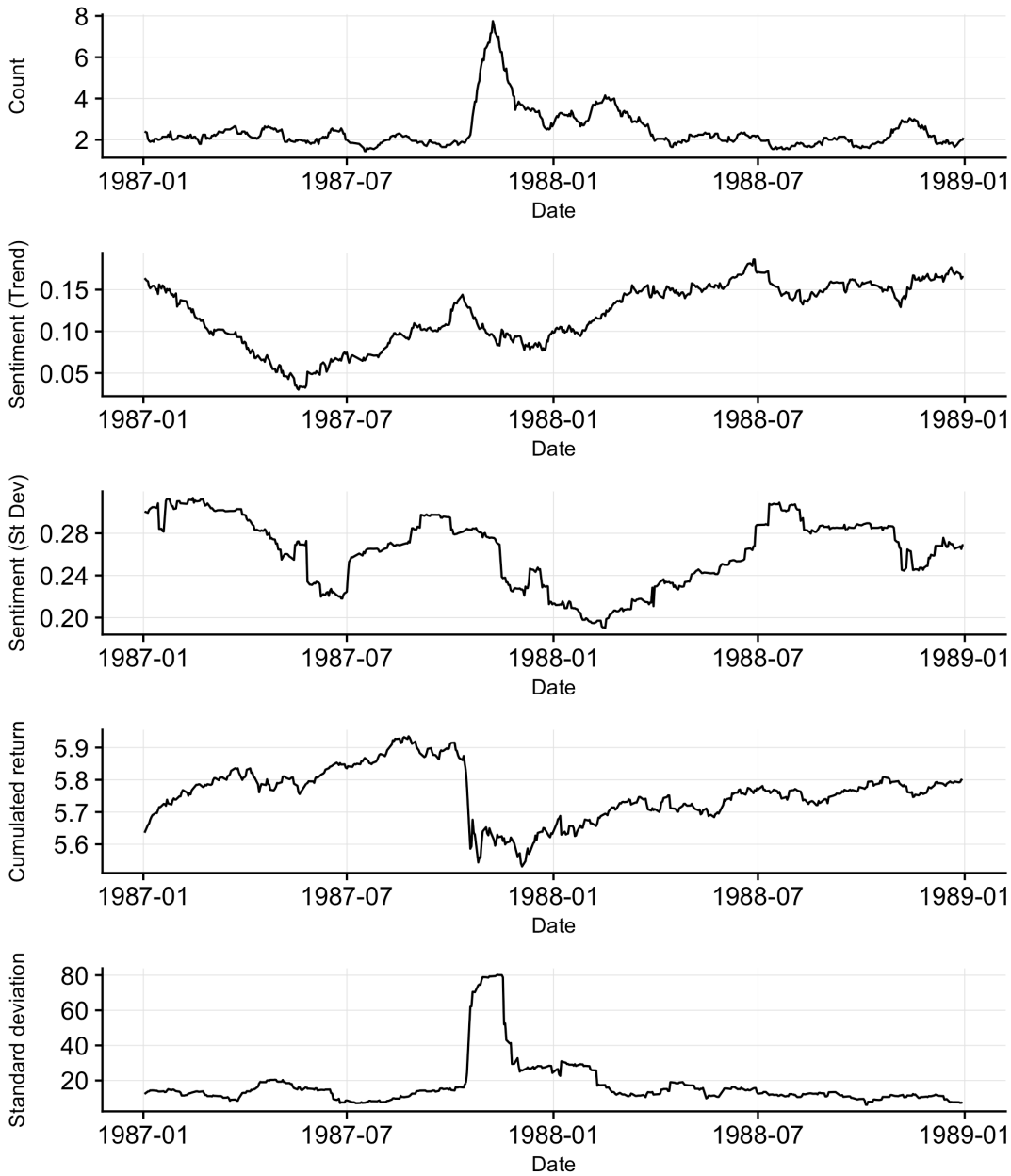
Recession 2000

Back



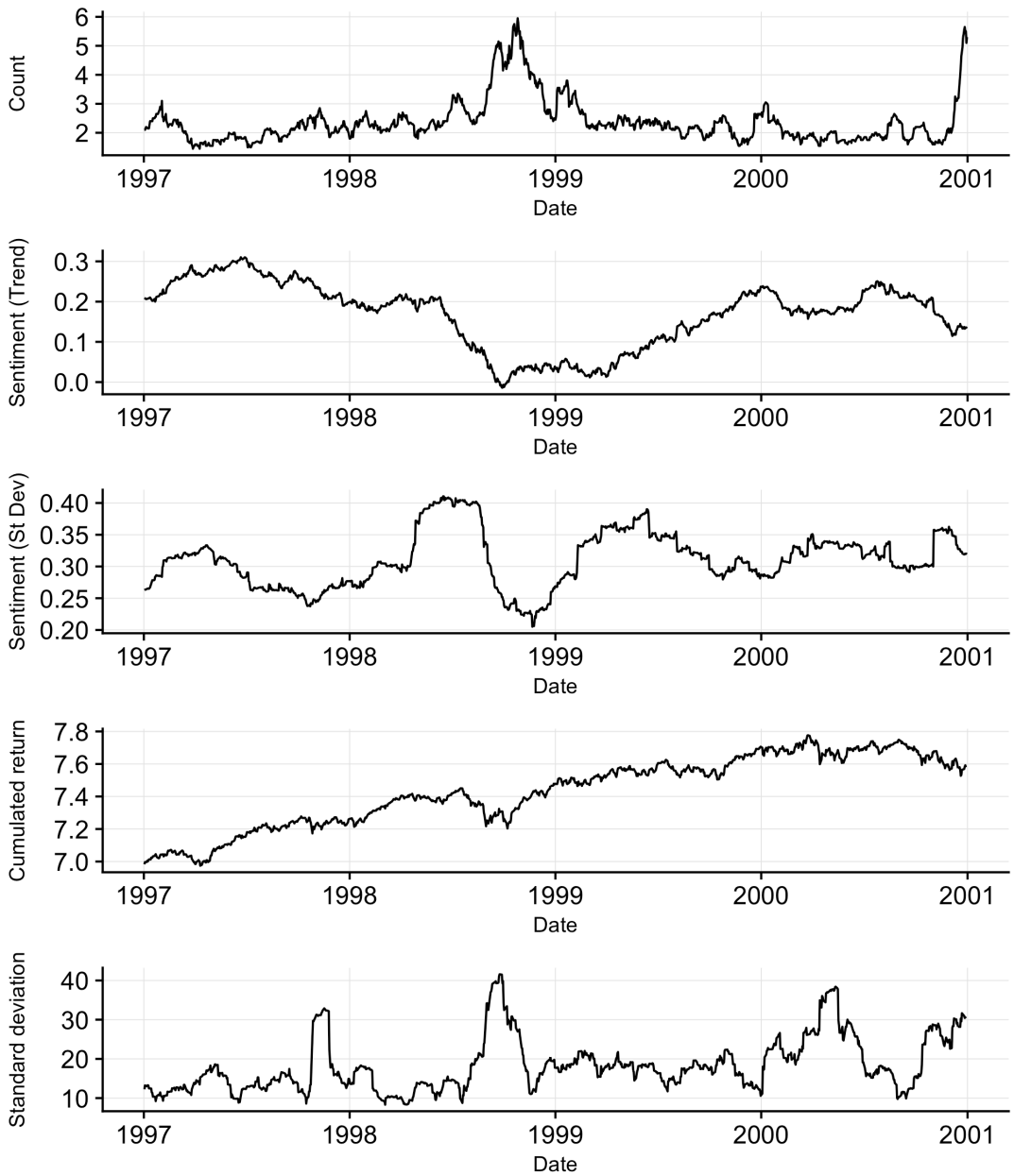
October 1987

Back



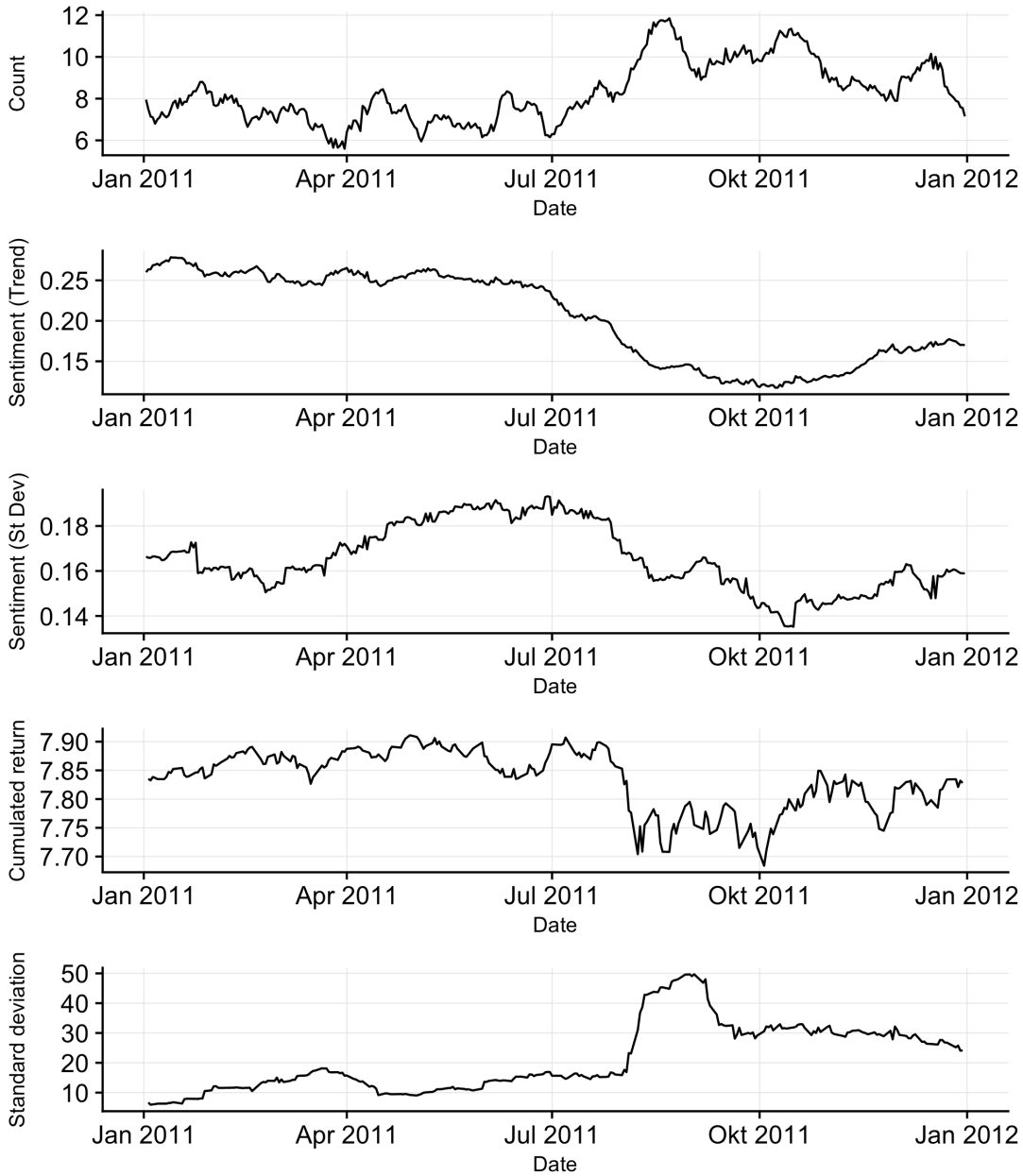
1990s

Back


Count
6
5
4
3
2
1997
1998
1999
2000
2001
Date
Sentiment (Trend)
0.3
0.2
0.1
0.0
1997
1998
1999
2000
2001
Date
Sentiment (St Dev)
0.40
0.35
0.30
0.25
0.20
1997
1998
1999
2000
2001
Date
Cumulated return
7.8
7.6
7.4
7.2
7.0
1997
1998
1999
2000
2001
Date
Standard deviation
40
30
20
10
1997
1998
1999
2000
2001
Date

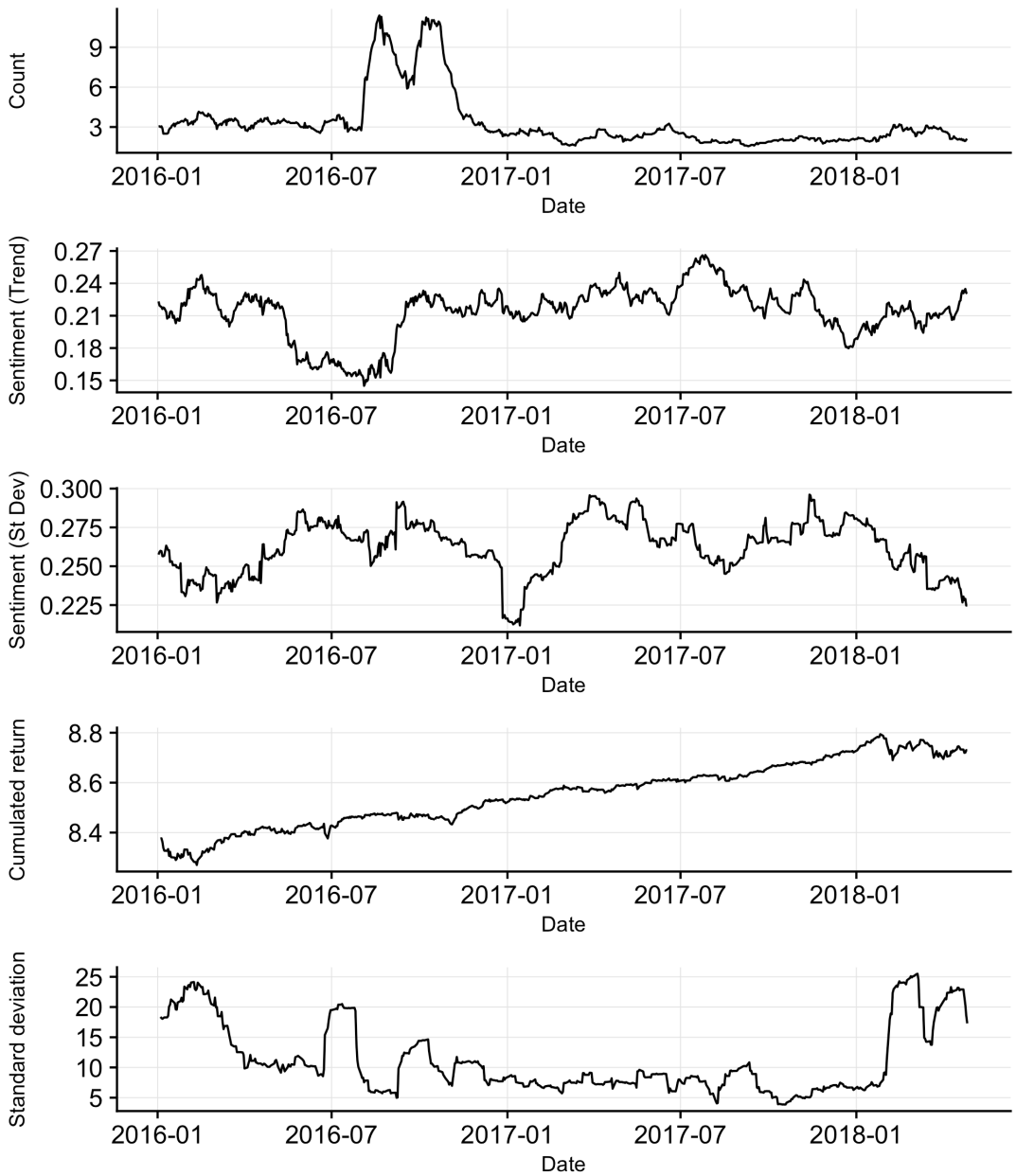
2011

Back


Count
12
10
8
6
Jan 2011
Apr 2011
Jul 2011
Okt 2011
Jan 2012
Date
Sentiment (Trend)
0.25
0.20
0.15
Jan 2011
Apr 2011
Jul 2011
Okt 2011
Jan 2012
Date
Sentiment (St Dev)
0.18
0.16
0.14
Jan 2011
Apr 2011
Jul 2011
Okt 2011
Jan 2012
Date
Cumulated return
7.90
7.85
7.80
7.75
7.70
Jan 2011
Apr 2011
Jul 2011
Okt 2011
Jan 2012
Date
Standard deviation
50
40
30
20
10
Jan 2011
Apr 2011
Jul 2011
Okt 2011
Jan 2012
Date

2016

Back


Count
9
6
3
2016-01
2016-07
2017-01
2017-07
2018-01
Date
Sentiment (Trend)
0.27
0.24
0.21
0.18
0.15
Sentiment (St Dev)
0.300
0.275
0.250
0.225
Cumulated return
8.8
8.6
8.4
Standard deviation
25
20
15
10
5

R packages I

- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
- **Check URL at NYT** / Hadley Wickham (NA). httr: Tools for Working with URLs and HTTP. R package version 1.3.1. https://github.com/r-lib/httr
- **Search data at NYT** / Scott Chamberlain (NA). rtimes: Client for New York Times 'APIs'. R package version 0.5.0.9151. https://github.com/ropengov/rtimes
- **Read HTML** / Hadley Wickham, James Hester and Jeroen Ooms (2018). xml2: Parse XML. R package version 1.2.0. https://CRAN.R-project.org/package=xml2
- **Parse HTML** / Duncan Temple Lang and the CRAN Team (2018). XML: Tools for Parsing and Generating XML Within R and S-Plus. R package version 3.98-1.11. https://CRAN.R-project.org/package=XML

R packages II

- **Get sentiment** / Jockers ML (2015). Syuzhet: Extract Sentiment and Plot Arcs from Text. https://github.com/mjockers/syuzhet.
- **Inference for estimated coefficients** / Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. https://CRAN.R-project.org/doc/Rnews/
- **Text mining** / Ingo Feinerer and Kurt Hornik (2017). tm: Text Mining Package. R package version 0.7-3. https://CRAN.R-project.org/package=tm
- **Topic modelling** / Grn B, Hornik K (2011). "topicmodels: An R Package for Fitting Topic Models. Journal of Statistical Software, 40(13), 1-30. doi: 10.18637/jss.v040.i13 (URL: http://doi.org/10.18637/jss.v040.i13).
- **Word cloud** / Ian Fellows (2014). wordcloud: Word Clouds. R package version 2.5. https://CRAN.R-project.org/package=wordcloud

R packages III

- **Word cloud** / Dawei Lang and Guan-tin Chien (2018). wordcloud2: Create Word Cloud by htmlwidget. R package version 0.2.1. https://CRAN.R-project.org/package=wordcloud2

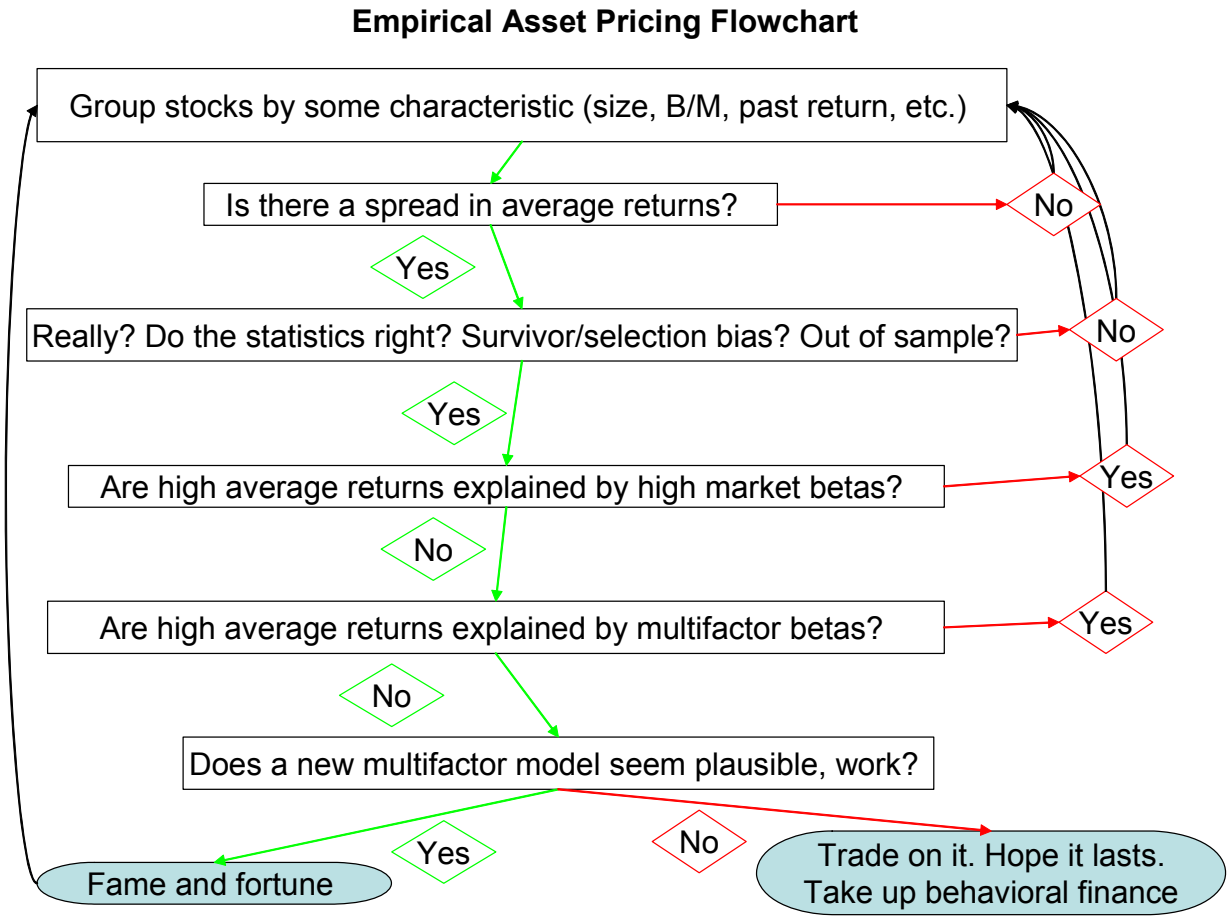
Back

APPENDIX CS

Appendix - Single securities

- Theory: Cochrane Old World ▸ Details and FMB ▸ Details
- Theory: Split ▸ Details
- Theory: One Tree ▸ Details
- Theory: RF ▸ Details and Model averaging ▸ Details
- Theory: Literature ▸ Details and People ▸ Details
- Theory: Algorithms ▸ Details
- Empirics: Data ▸ Details
- Empirics: Momentum Crash 2009 ▸ Details
- Empirics: Strategy returns against factor models ▸ Details
- Empirics: Variable Importance ▸ Details
- Empirics: Transaction costs ▸ Details
- Empirics: Different weighting ▸ Details
- Empirics: Value and Size Single ▸ Details
- Empirics: Linear regression with interactive terms ▸ Details
- Theory: Other Literature ▸ Details

The old world of thinking



John H. Cochrane, Lecture Notes, Winter 2013

Back

Fama-MacBeth approach

Fama-MacBeth approach as in Lewellen (2015):

Back

Fama-MacBeth approach

Fama-MacBeth approach as in Lewellen (2015):

1. In each cross-section, fit the regression

$$r_{i,t+1} = \beta_0^t + \sum_{g=1}^{24} \beta_g^t R_{it}(g,1) + \epsilon_{it}$$

Back

Fama-MacBeth approach

Fama-MacBeth approach as in Lewellen (2015):

1. In each cross-section, fit the regression

$$r_{i,t+1} = \beta_0^t + \sum_{g=1}^{24} \beta_g^t R_{it}(g,1) + \epsilon_{it}$$

2. Average coefficients over the past m months

$$\overline{\beta}_g^{t-1} = \frac{1}{m} \sum_{j=t-1-m}^{t-1} \hat{\beta}_g^t$$

Back

Fama-MacBeth approach

Fama-MacBeth approach as in Lewellen (2015):

1. In each cross-section, fit the regression

$$r_{i,t+1} = \beta_0^t + \sum_{g=1}^{24} \beta_g^t R_{it}(g,1) + \epsilon_{it}$$

2. Average coefficients over the past m months

$$\overline{\beta}_g^{t-1} = \frac{1}{m} \sum_{j=t-1-m}^{t-1} \hat{\beta}_g^t$$

3. Prediction is

$$\hat{r}_{i,t+1} = \overline{\beta}_0^{t-1} + \sum_{g=1}^{24} \overline{\beta}_g^{t-1} R_{it}(g,1),$$

▸ Back

Tree-based conditional portfolio sort

How are sorting variables and sorting values estimated?

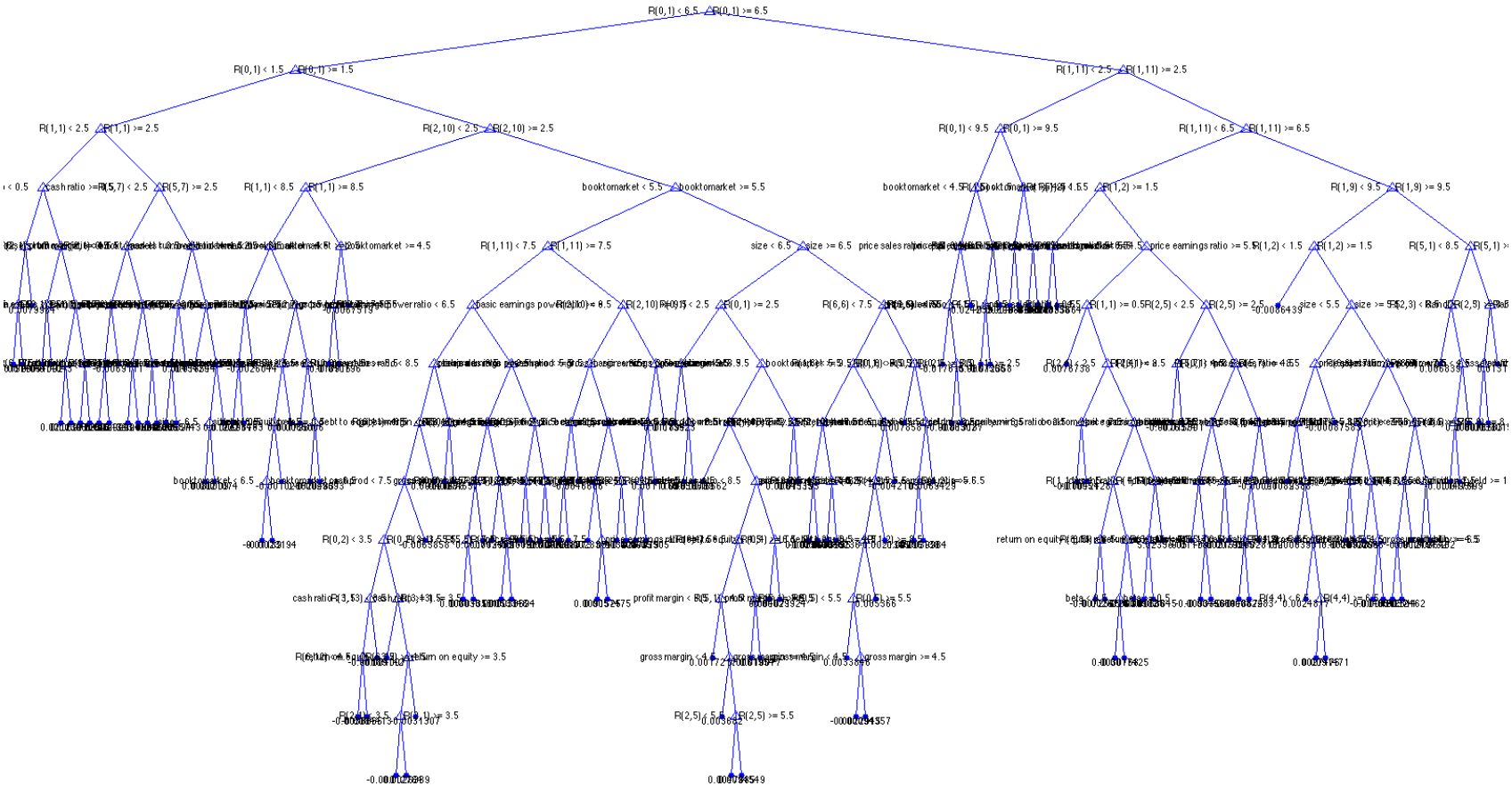
- Within each subset, find the variable and sorting value that minimize mean squared error by brute-force

$$\hat{\mu}_l = \text{Mean}(r_{i,t+1} | \text{Firm i } \in S_l \text{ in period t}). \tag{3}$$

$$(g^*, \tau^*) = \arg\min_{g,\tau} \left(\min_{\mu_1} \left(\sum_{R_{it}(g,1) \in S_1(R(g,1),\tau)} (r_{i,t+1} - \mu_1)^2 \right) + \min_{\mu_2} \left(\sum_{R_{it}(g,1) \in S_2(R(g,1),\tau)} (r_{i,t+1} - \mu_2)^2 \right) \right)$$

▸ Back

A Decision Tree on the whole data



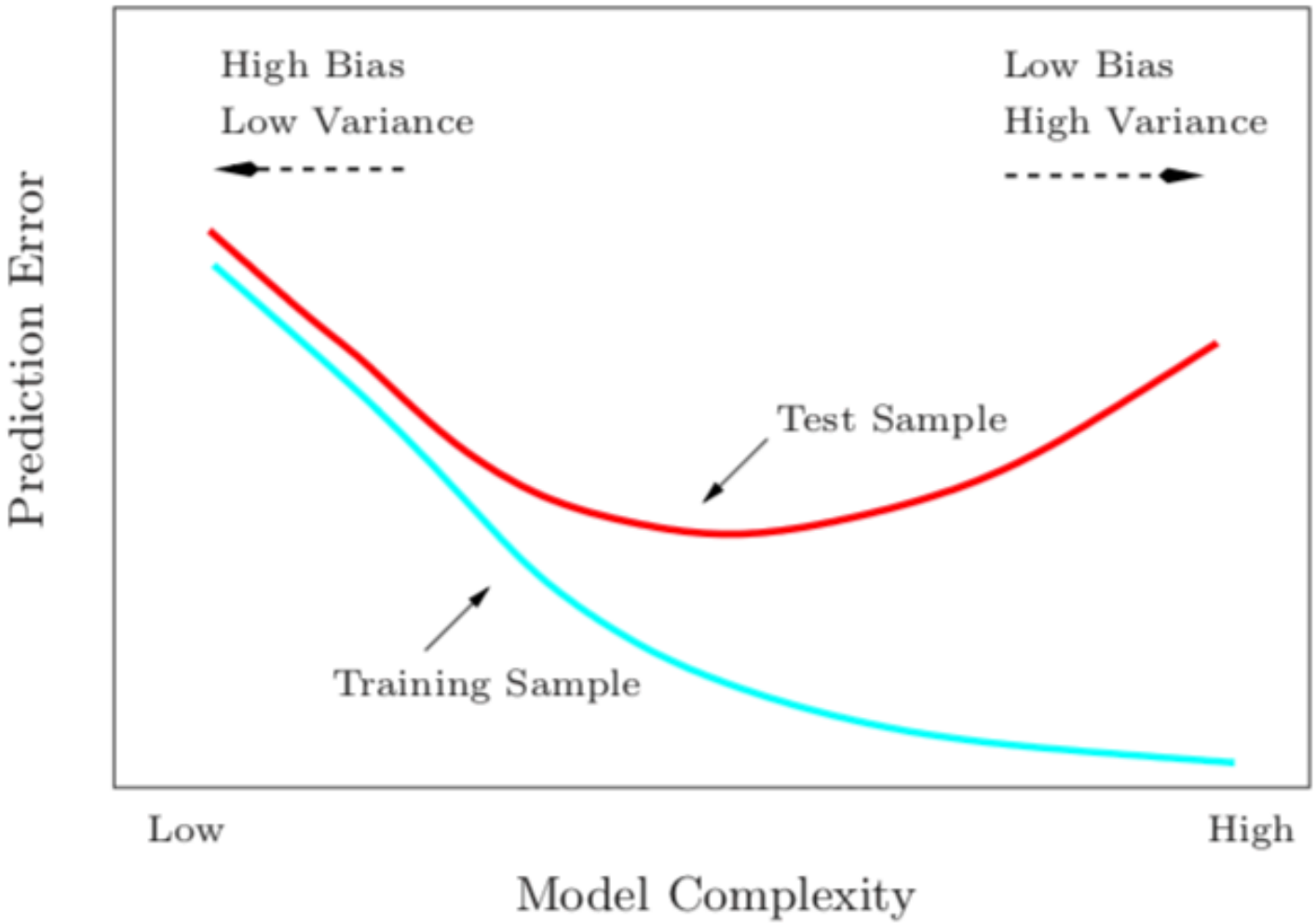
▶ Back

Random Forest - In a nutshell

Back

- The bias-variance tradeoff ($MSE = Variance + Bias^2$)
- The decision tree has a very high variance. It is instable w.r.t. changes in learning data. A solution: Random Forest
- Resampling involves repeatedly drawing samples from a training set and refitting a model of interest on each sample.
- Var(Mean) of uncorrelated variables:
  $Var(\overline{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}$
- Bootstrap = Sampling with replacement from one original data set. Each sample will contain approx. 63.2% of the original data. The other 36.8% are called ”out-of-bag” data.
  $\lim_{n\to\infty}(1-\frac{1}{n})^n = \frac{1}{e} \approx 0.368$
- Fitting on each bootstrap sample a tree and averaging the prediction is called bagging.
- Var(Mean) of correlated variables: $Var(\overline{X}) = \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2$
- Fitting on each bootstrap sample a tree where further the splitting variables in each split are randomly choosen (approximately 30% of all splitting variables), and averaging the prediction is called random forest.

Bias-Variance-Tradeoff



Source: Hastie, Tibshirani, Friedman "The Elements of Statistical Learning" (2008)

Back

Model averaging: Illustration

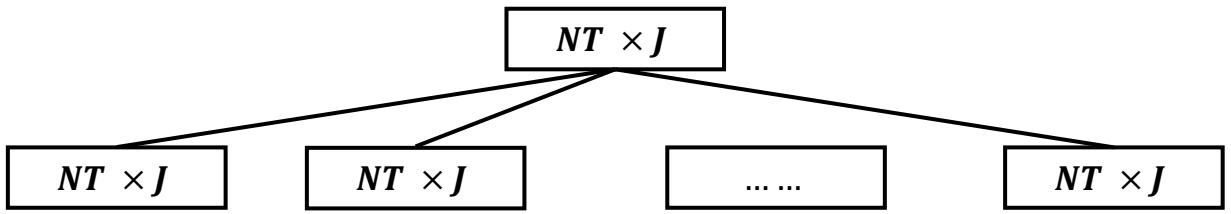
1. Estimation data

$$NT \times J$$

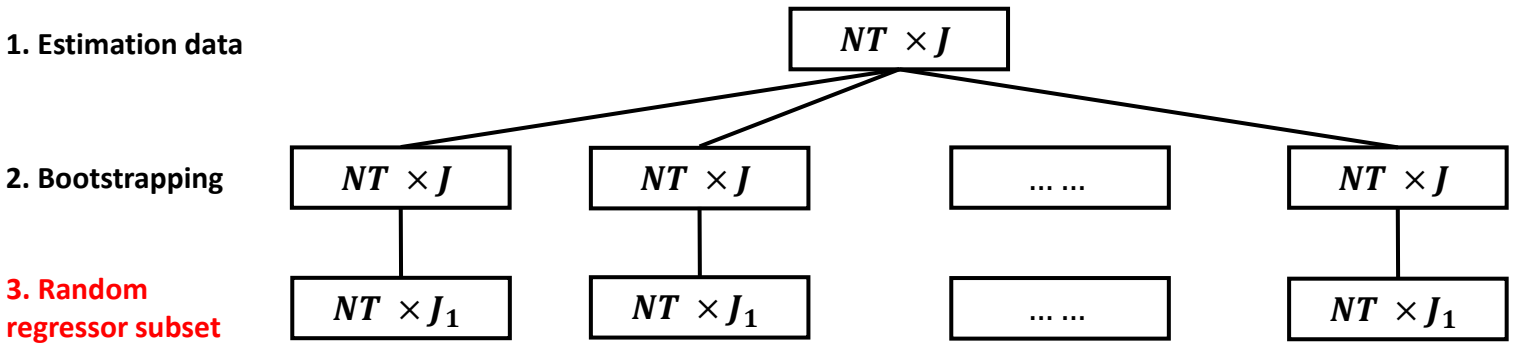
Model averaging: Illustration

1. Estimation data

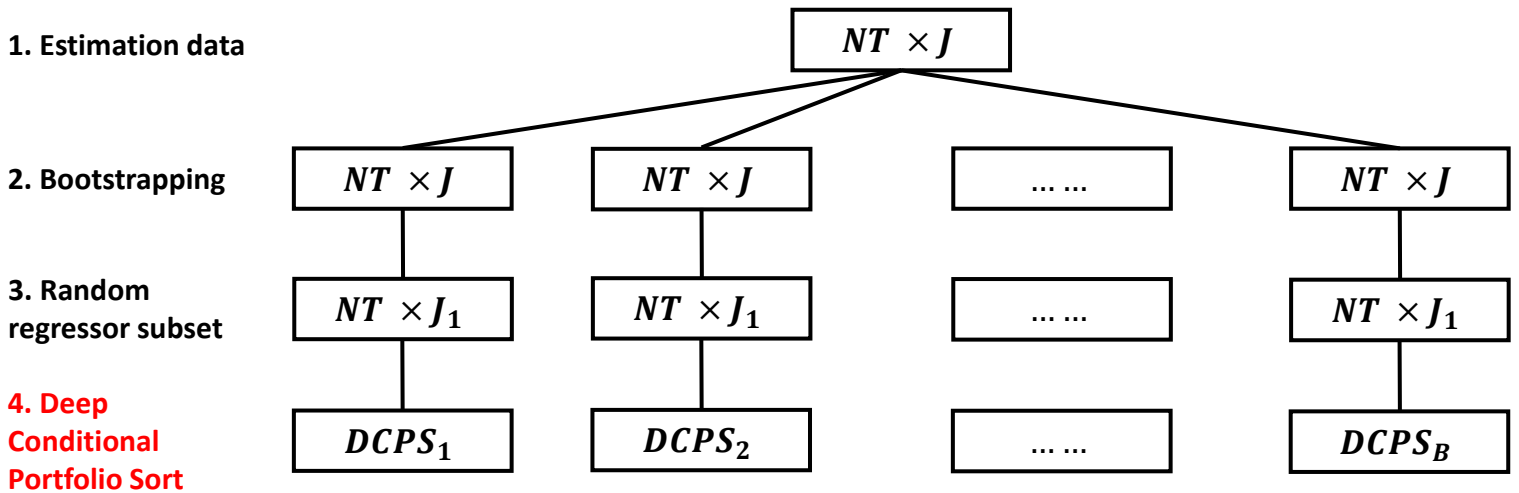
2. Bootstrapping



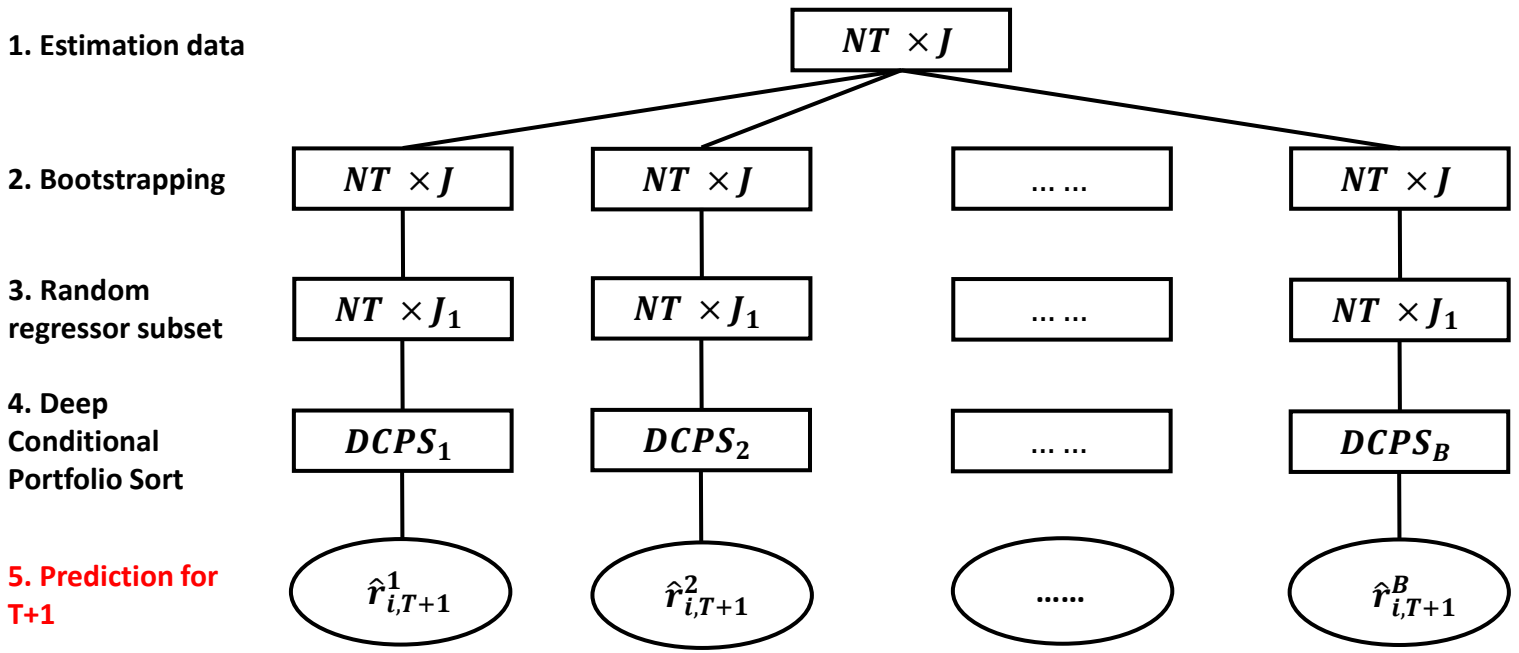
Model averaging: Illustration



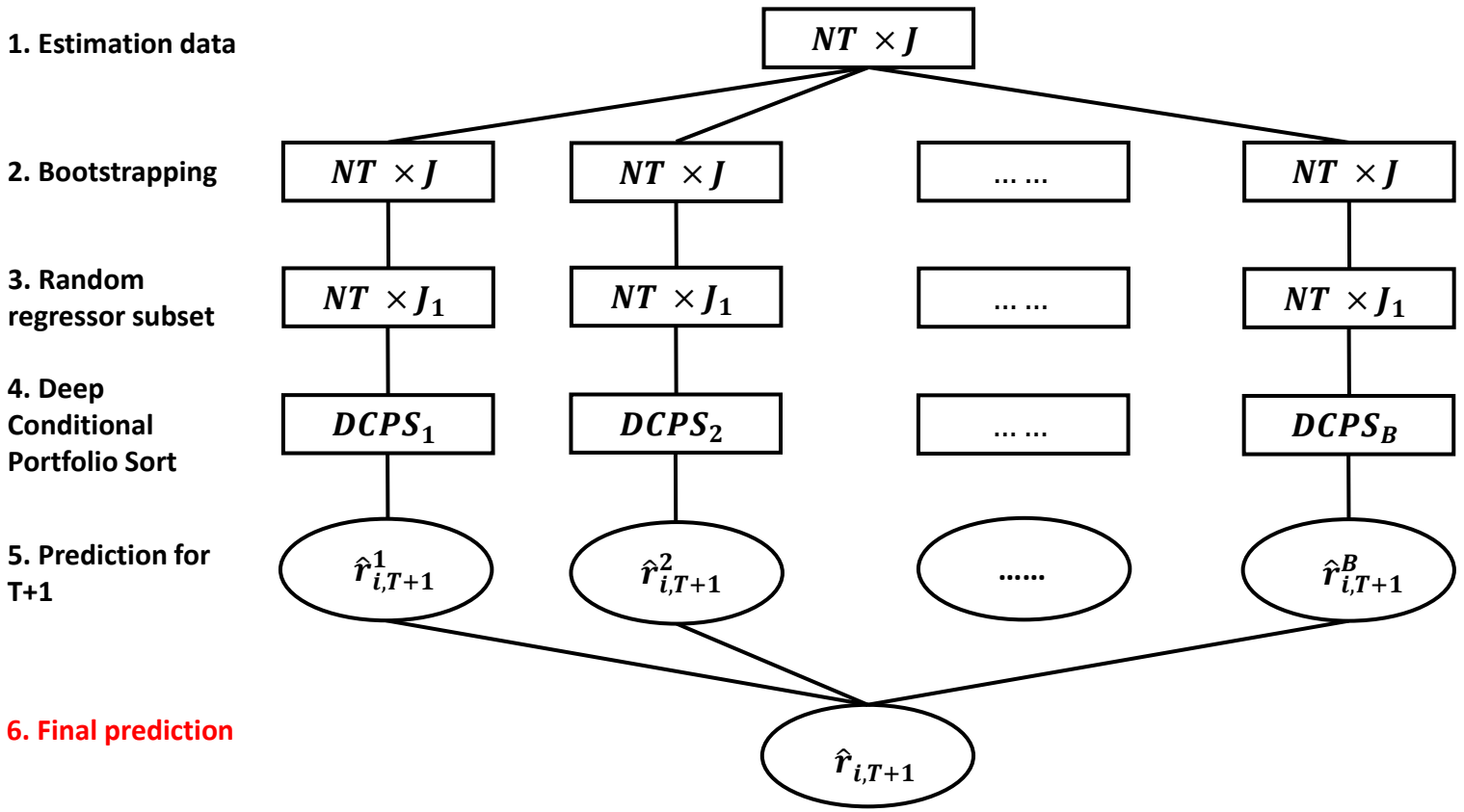
Model averaging: Illustration



Model averaging: Illustration



Model averaging: Illustration



Back

Literature

Key

- Breiman, Friedman, Olshen, Stone "Classification and Regression Trees" (1984)
- Breiman "Bagging Predictors", Machine Learning (1996)
- Breiman "Random Forests", Machine Learning (2001)

Further

- Breiman "Statistical Modeling. The Two Cultures", Statistical Science (2001)
- Hastie, Tibshirani, Friedman "The Elements of Statistical Learning" (2008) [1st ed.: 2001]
- James, Witten, Hastie, Tibshirani "An Introduction to Statistical Learning" (2013)

Back

Literature

Partial Dependence Plot

- Friedman "Greedy function approximation. a gradient boosting machine", Annals of Statistics (2001)

Seletion bias: the measure works particularly well when all regressors have the same scale and the same number of categories

- Strobl, Boulesteix, Zeileis, Hothorn "Bias in random forest variable importance measures. Illustrations, sources and a solution", BMC Bioinformatics (2007)

Correlated regressors

- Strobl, Boulesteix, Kneib, Augustin, Zeileis "Conditional variable importance for random forests", BMC Bioinformatics (2008)
- Gregorutti, Michel, Saint-Pierre "Correlation and variable importance in random forests", Statistics and Computing (2017)

▸ Back

Literature

More sophisticated tree-weighting schemes

- Biau "Analysis of a random forests model", Journal of Machine Learning Research (2012)
- Winham, Freimuth, Biernacka "A weighted random forests approach to improve predictive performance", Statistical Analysis and Data Mining (2013)

Adding irrelevant variables to an existing set of regressors does not change the importance measure of the variables in the existing set in large samples

- Ishwaran "Variable importance in binary regression trees and forests", Electronic Journal of Statistics (2007)
- Louppe, Wehenkel, Sutera, Geurts "Understanding variable importances in forests of randomized trees", In: Advances in Neural Information Processing Systems 26 (2013)

▸ Back

People

- **Leo Breiman (1928-2005, University of California, Berkeley)**
  https://www.stat.berkeley.edu/ breiman/
- **Jerome H. Friedman (1939, Stanford University)**
  http://statweb.stanford.edu/ jhf/
- **Trevor Hastie (1953, Stanford University)**
  http://web.stanford.edu/ hastie/
- **Richard A. Olshen (1942, Stanford University)**
  http://statweb.stanford.edu/ olshen/
- **Charles J. Stone (?, University of California, Berkeley)**
  https://statistics.berkeley.edu/people/chuck-stone
- **Robert Tibshirani (1956, Stanford University)**
  https://statweb.stanford.edu/ tibs/

Back

Algorithms

- **1st gen.** AID (Morgan and Sonquist, 1963), THAID (Messenger and Mandell, 1972), CHAID (Kass, 1980)
- **2nd gen.** CART (Breiman et al., 1984) , RECPAM (Ciampi et al., 1988), Segal (1988, 1992), LeBlanc and Crowley (1992), Alexander and Grimshaw (1996), Zhang (1998), MVPART (Death, 2002), Su et al. (2004); ID3 (Quinlan, 1986), M5 (Quinlan, 1992), C4.5 (Quinlan, 1993); FACT (Loh and Vanichsetakul, 1988)
- **3rd gen.** QUEST (Loh and Shih, 1997), CRUISE (Kim and Loh, 2001, 2003), Bayesian CART (Chipman et al., 1998; Denison et al., 1998)
- **4th gen.** GUIDE (Loh, 2002, 2009; Loh and Zheng, 2013; Loh et al., 2015), CTREE (Hothorn et al., 2006), MOB (Zeileis et al., 2008); Random forest (Breiman, 2001), TARGET (Fan and Gray, 2005; Gray and Fan, 2008), BART (Chipman et al., 2010)

Source: Loh (2014) and
http : *//washstat.org/presentations*/20150604/*loh_slides.pdf*

Algorithms

TABLE 1 | Comparison of Classification Tree Methods. A Check Mark Indicates Presence of a Feature

Feature	C4.5	CART	CHAID	CRUISE	GUIDE	QUEST
Unbiased Splits				✓	✓	✓
Split Type	*u*	*u,l*	*u*	*u,l*	*u,l*	*u,l*
Branches/Split	≥2	2	≥2	≥2	2	2
Interaction Tests				✓	✓	
Pruning	✓	✓		✓	✓	✓
User-specified Costs		✓	✓	✓	✓	✓
User-specified Priors		✓		✓	✓	✓
Variable Ranking		✓			✓	
Node Models	*c*	*c*	*c*	*c,d*	*c,k,n*	*c*
Bagging & Ensembles					✓	
Missing Values	*w*	*s*	*b*	*i,s*	*m*	*i*

b, missing value branch; *c*, constant model; *d*, discriminant model; *i*, missing value imputation; *k*, kernel density model; *l*, linear splits; *m*, missing value category; *n*, nearest neighbor model; *u*, univariate splits; *s*, surrogate splits; *w*, probability weights

Source: *https : //www.stat.wisc.edu/ loh/treeprogs/guide/wires*11.*pdf*

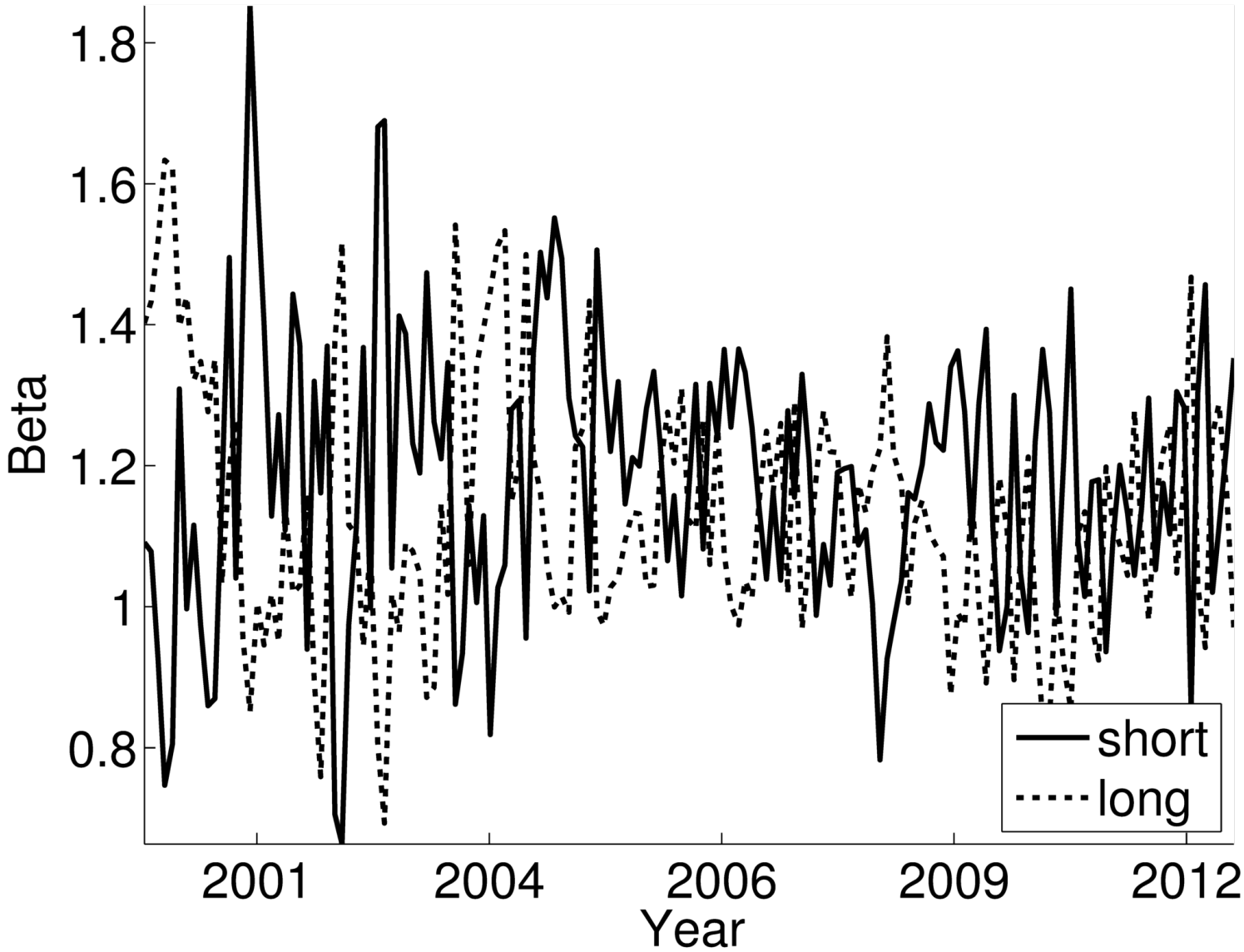
▸ Back

Data

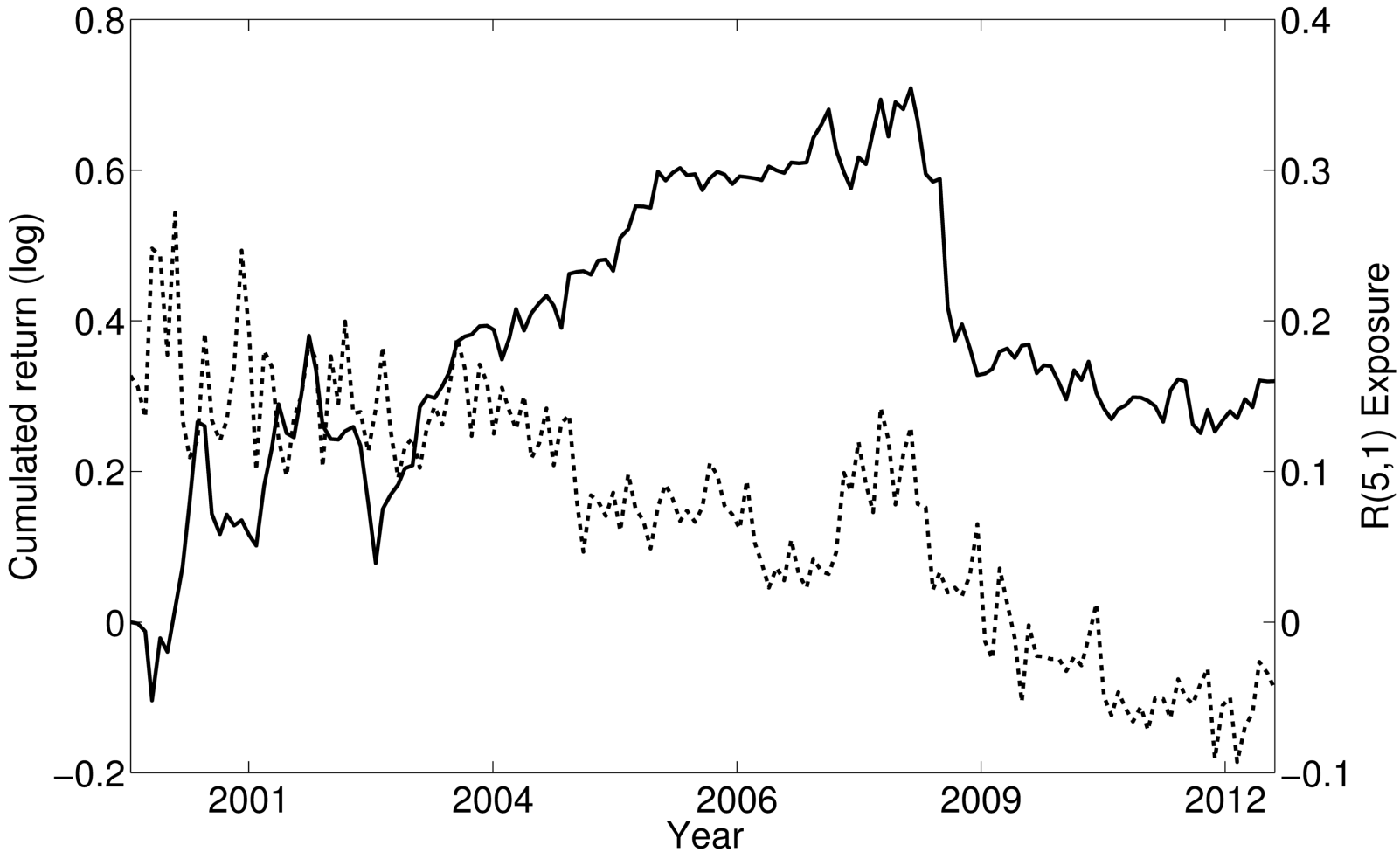
- CRSP monthly stock file
- Exclude penny stocks
- Common shares only
- Truncate .5% and 99.5% monthly return percentile
- Financial statement data (Compustat) for robustness checks
- Matched sample period 1963-2012
- FF factors and UMD

Back

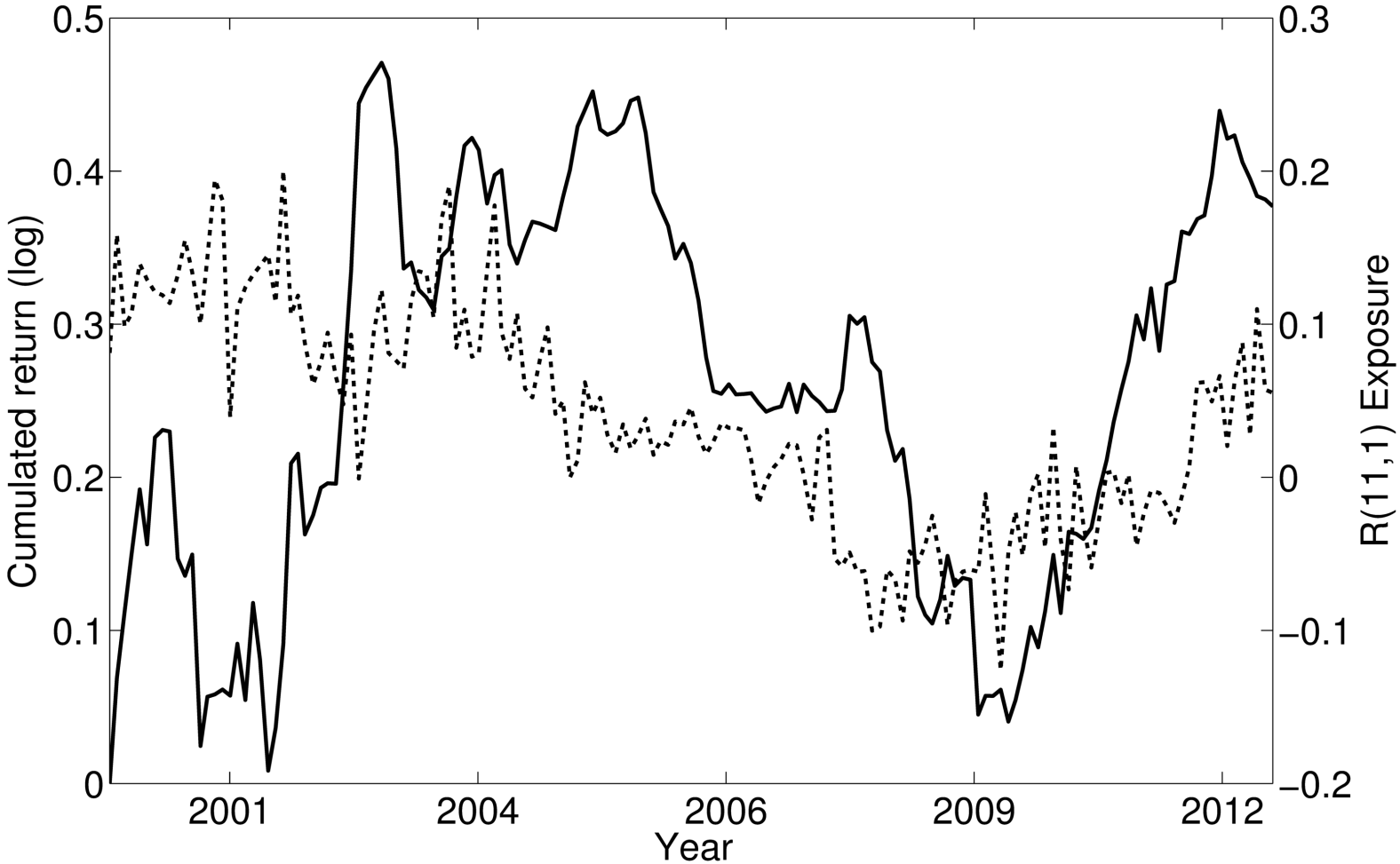
Momentum Crash 2009: Market beta



Momentum Crash 2009: Hedge return and exposure



Momentum Crash 2009: Hedge return and exposure



Back

Strategy returns against factor models

- Important to evaluate against equilibrium model of returns

$$\begin{aligned}\text{Strategy return}_t = \alpha &+ \beta^{MKT}\text{MKT}_t && \text{CAPM}\\ &+ \beta^{SMB}\text{SMB}_t + \beta^{HML}\text{HML}_t && \text{3F model}\\ &+ \beta^{UMD}\text{UMD}_t && \text{4F model}\\ &+ u_t\end{aligned}$$

Results: Tree-based conditional portfolio sort

Average return of 2.3 percent per month

	(1)	(2)	(3)	(4)
Intercept	2.30	2.23	2.25	2.05
	(16.75)	(16.04)	(16.51)	(14.54)
MKT		0.07	0.05	0.09
		(2.14)	(1.53)	(2.78)
SMB			0.08	0.09
			(1.40)	(1.69)
HML			-0.03	0.04
			(-0.39)	(0.61)
UMD				0.20
				(5.57)
R^2		0.02	0.03	0.13
IR		2.90	2.93	2.82
SR	2.96			
N	540	540	540	540

t-statistics in parentheses.

Decile factor loadings

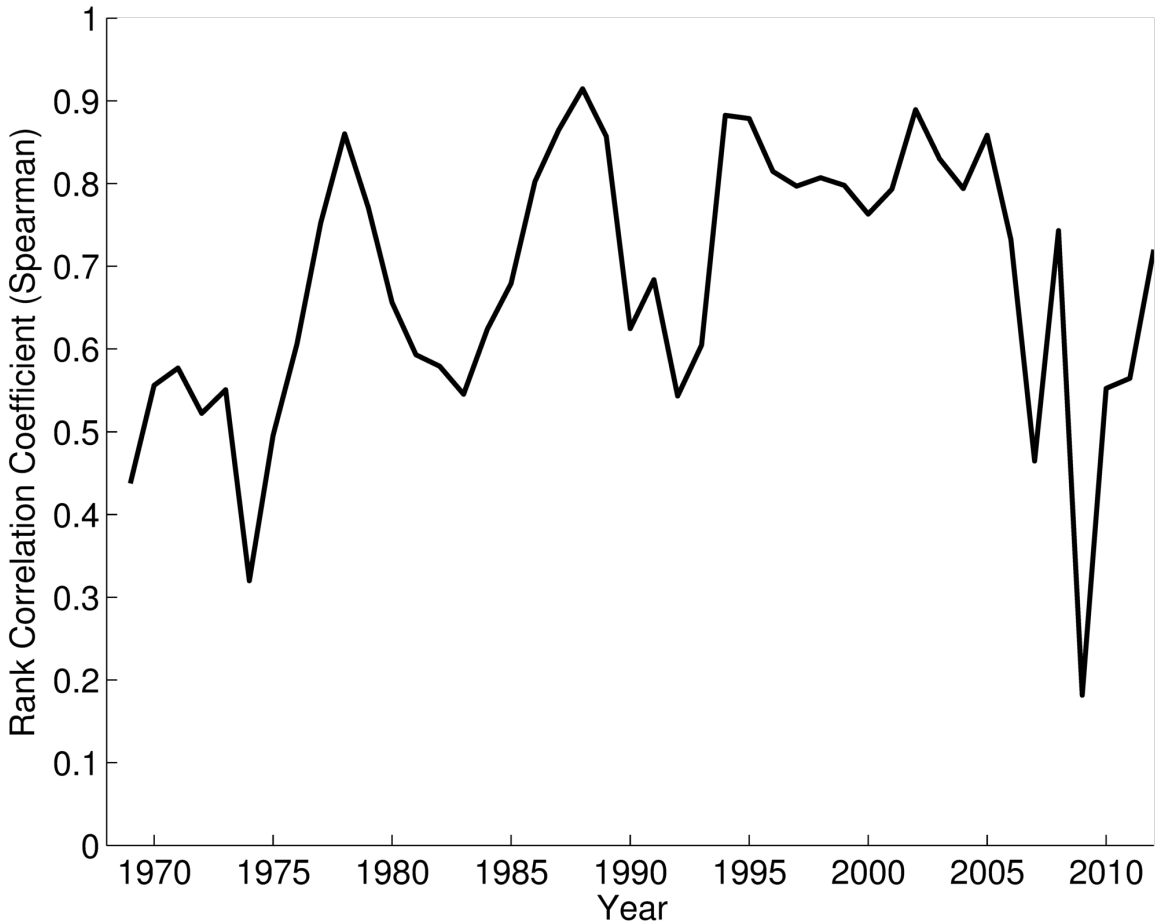
	Low	2	3	4	5	6	7	8	9	High	High-Low
Average return	-0.53	0.21	0.44	0.58	0.68	0.80	0.94	1.01	1.22	1.76	2.30
	(-1.74)	(0.77)	(1.66)	(2.29)	(2.61)	(3.10)	(3.52)	(3.78)	(4.27)	(5.54)	(16.75)
						Four factor model					
Intercept	-1.37	-0.67	-0.48	-0.36	-0.29	-0.21	-0.07	-0.02	0.13	0.69	2.05
	(-12.77)	(-7.10)	(-6.32)	(-4.96)	(-4.11)	(-3.10)	(-0.90)	(-0.25)	(1.40)	(5.13)	(14.54)
MKT	0.94	0.94	0.94	0.93	0.95	0.97	0.96	0.98	1.02	1.03	0.09
	(30.12)	(31.11)	(37.89)	(39.00)	(41.63)	(44.66)	(36.30)	(31.58)	(35.46)	(27.00)	(2.78)
SMB	0.86	0.73	0.69	0.67	0.65	0.65	0.69	0.71	0.80	0.95	0.09
	(11.69)	(11.04)	(10.88)	(10.14)	(10.15)	(9.58)	(9.15)	(9.38)	(10.52)	(13.16)	(1.69)
HML	0.15	0.19	0.23	0.26	0.25	0.26	0.25	0.23	0.24	0.18	0.04
	(2.47)	(3.39)	(4.40)	(4.91)	(4.96)	(4.65)	(4.53)	(3.62)	(3.86)	(2.34)	(0.61)
UMD	-0.27	-0.20	-0.15	-0.12	-0.10	-0.07	-0.06	-0.05	-0.03	-0.08	0.20
	(-7.71)	(-6.29)	(-4.87)	(-3.82)	(-2.90)	(-2.26)	(-1.86)	(-1.49)	(-0.93)	(-2.34)	(5.57)

t-statistics in parentheses.

▶ Back

Variable importance: Spearman rank correlation

Median rank correlation: .7



Variable importance

Including 86 firm fundamentals

- Construct firm fundamentals from Green, Hand and Zhang (2014)
- Popular characteristics that can be constructed with only Compustat, CRSP and IBES
- E.g. beta, size, value, earnings surprise, quality, volatility

Variable importance

Including 86 firm fundamentals

No Fundamentals	With Fundamentals
R(0,1)	R(0,1)
R(1,1)	R(1,1)
R(2,1)	R(2,1)
R(3,1)	R(3,1)
R(11,1)	R(4,1)
R(4,1)	R(5,1)
R(5,1)	R(11,1)
R(8,1)	R(7,1)
R(10,1)	R(6,1)
R(9,1)	R(8,1)

Variable importance

Including 86 firm fundamentals

No Fundamentals	With Fundamentals
R(0,1)	R(0,1)
R(1,1)	R(1,1)
R(2,1)	R(2,1)
R(3,1)	R(3,1)
R(11,1)	R(4,1)
R(4,1)	R(5,1)
R(5,1)	R(11,1)
R(8,1)	R(7,1)
R(10,1)	R(6,1)
R(9,1)	R(8,1)

Variable importance

By size category

- Size breakpoints as in Fama and French (2008)
- Based on NYSE stocks
  - Micro: Lowest 20%
  - Small: 20% to 50%
  - Large: upper 50%

Variable importance

By size category

All firms	Micro	Small	Large
R(0,1)	R(0,1)	R(0,1)	R(0,1)
R(1,1)	R(2,1)	R(2,1)	R(3,1)
R(2,1)	R(1,1)	R(1,1)	R(4,1)
R(3,1)	R(3,1)	R(3,1)	R(1,1)
R(11,1)	R(5,1)	R(4,1)	R(5,1)
R(4,1)	R(6,1)	R(5,1)	R(8,1)
R(5,1)	R(4,1)	R(6,1)	R(2,1)
R(8,1)	R(7,1)	R(7,1)	R(9,1)
R(10,1)	R(11,1)	R(8,1)	R(18,1)
R(9,1)	R(8,1)	R(11,1)	R(6,1)

Variable importance

By size category

All firms	Micro	Small	Large
R(0,1)	R(0,1)	R(0,1)	R(0,1)
R(1,1)	R(2,1)	R(2,1)	R(3,1)
R(2,1)	R(1,1)	R(1,1)	R(4,1)
R(3,1)	R(3,1)	R(3,1)	R(1,1)
R(11,1)	R(5,1)	R(4,1)	R(5,1)
R(4,1)	R(6,1)	R(5,1)	R(8,1)
R(5,1)	R(4,1)	R(6,1)	R(2,1)
R(8,1)	R(7,1)	R(7,1)	R(9,1)
R(10,1)	R(11,1)	R(8,1)	R(18,1)
R(9,1)	R(8,1)	R(11,1)	R(6,1)

Back

Transaction costs

- Past-return based strategies have high turnover (de Groot et al. (2012), Frazzini et al. (2015))
- Trading costs depend on investor type (Frazzini et al. (2015))
- We use the estimates from Frazzini et al to approximate the costs for an institutional investor

Transaction costs

Table: Turnover and trading costs

	Low	2	3	4	5	6	7	8	9	High	High-Low	9-2	8-3
Turnover (monthly)	1.56	1.74	1.76	1.78	1.78	1.8	1.78	1.76	1.76	1.62	3.18	3.5	3.52
Trading cost (annual)	3.71	4.09	4.13	4.17	4.17	4.21	4.17	4.13	4.13	3.83	7.13	7.81	7.85
Gross return (annual)	-6.18	2.55	5.41	7.19	8.47	10.03	11.88	12.82	15.66	23.29	31.37	12.82	7.06
Net return (annual)	-9.88	-1.54	1.28	3.02	4.30	5.82	7.71	8.69	11.53	19.46	24.24	5.01	-.79

▸ Back

Literature

- Kaminsky "Currency crises. Are they all the same", Journal of International Money and Finance (2006)
- Manasse, Roubini "Rules of thumb for sovereign debt crises", Journal of International Economics (2009)
- Duttagupta, Cashin "Anatomy of banking crises in developing and emerging market countries", Journal of International Money and Finance (2011)
- Ward "Spotting the danger zone. forecasting financial crises with classification tree ensembles and many predictors", Journal of Applied Econometrics (2017)
- Alessi, Detken "Identifying excessive credit growth and leverage", Journal of Financial Stability (2018)

Back

Robustness to tree weighting schemes

Back

Let MSE_b be the mean squared error of tree b and let MAE_b be the mean absolute error of tree b.

$$\hat{r}_{i,t+1} = \frac{1}{B} \sum_{b=1}^{B} w_b \hat{r}_{it}^b, \tag{4}$$

with

$$w_b = \frac{\frac{1}{MSE_b}}{\sum_{j=1}^{B} \frac{1}{MSE_j}} \tag{5}$$

and correspondingly for MAE.

	All relative to unweighted benchmark strategy			
Weighting	Annual return	Std Dev	Sharpe ratio	Max DD
MSE	0.25%	-0.15%	0.40%	-0.32%
MAE	0.10%	-0.08%	0.17%	0.00%

This table shows percentage deviations of strategy return moments when expected returns are based on weighted rather than unweighted trees (the benchmark case).

Tree-based sorts on single characteristics

Back

	Value				Size			
	Fama-French		Tree-based		Fama-French		Tree-based	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	0.39	0	0.83	0.63	0.18	0	1.41	1.15
	(2.58)	(0)	(5.53)	(4.37)	(1.31)	(0)	(6.46)	(5.58)
Mkt		0		0.11		0		0.15
		(0)		(3.75)		(0)		(3.42)
SMB		0		0.1		1		0.28
		(0)		(1.88)		(-)		(4.46)
HML		1		0.2		0		0.1
		(-)		(3.34)		(0)		(1.55)
UMD		0		0.01		0		0.03
		(0)		(0.17)		(0)		(0.68)
R^2		1		0.06		1		0.07
IR		0		0.75		0		0.89
SR		0		0.95		0		1.06
N	540	540	540	540	540	540	540	540

This table shows time-series regressions of strategy returns on factors. Returns are specified in percent per month. Columns labeled *Fama-French* have as dependent variable the standard value and size factors. Columns labeled *Tree-based* have as dependent variable strategy

Linear regression with interactive terms

Back

	Coefficients	t-stats
R(0, 1)	-2.40	-13.50
R(1, 1)	-0.03	-0.28
R(2, 1)	0.59	5.02
R(3, 1)	0.33	2.94
R(4, 1)	0.45	3.99
R(5, 1)	0.50	-4.99
R(6, 1)	0.27	2.68
(R(0, 1) $>$ 5) X R(1, 1)	0.39	4.98
(R(0, 1) $\leq$ 5) X R(1, 1)	-0.15	-1.93
(R(0, 1) $>$ 5) X R(2, 1)	0.30	4.78
(R(0, 1) $\leq$ 5) X R(2, 1)	-0.14	-2.08
(R(0, 1) $>$ 5) X R(3, 1)	0.22	3.06
(R(0, 1) $\leq$ 5) X R(3, 1)	-0.01	-0.20
(R(0, 1) $>$ 5) X R(4, 1)	-0.04	-0.60
(R(0, 1) $\leq$ 5) X R(4, 1)	0.03	0.51
(R(0, 1) $>$ 5) X R(5, 1)	0.04	0.58
(R(0, 1) $\leq$ 5) X R(5, 1)	0.21	2.98
(R(0, 1) $>$ 5) X R(6, 1)	0.09	1.34
(R(0, 1) $\leq$ 5) X R(6, 1)	0.19	2.69

This table shows coefficient estimates and t-statistics for the linear regression model. Past returns include return-based functions R(g,l) with length equal to

REFERENCES

References I

Antweiler, W. and M. Z. Frank (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance 59*(3), 1259–1294.

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics 131*(4), 1593–1636.

Bandarchuk, P. and J. Hilscher (2012). Sources of Momentum Profits: Evidence on the Irrelevance of Characteristics. *Review of Finance 17*(2), 809–845.

Bollerslev, T., R. F. Engle, and J. M. Wooldridge (1988). A capital asset pricing model with time-varying covariances. *Journal of political Economy 96*(1), 116–131.

Campbell, G., W. Quinn, J. D. Turner, and Q. Ye (2018). What moved share prices in the nineteenth-century london stock market? *The Economic History Review 71*(1), 157–189.

Campbell, J. Y. (1987). Stock returns and the term structure. *Journal of Financial Economics 18*(2), 373–399.

References II

Cornell, B. (2013). What moves stock prices: Another look. *Journal of Portfolio Management 39*(3), 32.

Cutler, D., J. Poterba, and L. Summers (1989). What moves stock prices? *Journal of Portfolio Management 15*(2).

Da, Z., J. Engelberg, and P. Gao (2014). The sum of all fears investor sentiment and asset prices. *The Review of Financial Studies 28*(1), 1–32.

de Groot, W., J. Huij, and W. Zhou (2012). Another look at trading costs and short-term reversal profits. *Journal of Banking & Finance 36*(2), 371–382.

Doms, M. E. and N. J. Morin (2004). Consumer sentiment, the economy, and the news media.

Fair, R. C. (2002). Events that shook the market. *The Journal of Business 75*(4), 713–731.

Fama, E. and K. French (1992). The cross-section of expected stock returns. *The Journal of Finance XLVII*(2), 427–467.

References III

Fama, E. and K. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(5), 1–22.

Fama, E. F. and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*(3), pp. 607–636.

Frazzini, A., R. Israel, and T. J. Moskowitz (2015). Trading costs of asset pricing anomalies. Unpublished manuscript.

French, K. R., G. W. Schwert, and R. F. Stambaugh (1987). Expected stock returns and volatility. *Journal of Financial Economics 19*(1), 3–29.

Ghysels, E., P. Guérin, and M. Marcellino (2014). Regime switches in the risk–return trade-off. *Journal of Empirical Finance 28*, 118–138.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2005). There is a risk-return trade-off after all. *Journal of Financial Economics 76*(3), 509–548.

References IV

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance 48*(5), 1779–1801.

Goyal, A. and S. Wahal (2015). Is momentum an echo? *Journal of Financial and Quantitative Analysis 50*(6), 1237–1267.

Harvey, C. R., Y. Liu, and H. Zhu (2016). ... and the cross-section of expected returns. *Review of Financial Studies 29*(1), 5–68.

Heston, S. L. and R. Sadka (2008). Seasonality in the cross-section of stock returns. *Journal of Financial Economics 87*(2), 418–445.

Hou, K., C. Xue, and L. Zhang (2017). Replicating anomalies. Technical report, National Bureau of Economic Research.

Ibbotson, R., R. J. Grabowski, J. P. Harrington, and C. Nunes (2016). *2016 Stocks, Bonds, Bills, and Inflation (SBBI) Yearbook*. John Wiley & Sons.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of Finance 45*(3), 881–898.

References V

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance 48*(1), 65–91.

Lawrence, S., S. Cates, C. Penado, and V. Samatova (2017). A machine learning approach to research curation for investment process. *Journal of Investment Management*.

Lehmann, B. (1990). Fads, martingales, and market efficiency. *The Quarterly Journal of Economics 105*(1), 1–28.

Lewellen, J. (2015). The cross-section of expected stock returns. *Critical Finance Review 4*, 1–44.

Loh, W.-Y. (2014). Fifty years of classification and regression trees. *International Statistical Review 82*(3), 329–348.

Lundblad, C. (2007). The risk return tradeoff in the long run: 1836–2003. *Journal of Financial Economics 85*(1), 123–150.

Manela, A. and A. Moreira (2017). News implied volatility and disaster concerns. *Journal of Financial Economics 123*(1), 137–162.

References VI

Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica*, 867–887.

Moreira, A. and T. Muir (2017). Volatility-managed portfolios. *The Journal of Finance 72*(4), 1611–1644.

Niederhoffer, V. (1971). The analysis of world events and stock prices. *The Journal of Business 44*(2), 193–219.

Novy-Marx, R. (2012). Is momentum really momentum? *Journal of Financial Economics 103*(3), 429–453.

Panetta, F., P. Angelini, R. Perli, M. Scatigna, S. Ramaswamy, S. Gerlach, A. Levy, P. Yesin, and G. Grande (2006). The recent behaviour of financial markets volatility. *BIS Papers*.

Roll, R. (1988). R2. *The Journal of Finance 43*(3), 541–566.

Schwert, G. W. (1989a). Business cycles, financial crises, and stock volatility. In *Carnegie-Rochester Conference Series on Public Policy*, Volume 31, pp. 83–125. Elsevier.

Schwert, G. W. (1989b). Why does stock market volatility change over time? *The Journal of Finance 44*(5), 1115–1153.

References VII

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance 62*(3), 1139–1168.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 817–838.

Whitelaw, R. F. (1994). Time variations and covariations in the expectation and volatility of stock market returns. *The Journal of Finance 49*(2), 515–541.