Distillation of News Flow into Analysis of Stock Reactions

Junni Zhang
Cathy Chen
Wolfgang Karl Härdle
Elisabeth Bommes

Guanghua School of Management
Peking University
Chung Hua University
Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics
and Economics
Humboldt–Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de

News moves Markets...

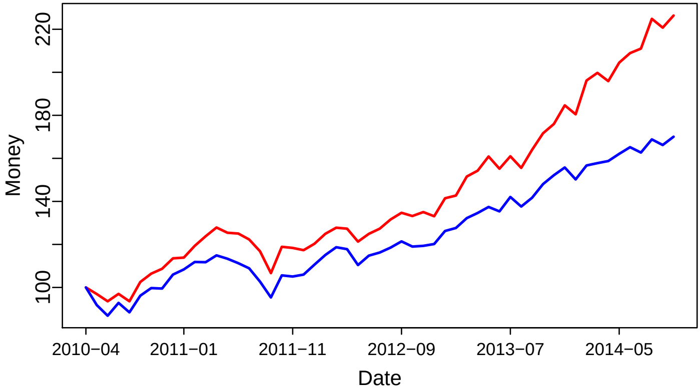

Figure 1: Investment in: S&P 500, Sentiment Strategy

... but there is a lot of News

Dimensions of News

- Source of news
  - Official channel: government, federal reserve bank/central bank, financial institutions
  - Internet: blog, social media, message board
- Content of news
  - Signal v.s. noise

Dimension of News ctd

- Type of news
  - Scheduled v.s. non-scheduled
  - Expected v.s. unexpected
  - Specific-event v.s. continuous news flows

Challenge

- Interpret news
- Evaluate news impact from different news dimensions

Sentiment Projection

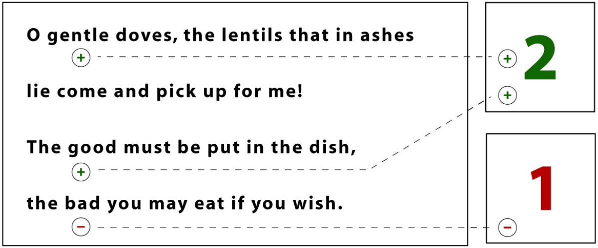

Figure 2: Example of Text Numerisization

- Many texts are numerisized via lexical projection
- Goal: Accurate values for positive and negative sentiment

Sentiment Lexica

- *Opinion Lexicon* (BL)
  Hu and Liu (2004)
- *Financial Sentiment Dictionary* (LM)
  Loughran and McDonald (2011)
- *Multi-Perspective Question Answering Subjectivity Lexicon* (MPQA)
  Wilson et al. (2005)

Research Questions

- Do opinions of small traders contribute to stock markets and create news-driven stock reactions?
  - Small traders v.s. financial institutions
  - Opinions of small traders v.s. financial analysts
- Concerns for analyst recommendation
  - Career
  - Compensation scheme
  - Strategic alliance

Research Questions ctd

- Are there differences regarding
  1. stock reaction indicators: volatility, trading volume, returns?
  2. degree of asymmetric response (leverage effect)?
  3. high and low attention companies?
  4. specific sectors?

Outline

1. Motivation ✓
2. Data Collection
3. Sentiment Projection
4. Panel Regression
5. Simulation
6. Conclusion

How to gather Sentiment Variables?

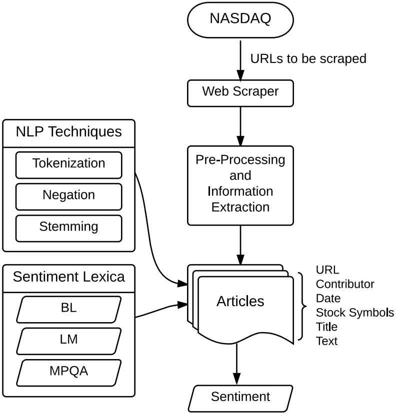

Figure 3: Flowchart of Data Gathering Process

NASDAQ Articles

- Terms of Service permit web scraping
- 116,691 articles in total
- 43,459 articles about 100 selected S&P 500 stocks in 9 major GICS sectors Frequency Table: GICS
- Time frame: October 2009 - October 2014
- Data available at RDC

Sentiment Lexica ctd

- Number of entries in each lexicon:

Lexicon	Positive	Negative
BL	2,006	4,783
LM	354	2,329
MPQA	2,718	4,911

- Some words appear only in one lexicon Unique Words
- Other words are only found in two lexica Shared Words

Sentiment Variables

- $I_{i,t}$ - article indicator
- $Pos_{i,t}$ - average proportion of positive words
- $Neg_{i,t}$ - average proportion of negative words

for stock i on day t

Comparison of Lexical Projections

- Average sentiment values are smaller for LM than for BL and MPQA
- *BL* and *MPQA* relatively similar
- *LM* only contains finance specific words
- *BL* and *MPQA* also contain more general words (e.g. "cancer")

Summary Statistics Correlation - Sentiment Tagging Example

- Combination of projections might improve results
  - PCA on sentiment scores
  - Use first principal component of $Pos_{i,t}$ and $Neg_{i,t}$

How good are the Projections?

- Random selection of 100 articles, manual labeling of polarity and comparison with polarity of lexical projections
- *BL* and *MPQA* recognize fewer negative articles but good in detection of positive articles
- *LM* accurately detects negative articles, recognizes fewer positive articles

Classification Evaluation Table

Stock Reaction Indicators

Range-based measure of volatility by Garman and Klass (1980)

- Notation: $\sigma_{i,t}$ Computation
- Based on open-high-low-close prices
- Equivalent results to realized volatility
- More robust in case of microstructure effects

Detrended log trading volume by Girard and Biswas (2007)

$$V_{i,t} = V^*_{i,t} - (\alpha + \beta_{1,i}\, t + \beta_{2,i}\, t^2) \tag{1}$$

with raw log trading volume $V^*_{i,t}$ and detrended log trading volume $V_{i,t}$ for stock i on day t, rolling window estimation (size: 120 days)

Returns

$$R_{i,t} = \log(P^C_{i,t}) - \log(P^C_{i,t-1}) \tag{2}$$

with $P^C_{i,t}$ as closing price of stock i on day t

Panel Regression

$$
\begin{aligned}
\sigma_{i,t+1} &= \alpha + \beta_1 I_{i,t} + \beta_2 Pos_{i,t} + \beta_3 Neg_{i,t} + \beta_4^\top X_{i,t} + \gamma_i + \varepsilon_{i,t} \qquad (3) \\
V_{i,t+1} &= \alpha + \beta_1 I_{i,t} + \beta_2 Pos_{i,t} + \beta_3 Neg_{i,t} + \beta_4^\top X_{i,t} + \gamma_i + \varepsilon_{i,t} \qquad (4) \\
R_{i,t+1} &= \alpha + \beta_1 I_{i,t} + \beta_2 Pos_{i,t} + \beta_3 Neg_{i,t} + \beta_4^\top X_{i,t} + \gamma_i + \varepsilon_{i,t} \qquad (5)
\end{aligned}
$$

for stock i on day t with separate estimation of (3) to (5).

$X_{i,t}$ - control variables
γ_i - company specific fixed effect satisfying $\sum_i \gamma_i = 0$

Control Variables

- $R_{M,t}$ - S&P 500 index return
- VIX_t - CBOE VIX
- $\sigma_{i,t}$ - Range-based volatility
- $V_{i,t}$ - Detrended trading volume
- $R_{i,t}$ - Return

Entire Panel Regression Results

Variable	BL	LM	MPQA	PCA
	Panel A: Future Volatility $\sigma_{i,t+1}$			
$I_{i,t}$	-0.000	-0.000	-0.000	-0.000
$Pos_{i,t}$	-0.002	-0.001	-0.001	-0.001
$Neg_{i,t}$	0.005^{*}	0.006^{**}	0.004	0.004^{**}
	Panel B: Future Detrended Log Trading Volume $V_{i,t+1}$			
$I_{i,t}$	0.047^{***}	0.032^{***}	0.050^{***}	0.049^{***}
$Pos_{i,t}$	-0.671^{***}	-0.233	-0.618^{***}	-0.470^{***}
$Neg_{i,t}$	0.888^{***}	0.768^{***}	0.907^{***}	0.589^{***}
	Panel C: Future Returns $R_{i,t+1}$			
$I_{i,t}$	-0.001^{**}	-0.000	-0.000	-0.001^{**}
$Pos_{i,t}$	0.021^{***}	0.016^{***}	0.016^{**}	0.015^{***}
$Neg_{i,t}$	-0.000	-0.006	-0.006	-0.003

*** p value < 0.01, ** $0.01 \leq p$ value < 0.05, * $0.05 \leq p$ value < 0.1

Does Attention matter?

- Number of days with articles differs between firms
- High attention: Faster incorporation of news?

$$\text{attention ratio} \stackrel{def}{=} N_i / T \tag{6}$$

with N_i as number of days with at least one article for company i and T as total number of trading days

Grouping

Use attention ratio quartiles to group firms:

Low		attention ratio $<$ Q1
Median	Q1 $\leq$	attention ratio $<$ Q2
High	Q2 $\leq$	attention ratio $<$ Q3
Extremely High	Q3 $\leq$	attention ratio

with Q1, Q2, Q3 as first, second and third quartile

Attention Analysis Regression Results

	BL		LM		MPQA	
Attention	Low	Extr. High	Low	Extr. High	Low	Extr. High
	Panel A: Future Volatility $\sigma_{i,t+1}$					
$I_{i,t}$	0.000	0.000	0.000	-0.000	0.000	0.000
$Pos_{i,t}$	-0.000	-0.001	-0.002	-0.002	-0.001	-0.001
$Neg_{i,t}$	0.001	0.005***	0.001	0.007***	0.001	0.004**
	Panel B: Future Detrended Log Trading Volume $V_{i,t+1}$					
$I_{i,t}$	0.072***	0.033***	0.048***	0.025**	0.067***	0.049***
$Pos_{i,t}$	-1.185***	-0.242	-1.077*	0.327	-0.815**	-0.623*
$Neg_{i,t}$	0.328	0.764**	0.200	0.709**	-0.900	0.936**

*** p value < 0.01, ** $0.01 \leq p$ value < 0.05, * $0.05 \leq p$ value < 0.1

- Parameters regarding $R_{i,t+1}$ only significant for $Neg_{i,t}$ (*LM*, Extr. High)

Attention Analysis Regression Results ctd

- Similar results for median and high attention groups regarding $\sigma_{i,t+1}$ and $V_{i,t+1}$
- Differences for $R_{i,t+1}$:

	BL		LM		MPQA	
Attention	Median	High	Median	High	Median	High
	Panel C: Future Returns $R_{i,t+1}$					
$I_{i,t}$	-0.001	-0.000	0.000	0.000	0.001^{*}	-0.000
$Pos_{i,t}$	0.025	0.025^{*}	0.032	0.034	0.039^{**}	0.026^{**}
$Neg_{i,t}$	0.008	-0.031^{*}	-0.037	-0.050^{***}	0.002	-0.042^{**}

*** p value < 0.01, ** $0.01 \leq p$ value < 0.05, * $0.05 \leq p$ value < 0.1

Sector Analysis

- Compare financials sector with health care sector
- Attention ratio high for financials (0.413) and low for health care (0.287)
- *BL*, *MPQA*: no leverage effect of negative news for health care
- *LM*: very effective in financials not so much in health care

Entire Panel Results

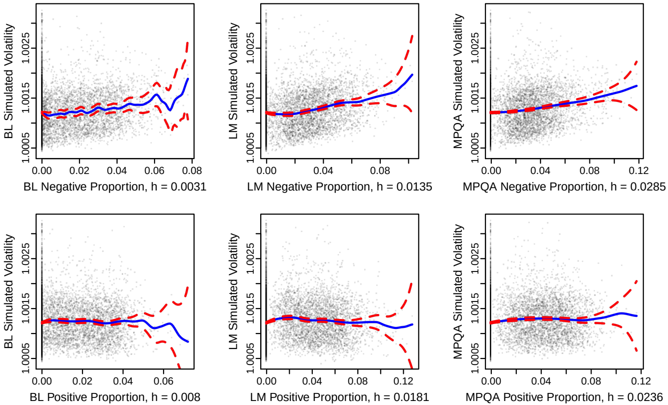

Figure 4: Volatility Simulation for Entire Panel: Mean curve, 95% Uniform Confidence Bands TXTSimulation Simulation Setup

Entire Panel Results ctd

- *LM* and *MPQA*: curve for $Neg_{i,t}$ significantly differs from curve for $Pos_{i,t}$
  - Range *LM*: 0.042 - 0.094
  - Range *MPQA*: 0.051 - 0.091
- Not the case for *BL*

Low Attention Results

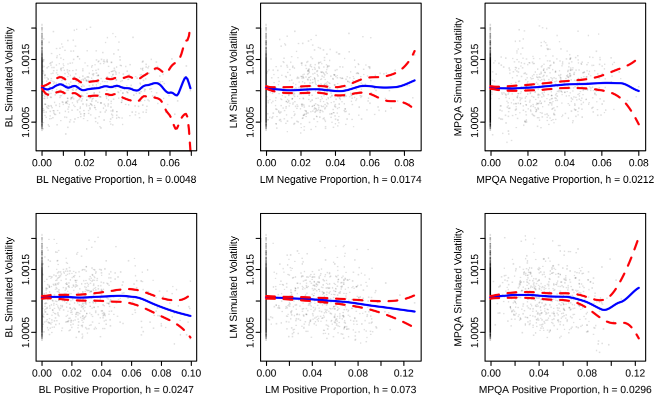

Figure 5: Volatility Simulation for Low Attention Group: Mean curve, 95% Uniform Confidence Bands TXTSimulationAttention Setup

Extremely High Attention Results

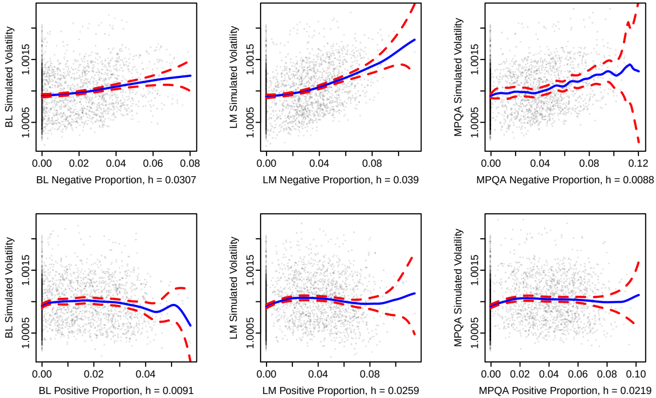

Figure 6: Volatility Simulation for Extremely High Attention Group: Mean curve, 95% Uniform Confidence Bands TXTSimulationAttention

Extremely High Attention Results ctd

- *BL* and *LM*: Curve for $Neg_{i,t}$ significantly differs from curve for $Pos_{i,t}$
- Not the case for *MPQA*

Are the Bands too narrow?

- Before: confidence bands based on asymptotic properties of normal distribution
- Alternative: bootstrap confidence bands for M-Smoother by Härdle (2015) Algorithm

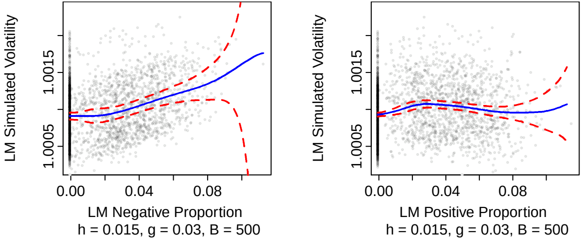

Figure 7: Volatility Simulation for Extremely High Attention Group: Mean curve, 95% Uniform Bootstrap Confidence Bands

Conclusion

- Sentiment measures: incremental information about future stock reactions
- Asymmetric impact of positive and negative sentiment
- Degree of incremental information and asymmetry is sector and attention specific
- Choice of lexicon matters

Distillation of News Flow into Analysis of Stock Reactions

Junni Zhang
Cathy Chen
Wolfgang Karl Härdle
Elisabeth Bommes

Guanghua School of Management
Peking University
Chung Hua University
Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics
and Economics
Humboldt–Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de

Frequency Table: GICS Sectors

GICS Sector	No. Stocks
Consumer Discretionary	21
Consumer Staples	9
Energy	6
Financials	12
Health Care	15
Industrials	10
Information Technology	21
Materials	4
Telecommunication Services	2

Back

Number of unique Words

- Some words only in one lexicon: "unique words"
- Number of unique words that appear at least three time in the articles:

Lexicon	Positive	Negative
BL	470	918
LM	267	916
MPQA	512	181

Most frequent Words unique to one Lexicon

BL		LM		MPQA	
Positive	Negative	Positive	Negative	Positive	Negative
Available	Debt	Opportunities	Declined	Just	Low
(5,836)	(12,540)	(4,720)	(9,809)	(17,769)	(12,739)
Led	Fell	Strength	Dropped	Help	Division
(5,774)	(9,274)	(4,393)	(4,894)	(17,334)	(5,594)
Lead	Fool	Profitability	Late	Profit	Least
(4,711)	(5,473)	(4,174)	(4,565)	(15,253)	(5,568)
Recovery	Issues	Highest	Claims	Even	Stake
(4,357)	(3,945)	(3,409)	(3,785)	(13,780)	(4,445)
Work	Risks	Greater	Closing	Deal	Slightly
(3,808)	(2,850)	(3,321)	(3,604)	(13,032)	(3,628)

Words only appear in one of the lexica and frequencies are given in parentheses.

Back

Number of shared Words

- Some words are only shared by two lexica
- Number of shared words that appear at least three time in the articles:

Lexicon	Positive	Negative
BL and LM	131	322
BL and MPQA	971	1,164
LM and MPQA	32	30

Most frequent shared Words

BL and LM		BL and MPQA		LM and MPQA	
Positive	Negative	Positive	Negative	Positive	Negative
Gains	Losses	Free	Gross	Despite	Against
(7,604)	(5,938)	(133,395)	(8,228)	(7,413)	(8,877)
Gained	Missed	Well	Risk	Able	Cut
(7,493)	(3,165)	(3,0270)	(7,471)	(5,246)	(3,401)
Improved	Declining	Like	Limited	Opportunity	Challenge
(7,407)	(3,053)	(24,617)	(5,884)	(4,398)	(1,042)
Improve	Failed	Top	Motley	Profitable	Serious
(5,726)	(2,421)	(14,899)	(5,165)	(3,580)	(1,022)
Restructuring	Concerned	Guidance	Crude	Efficiency	Contrary
(3,210)	(1,991)	(11,715)	(5,109)	(2,615)	(401)

Words are shared by only two lexica and frequencies are given in parentheses.

Back

Comparison of Lexical Projections

Variable	$\widehat{\mu}$	$\widehat{\sigma}$	Max	Q1	Q2	Q3	Polarity
Pos (BL)	0.033	0.012	0.134	0.025	0.032	0.040	88.04%
Neg (BL)	0.015	0.010	0.091	0.008	0.014	0.020	10.51%
Pos (LM)	0.014	0.007	0.074	0.009	0.013	0.018	55.70%
Neg (LM)	0.012	0.009	0.085	0.006	0.011	0.016	40.17%
Pos (MPQA)	0.038	0.012	0.134	0.031	0.038	0.045	96.26%
Neg (MPQA)	0.013	0.008	0.133	0.007	0.012	0.017	2.87%

Sample mean, sample standard deviation, maximum value, 1st, 2nd and 3rd quartiles, and polarity as relative dominance between positive and negative sentiment.

Back

Classification Evaluation

Manual Label	BL Label Pos	BL Label Neg	BL Label Neu	LM Label Pos	LM Label Neg	LM Label Neu	MPQA Label Pos	MPQA Label Neg	MPQA Label Neu	Total
Pos	56	4	1	41	12	8	61	0	0	61
Neg	9	2	1	0	9	3	9	2	1	12
Neu	22	5	0	10	15	2	26	0	1	27
Total	87	11	2	51	36	13	96	2	2	100

Back

Tagging Example - BL

... McDonald's has an obesity **problem** that continues to get **worse**. And that's nothing to do with the food itself, but rather the huge menus that can now double as medieval fortification. For perspective, the chain's menu has grown 70% since 2007. And while more offerings might seem **like** a **good** thing, large menus result in **slower** service and more flare-ups between franchisees and the corporation.
Bloated menus raise inventory costs for smaller franchisees and **lead** to lower profit margins. The McDonald's corporate franchise fee is based upon sales instead of profits, making it a smaller **concern** for the company overall. ...

3 **positive words** and 5 **negative words**

Article source

Tagging Example - LM

... McDonald's has an obesity **problem** that continues to get **worse**. And that's nothing to do with the food itself, but rather the huge menus that can now double as medieval fortification. For perspective, the chain's menu has grown 70% since 2007. And while more offerings might seem like a **good** thing, large menus result in **slower** service and more flare-ups between franchisees and the corporation.
Bloated menus raise inventory costs for smaller franchisees and lead to lower profit margins. The McDonald's corporate franchise fee is based upon sales instead of profits, making it a smaller **concern** for the company overall. ...

1 **positive word** and 4 **negative words**

Tagging Example - MPQA

... McDonald's has an obesity **problem** that continues to get **worse**. And that's nothing to do with the food itself, but rather the huge menus that can now double as medieval fortification. For perspective, the chain's menu has grown 70% since 2007. And while more offerings might seem **like** a **good** thing, **large** menus result in **slower** service and more flare-ups between franchisees and the corporation.
Bloated menus raise inventory costs for smaller franchisees and **lead** to lower **profit** margins. The McDonald's corporate franchise fee is based upon sales instead of **profits**, making it a smaller **concern** for the company overall. ...

5 **positive words** and 5 **negative words**

Back

Correlation - Positive Sentiment

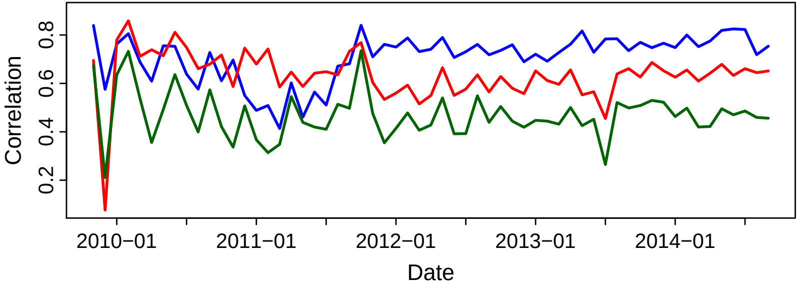

Figure 8: Monthly correlation between positive sentiment: BL and LM , BL and MPQA, LM and MPQA

Correlation - Negative Sentiment

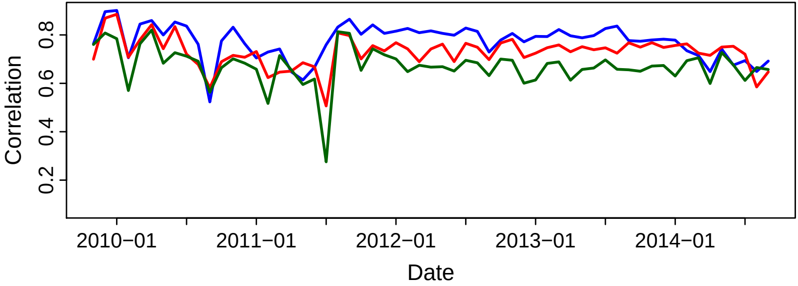

Figure 9: Monthly correlation between negative sentiment: BL and LM, BL and MPQA, LM and MPQA Back

Garman and Klass range-based Measure of Volatility

$$\sigma^2_{i,t} = 0.511(u-d)^2 - 0.019\left\{c(u+d) - 2ud\right\} - 0.383c^2 \qquad (7)$$

with $u = \log(P^H_{i,t}) - \log(P^L_{i,t}), \quad d = \log(P^L_{i,t}) - \log(P^O_{i,t}),$

$c = \log(P^C_{i,t}) - \log(P^O_{i,t})$

for company i on day t with $P^H_{i,t}$, $P^L_{i,t}$, $P^O_{i,t}$, $P^C_{i,t}$ as highest, lowest, opening and closing stock prices, respectively.

Back

Simulation Setup

- Evaluate the asymmetric reaction of volatility to sentiment
- $I_{i,t} \sim B(1, p_i)$
- $Pos_{i,t} \sim U(0, m_{Pos,i})$, $m_{Pos,i} = \max(Pos_i)$
- $Neg_{i,t} \sim U(0, m_{Neg,i})$, $m_{Neg,i} = \max(Neg_i)$
- Correlation of $Pos_{i,t}$ and $Neg_{i,t}$: Cholesky Decomposition

Simulation Setup ctd

- $R_{M,t} \sim G_\gamma(\mu, \sigma)$
  - Generalized Extreme Value Distribution
  - Estimate parameters from sample period
  - $\mu = 0.64$, $\sigma = 0.35$ and $\gamma = 0.20$

Simulation Setup ctd

- $R_{i,t} - R_{f,t} = \beta_i(R_{M,t} - R_{f,t})$
  - CAPM by Sharpe (1964) and Lintner (1965)
  - Systematic risk β_i
  - Risk-free rate $R_{f,t} = 1\%$ p.a.

Back to Entire Panel Back to Attention Panel

Algorithm: Bootstrap Confidence Bands I

1) Compute $\hat{m}_h(x)$ by using the curve estimator proposed by Nadaraya(1964) and Watson(1964):

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i) Y_i}{\sum_{i=1}^{n} K_h(x - X_i)}$$

with $K_h(u) = \varphi(u/h)/h$ denoting the Gaussian kernel and set $\hat{\varepsilon}_i \stackrel{def}{=} Y_i - \hat{m}_h(X_i)$. To ensure robustness against outliers, this estimator is adjusted as proposed by Brillinger (1977).

Algorithm: Bootstrap Confidence Bands II

2) Compute the estimated conditional distribution function $\hat{F}_{(\varepsilon|X)}(\cdot)$ with Gaussian kernel.
3) Construct $j = 1, \dots, J$ samples by generating the random variables $\varepsilon_i^* \sim \hat{F}_{(\varepsilon|X=X_i)}$ with $i = 1, \dots, n$ for each sample. Compute

$$Y_i^* = \hat{m}_g(X_i) + \varepsilon_i^*$$

with g chosen such that $\hat{m}_g(X_i)$ is slightly oversmoothed.

Algorithm: Bootstrap Confidence Bands III

4) For each bootstrap sample $\{X_i, Y_i^*\}_{i=1}^n$, compute $\hat{m}^*_{h,g}(\cdot)$ and the random variable

$$d_j \stackrel{def}{=} \sup_{x \in B} [|\hat{m}^*_{h,g}(x) - \hat{m}_g(x)| \sqrt{\hat{f}_X(x) \hat{f}_{(\varepsilon|X)}(x)\}} / \sqrt{\widehat{\mathsf{E}}_{\varepsilon|X}\{\psi^2(\varepsilon)\}}],$$

$$j = 1, \ldots, J$$

for a finite number of points in the compact set B. Both $\hat{f}_{(\varepsilon|X)}(x)$ and $\widehat{\mathsf{E}}_{\varepsilon|X}\{\psi^2(\varepsilon)\}$ are computed using the estimated residuals $\hat{\varepsilon}_i$. $\psi(\cdot)$ denotes the ψ-function by Huber(1981) with $\psi(u) = \max\{-c, \min(u, c)\}$ for $c > 0$.

Algorithm: Bootstrap Confidence Bands IV

5) Calculate the $1-\alpha$ quantile d_α^* of $d_1, \ldots, d_J$.
6) Construct the bootstrap uniform band centered around $\hat{m}_h(x)$

$$\hat{m}_h(x) \pm [\sqrt{\hat{f}_X(x)}\hat{f}_{(\varepsilon|X)}(x)\}/\sqrt{\widehat{\mathsf{E}}_{\varepsilon|X}\{\psi^2(\varepsilon)\}}]^{-1} d_\alpha^*.$$

Back

Bibliography I

Chen, Z., Daigler, R. T., and Parhizgari, A. M.
Persistence of volatility in futures markets
J. Futures Markets, 2006

Brillinger, D. R.
In discussion of "consistent nonparametric regression" by C.J. Stone
Ann. Statist., 1977

Garman, M. and Klass, M.
On the Estimation of Security Price Volatilities from Historical Data
J. Bus., 1980

Bibliography II

Girard, E. and Biswas, R.
Trading volume and market volatility
Financ. Rev., 2007

Härdle, W. K. and Ritov, Y. and Wang, W.
Tie the straps
J. Multivariate Anal., 2015

Hu, M. and Liu, B.
Mining and Summarizing Customer Reviews
10th ACM SIGKDD, 2004

Bibliography III

Huber, P. J.
Robust statistics
Wiley, 1981

Loughran, T. and McDonald, B.
When is a liability not a liability?
J. Financ., 2011

Nadaraya, E. A.
On estimating regression
Theor. Probab. Appl., 1964

Bibliography IV

Shu, J. and Zhang, J. E.
Testing range estimators of historical volatility
J. Futures Markets, 2006

Watson, G. S.
Smooth regression
Indian J. Statist., 1964

Wilson, T. and Wiebe, J. and Hoffmann, P.
Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis
HLT-EMNLP, 2005