Textual sentiment and sector-specific reaction

Wolfgang Karl Härdle
Cathy Yi-Hsuan Chen
Elisabeth Bommes

Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin
http://lvb.wiwi.hu-berlin.de

News moves Markets

- Zhang et al. (2016): textual sentiment provides incremental information about future stock reactions
- Sectors react differently to sentiment
- Unsupervised vs. supervised approach in sentiment projection

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 vs. noise
  - Signal: nuance of context
  - Noise: increasing imprecision of deep parsing

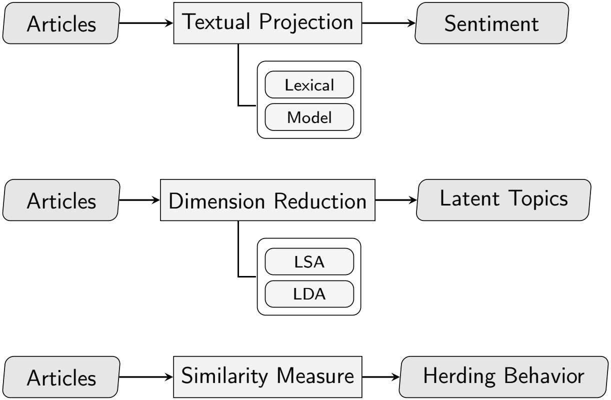
Dimensions of News ctd

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

Challenge

- News are sector-specific
- How to distill sentiment across various sectors

The Power of Words: Textual Analytics



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)

Lexicon Correlation

Unsupervised Projection

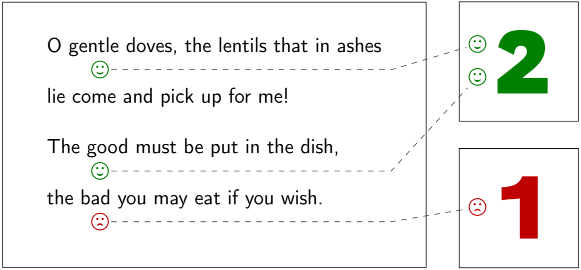


Figure 1: Example of Text Numerisization

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

Examples

Supervised Projection

- Training data: Financial Phrase Bank by Malo et al. (2014)
  - Sentence-level annotation of financial news
  - Manual annotation of 5,000 sentences by 16 annotators

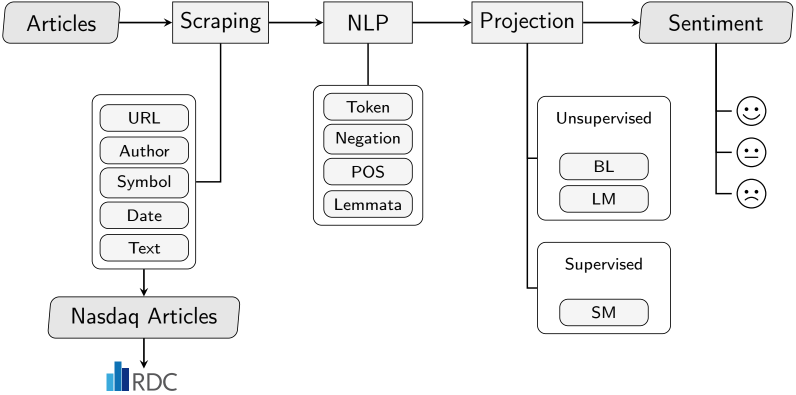
Research Questions

- Is the sentiment effect sector specific?
- Is supervised learning an effective approach in text classification?

Outline

1. Motivation ✓
2. Data Collection
3. Sentiment Projection
4. Panel Regression
5. ARIMA-GARCH
6. Outlook

How to gather Sentiment Variables?



Nasdaq Articles

- Terms of Service permit web scraping
- Currently $>$ 440k articles between October 2009 and January 2016
- Data available at RDC

Sector-specific articles

Sector	Abbr.	# Articles	# Comp.
Consumer Discretionary	CD	44,454	84
Consumer Staples	CS	19,435	40
Energy	EN	18,069	43
Financials	FI	37,614	85
Health Care	HC	23,838	55
Industrials	IN	24,124	64
Information Technology	IT	44,967	65
Materials	MA	10,947	30
Telecommunication Services	TE	5,963	5
Utilities	UT	6,078	30

Table 1: Number of Articles per Sector between 10/2009 and 01/2016

Top Word Frequencies

Word	Freq. (in k)	Sector Freq. Top 5	Sector Freq. Top 10
free	649	10	10
well	238	9	10
gold	235	1	1
best	207	9	10
fool	200	5	8
strong	196	5	10
like	172	5	10
top	167	3	10
better	162	0	9
motley	152	2	7

Table 2: Most frequent words of either BL or LM

Article Timeline

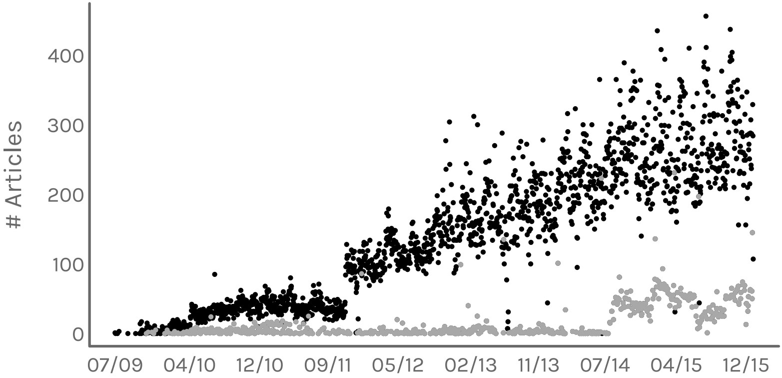


Figure 2: Number of Sector-specific Articles per Day (no Trading)

Lexicon-based Sentiment

Consider document i, positive sentiment Pos_i, positive lexicon entries W_j ($j = 1, \ldots, J$) and count frequency of those entries w_j:

$$Pos_i = n_i^{-1} \sum_{j=1}^{J} \mathbf{I}\left(W_j \in L\right) w_j \tag{1}$$

with n_i: number of words in document i (e.g. sentence)

Equivalent calculation of negative sentiment Neg_i

Sentence-level Polarity

$$Pol_i = \begin{cases} 1, & \text{if } Pos_i > Neg_i \\ 0, & \text{if } Pos_i = Neg_i \\ -1, & \text{if } Pos_i < Neg_i \end{cases} \tag{2}$$

for sentence i.

- Measure sentiment on sentence-level

Regularized Linear Models (RLM)

- Training data $(X_1, y_1) \dots (X_n, y_n)$ with $X_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$
- Linear scoring function $s(X) = \beta^\top X$ with $\beta \in \mathbb{R}^p$

Example

Regularized training error:

$$n^{-1} \sum_{i=1}^{n} \underbrace{L\{y_i, s(X)\}}_{\text{Loss Function}} + \lambda \underbrace{R(\beta)}_{\text{Regularization Term}} \tag{3}$$

with hyperparameter $\lambda \geq 0$.

RLM Estimation

- Optimize via Stochastic Gradient Descent More
- 5-fold cross validation More
- Oversampling More
- Choice of: $L(\cdot), R(\cdot), \lambda$, X (n-gram range, features) ...
- Three categories: one vs. all sub-models

Bullishness

$$B = \log[\{1 + n^{-1} \sum_{j=1}^{n} \mathrm{I}\left(Pol_j = 1\right)\}/\{1 + n^{-1} \sum_{j=1}^{n} \mathrm{I}\left(Pol_j = -1\right)\}] \tag{4}$$

by Antweiler and Frank (2004) with $j = 1, \ldots, n$ sentences in document.

- $B_{i,t}$ accounts for bullishness of company i on day t
- Consider $|B_{i,t}|$ and $BN_{i,t} = \mathrm{I}\left(B_{i,t} < 0\right) B_{i,t}$

Model Accuracy - Polarity

Supervised Learning

- Chosen model: Hinge loss, L1 norm, $\lambda = 0.0001$, ...
- Mean accuracy (oversampling): 0.80
- Mean accuracy (normal sample): 0.82

Lexicon-based

- Mean accuracy BL: 0.58
- Mean accuracy LM: 0.64

Evaluation BL

True \ Pred	-1	0	1	Total
-1	**214**	268	32	514
0	203	**1,786**	546	2,535
1	89	627	**452**	1,168
Total	506	2,681	1,030	4,217

Table 3: Confusion Matrix - BL Lexicon TXTfpblexical

Evaluation LM

True \ Pred	-1	0	1	Total
-1	**213**	289	12	514
0	200	**2,187**	148	2,535
1	111	772	**285**	1,168
Total	524	3,248	445	4,217

Table 4: Confusion Matrix - LM Lexicon TXTfpblexical

Evaluation SM

Pred / True	-1	0	1	Total
-1	**389**	67	58	514
0	96	**2,134**	305	2,535
1	105	198	**916**	1,168
Total	539	2,399	1,279	4,217

Table 5: Confusion Matrix - Supervised Learning, estimated with Oversampling and evaluated on total Sample TXTfpbsupervised

Confusion Matrix with Oversampling | Choice of λ | Results Logistic Loss

Sectors as Panels

$$\log \sigma_{i,t} = \alpha + \beta_1 |B_{i,t}| + \beta_2 BN_{i,t} + \beta_3^\top X_{i,t} + \gamma_i + \varepsilon_{i,t} \quad (5)$$

$$R_{i,t} = \alpha + \beta_1 B_{i,t} + \beta_2^\top X_{i,t} + \gamma_i + \varepsilon_{i,t} \quad (6)$$

for stock i on day t with separate estimation of (5) and (6).

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

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

Returns

$$R_{i,t} = \log(P^C_{i,t}) - \log(P^C_{i,t-1}) \tag{7}$$

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

Regression - GK Log Volatility

		CD	CS	EN	FI	HC	IN	IT	MA	TE	UT
$\lvert B_{i,t} \rvert$	BL		0.01	0.01	0.05	0.01	0.1	0.05	0.01		0.01
	LM		0.01	0.01		0.01			0.05	0.01	
	SM	0.01	0.01	0.01	0.01	0.01	0.05	0.01	0.01		0.01
$BN_{i,t}$	BL			0.01						0.01	0.01
	LM			0.1					0.05	0.01	
	SM		0.01				0.01				0.05

Table 6: Significance codes 0.01 0.05 0.1

Abbreviations

Regression - Returns

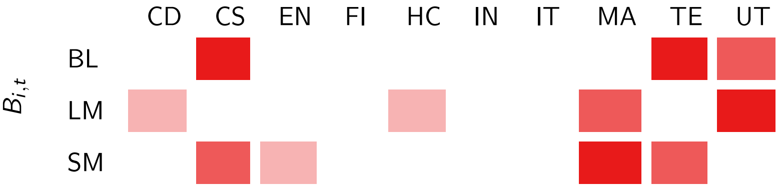


Table 7: Significance codes 0.01 0.05 0.1

Abbreviations

S&P 500 Sector Indices

AR(1)-GARCH(1, 1) model with control variables

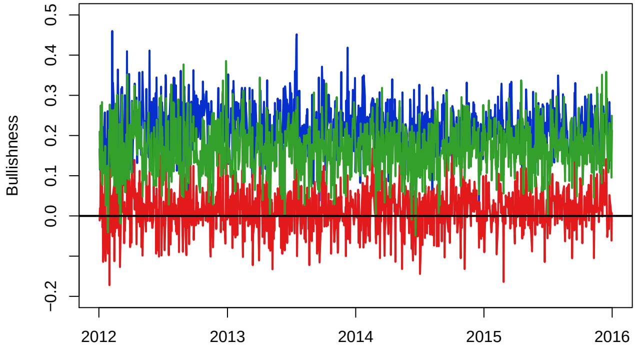
$$R_{i,t} = c_i + \varphi R_{i,t} + \varepsilon_{i,t} \tag{8}$$

$$\sigma^2_{i,t} = \omega_i + \alpha_i \varepsilon^2_{i,t-1} + \beta_i \sigma^2_{i,t-1} + \theta_i PF_{i,t-1} + \gamma_i NF_{i,t-1} \tag{9}$$

for sector index i on day t.

$PF_{i,t}$ - Fraction of positive words
$NF_{i,t}$ - Fraction of negative words

Why not Bullishness?



- Financial sector, BL (green), LM (red), SM (blue)
- Aggregated news for markets are very bullish
- Potential news bias?

Regression Results

		CD	CS	EN	FI	HC	IN	IT	MA	TE	UT
PF	BL		■	■		■	■				■
	LM										■
	SM		■			■		■			■
NF	BL	■									■
	LM		■								■
	SM		■		■		■				

Table 8: Significance codes ■ 0.01 ■ 0.05 ■ 0.1

Financials Lags

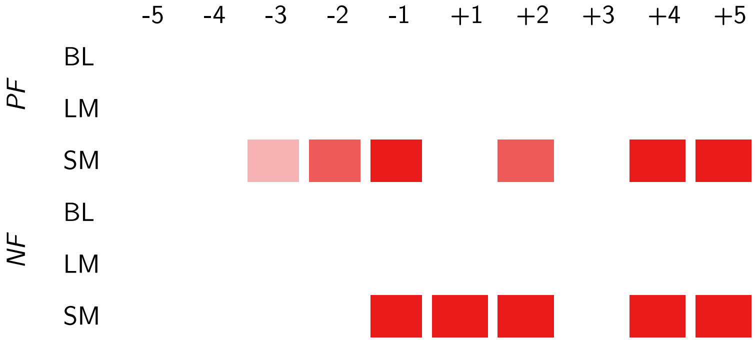


Table 9: Significance codes 0.01 0.05 0.1

What's next?

- Closer look at sectors : sectoral attributes, concentration, competition...
- Textual sentiment spillover : network modelling

Textual sentiment and sector-specific reaction

Wolfgang Karl Härdle
Cathy Yi-Hsuan Chen
Elisabeth Bommes

Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin
http://lvb.wiwi.hu-berlin.de

Bibliography

Antweiler, W. and Frank, M. Z.
Is All That Talk Just Noise?
J. Fin., 2004

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

Härdle, W. K. and Lee, Y. J. and Schäfer D. and Yeh Y. R.
Variable Selection and Oversampling in the Use of Smooth Support Vector Machines for Predicting the Default Risk of Companies
J. Forecast., 2009

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

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

Malo, Pekka and Sinha, Ankur and Korhonen, Pekka and Wallenius, Jyrki and Takala, Pyry
Good debt or bad debt
Journal of the Association for Information Science and Technology, 2014

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

Zhang, J., Chen C. Y., Härdle, W. K. and Bommes, E.
Distillation of News into Analysis of Stock Reactions
JBES, 2016

Zhang, X., Yichao, W., Wang, L. and Runze, L.
A Consistent Information Criterion for Support Vector Machines in Diverging Model Spaces
J. Mach. Learn. Res., 2016

Appendix

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**

TXTMcDbm
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**

TXTMcDlm

Back

Correlation - Positive Sentiment

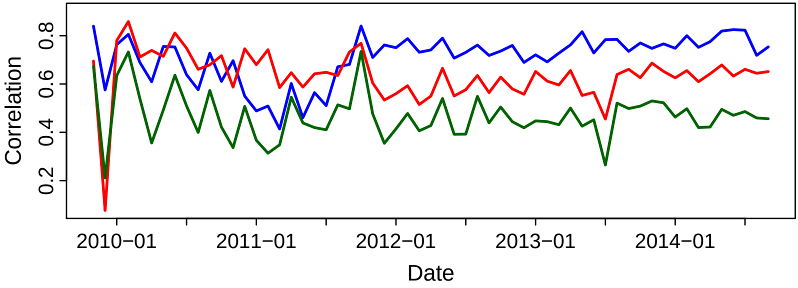


Figure 3: Monthly correlation between positive sentiment: BL and LM , BL and MPQA, LM and MPQA. Source: Zhang et al. (2016)

Correlation - Negative Sentiment

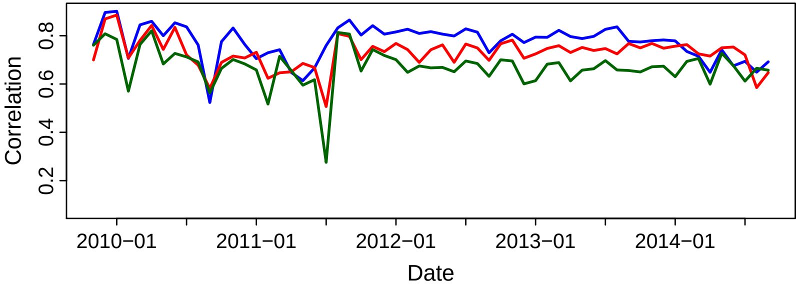


Figure 4: Monthly correlation between negative sentiment: BL and LM, BL and MPQA, LM and MPQA. Source: Zhang et al. (2016) Back

Web Scraping

- Databases to buy?
- Automatically extract information from web pages
- Transform unstructured data (HTML) to structured data
- Use HTML tree structure to parse web page
- Legal issues
  - Websites protected by copyright law
  - Prohibition of web scraping possible
  - Comply to Terms of Service (TOS)

Back

Natural Language Processing (NLP)

- Text is unstructured data with implicit structure
  - Text, sentences, words, characters
  - Nouns, verbs, adjectives, ..
  - Grammar
- Transform implicit text structure into explicit structure
- Reduce text variation for further analysis
- Python Natural Language Toolkit (NLTK)
- TXTnlp

Tokenization

- String

  ``McDonald's has its work cut out for it. Not only are sales falling in the U.S., but the company is now experiencing problems abroad.''

- Sentences

  ``McDonald's has its work cut out for it.'',
  ``Not only are sales falling in the U.S., but the company is now experiencing problems abroad.''

- Words

  ``McDonald'', ``'s'', ``has'', ``its'', ``work'', ``cut'', ``out'' ...

Negation Handling

- “not good” $\neq$ “good”
- Reverse polarity of word if negation word is nearby
- Negation words
  `"n’t", "not", "never", "no", "neither", "nor", "none"`

Part of Speech Tagging (POS)

- Grammatical tagging of words
  - `dogs - noun, plural (NNS)`
  - `saw - verb, past tense (VBD) or noun, singular (NN)`
- Penn Treebank POS tags
- Stochastic model or rule-based

Lemmatization

- Determine canonical form of word
  - `dogs - dog`
  - `saw (verb) - see and saw (noun) - saw`
- Reduces dimension of text
- Takes POS into account
  - Porter stemmer: `saw (verb and noun) - saw`

Back

Loss Functions for Classification

- Logistic: Logit

$$L\{y, s(X)\} = \log(2)^{-1} \log[1 + \exp\{-s(X)y\}] \tag{10}$$

- Hinge: Support Vector Machines

$$L\{y, s(X)\} = \max\{0, 1 - s(X)y\} \tag{11}$$

Back

Regularization Term

- L2 norm

$$R(\beta) = 2^{-1} \sum_{i=1}^{p} \beta_i^2 \tag{12}$$

- L1 norm

$$R(\beta) = \sum_{i=1}^{p} |\beta_i| \tag{13}$$

Back

RLM Example

Sentence 1: "The profit of Apple increased."
Sentence 2: "The profit of the company decreased."

$$y = (1, -1) \quad (14)$$

$$X = \begin{array}{r} \\ \textit{the} \\ \textit{profit} \\ \textit{of} \\ \textit{Apple} \\ \textit{increased} \\ \textit{company} \\ \textit{decreased} \end{array} \begin{array}{c} \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array} \quad (15)$$

Back

k-fold Cross Validation (CV)

- Partition data into k complementary subsets
- No loss of information as in conventional validation
- Stratified CV: equally distributed response variable in each fold

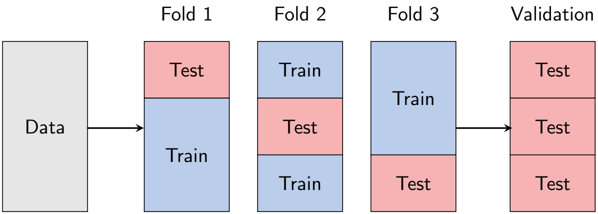


Figure 5: 3-fold Cross Validation

Back

Oversampling

- Härdle et al. (2009) Trade-off between Type I and Type 2 error in classification Error types
- Balance size of neutral sentences and ones with polarity in sample
- Duplicate sentences within folds of stratified cross validation until the sample is balanced

Back

Classification Error Rates

- Type I error rate $= FN/(FN + TP)$
- Type II error rate $= FP/(FP + TN)$
- Total error rate $= (FN + FP)/(TP + TN + FP + FN)$

with TP as true positive, TN as true negative, FP as false positive and FN as false negative.

Back

Stochastic Gradient Descent (SGD)

- Approximately minimize loss function

$$L(\theta) = \sum_{i=1}^{n} L_i(\theta) \tag{16}$$

- Iteratively update

$$\theta_i = \theta_{i-1} - \eta \, \frac{\partial L_i(\theta)}{\partial \theta} \tag{17}$$

SGD Algorithm

1. Choose learning rate η
2. Shuffle data
3. For $i = 1, \ldots, n$, do:

$$\theta_i = \theta_{i-1} - \eta \frac{\partial L_i(\theta)}{\partial \theta}$$

Repeat 2 and 3 until approximate minimum obtained.

SGD Example

$X \sim \mathrm{N}(\mu, \sigma)$ and $x_1, ..., x_n$ as randomly drawn sample

$$\min_{\theta} \; n^{-1} \sum_{i=1}^{n} (\theta - x_i)^2$$

Update step

$$\theta_i = \theta_{i-1} - 2\eta(\theta_{i-1} - x_i)$$

Optimal gain

Set $2\eta = 1/i$ and obtain $\theta_n = \bar{x}$ with $\bar{x}$ as sample mean.

SGD Example ctd

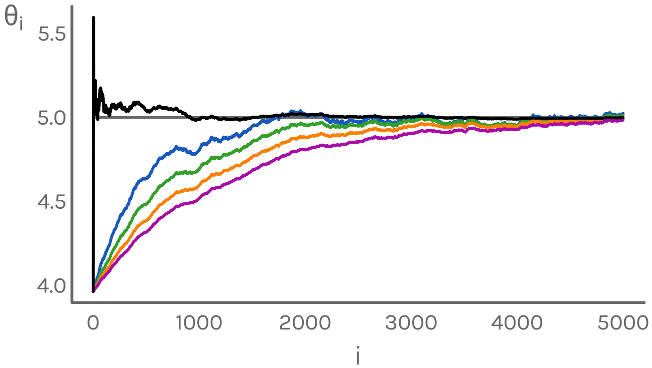


Figure 6: Estimate Mean via SGD, $x_t \sim \mathrm{N}(5, 1)$

$\eta \in \{1/t,\ 1/1000,\ 1/1500,\ 1/2000,\ 1/2500\}$ TXTSGD

Back

Garman and Klass range-based Measure of Volatility

$$\sigma^2_{i,t} = 0.511(u-d)^2 - 0.019\left\{c(u+d) - 2ud\right\} - 0.383c^2 \qquad (18)$$

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

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

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

Evaluation Supervised Learning

True \ Pred	-1	0	1	Total
-1	**1,983**	298	254	2,535
0	96	**2,134**	305	2,535
1	105	469	**1,961**	2,535
Total	2,184	2,901	2,520	7,605

Table 10: Confusion Matrix - Supervised Learning with Oversampling

Back

Choice of λ

- Fine grid with $\lambda_i \in [5 \cdot 10^{-6}, 0.05]$, $i = 1, \ldots, 9999$
- Estimate penalized SVM model
- Results remain stable
  - $\hat{\lambda}_{CV} = 0.000155$
  - Accuracy: 0.8

Choice of λ also possible via information criterion, e.g. Zhang et al. (2016)

Back

Evaluation Logistic Loss Function

True \ Pred	-1	0	1	Total
-1	**397**	55	62	514
0	103	**2,115**	317	2,535
1	58	193	**917**	1,168
Total	558	2,363	1,296	4,217

Table 11: Confusion Matrix - Supervised Learning, estimated with Over-sampling and evaluated on total Sample, Accuracy: 0.80

Back

Abbreviations

Sector	Abbreviation
Consumer Discretionary	CD
Consumer Staples	CS
Energy	EN
Financials	FI
Health Care	HC
Industrials	IN
Information Technology	IT
Materials	MA
Telecommunication	TE
Utilities	UT

Table 12: Sector Abbreviations

Volatility Regression Returns Regression

Control Variables

$R_{M,t}$ - S&P 500 index return

$\log VIX_t$ - CBOE VIX More Information

$\log \sigma_{i,t}$ - Range-based volatility

$R_{i,t}$ - Return

Back

VIX

- Implied volatility
- Measures market expectation of S&P 500
- Calculated by Chicago Board Options Exchange (CBOE)
- Measures 30-day expected volatility
- Calculated with put and call options with more than 23 days and less than 37 days to expiration

Back