Tales of sentiment driven tails

Jozef Baruník
Cathy Yi-Hsuan Chen
Wolfgang Karl Härdle

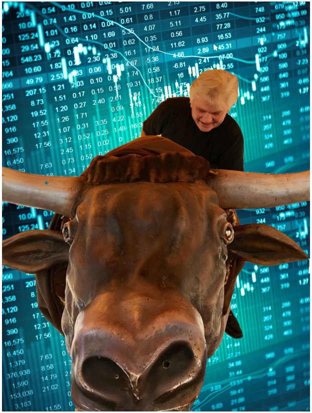
Institute of Economic Studies
Charles University in Prague
Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin

http://ies.fsv.cuni.cz
http://lvb.wiwi.hu-berlin.de

"Forget the dot-com boom with its irrational exuberance and the real estate bubble that was supposed to be invincible: Current market sentiment eclipses all of that"

Jeff Cox, CNBC, March 1 2017

Sentiment moves market



John Maynard Keynes (1936): markets can fluctuate wildly under the influence of investors' "animal spirits," which move prices in a way unrelated to fundamentals.

Sentiment can cause mispricing

Fifty years later...

De Long, Shleifer, Summers, and Waldmann (1990) formalized the role of investor sentiment in financial markets.

- uninformed noise traders base their decisions on sentiment
  - greater mispricing (Stambaugh et al., 2012)
  - excess volatility (Dumas et al., 2009)

"Now, the question is no longer, as it was a few decades ago, whether investor sentiment affects stock prices, but rather how to measure investor sentiment and quantify its effects."

(Baker and Wurgler, 2007)

News moves markets

- Baker and Wurgler (2007) investor sentiment affects securities whose valuations are highly subjective
- Large literature Huang et al. (2014), Da et al. (2015), Shefrin (2007+)
- Zhang et al. (2016) textual sentiment provides incremental information about future stock reactions

Is average enough?

- Sentiment affects cross section of returns or volatility
- Grand average is OK for expected payoffs
- Though...
  - bear vs. bull markets
  - extreme negative vs. positive returns

Is average man enough?

Contrarians vs. Trend followers

We already know that we can measure sentiment...

but how to quantify its effect on prices?

Tales of sentiment driven tails

Contribution

- Step forward from classical asset pricing (EU based)
- Provide decision-theoretic foundations of pricing in quantiles
- Link sentiment with quantiles of the return distributions
- Nonlinear dynamic quantile asset pricing model
- Confirm empirically on Panel of 100 US stocks

Outline

1. Motivation ✓
2. Theoretical Framework
3. Data Collection
4. Sentiment Projection
5. Calibration of weighting function
6. Quantile Panel Regressions
7. Outlook

Classical asset pricing

Investor maximizes utility subject to budget constraint. The FOC (Euler equation):

$$\mathsf{E}_F\left[M \times (1+R)\right] = 1, \qquad (1)$$

where M is a pricing kernel (PK), or stochastic discount factor (SDF), R is the total return on a risky asset with physical distribution $F(R)$.

Probability weighting

Decisions under risk are more sensitive to changes in probability of events at extremes, Tversky and Kahneman (1992).
Polkovnichenko and Zhao (2013) use the rank-dependent expected utility (RDEU) $\mathcal{U}(R) = \mathsf{E}_F[u(R)g\{F(R)\}]$ with PK

$$M = u'(R)g\{F(R)\}, \tag{2}$$

where $g\{F(R)\} = G'\{F(R)\}$ is a probability weighting function.

Euler equation reads as:

$$\mathsf{E}_F[u'(R)g\{F(R)\}\ (1+R)] = 1. \tag{3}$$

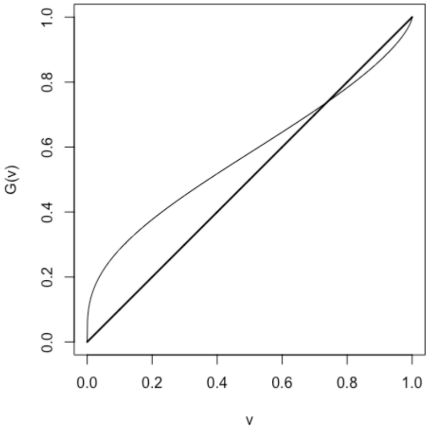


Figure: Probability weighting function $G(v) = \exp\{-(-\beta \log v)^{\alpha}\}$ with $\alpha = 0.7$ and $\beta = 0.6$

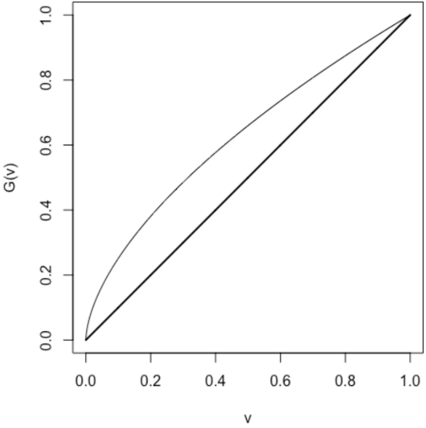


Figure: Probability weighting function $G(v) = \exp\{-(-\beta \log v)^{\alpha}\} = v^{0.6}$ ($\alpha = 1$ and $\beta = 0.6$)

A route towards quantile preferences

- X is preferred to Y if there exist utility function $\mathcal{U}(.)$ such that

$$X \succeq Y \text{ iff } \mathrm{E}_F[\mathcal{U}(X)] \geq \mathrm{E}_F[\mathcal{U}(Y)] \tag{4}$$

- Manski (1988), Rostek (2010) look at τ-quantile preferences

$$X \succeq Y \text{ iff } \mathrm{Q}_\tau[\mathcal{U}(X)] \geq \mathrm{Q}_\tau[\mathcal{U}(Y)] \tag{5}$$

- Maximising lower quantile is more risk-averse than higher quantile (example of portfolio), de Castro et al. (2017)

Example

Utility function $u(x) \stackrel{\text{def}}{=} x$

$$X = \begin{cases} 10^7 & \text{with } p = 10^{-6} \\ -1 & \text{with } q = 1 - p \end{cases} \qquad Y = \begin{cases} 10 & \text{with } p = 9/10 \\ -1 & \text{with } q = 1 - p \end{cases}$$

$X \succeq_{\mathrm{E}} Y$ since $\mathrm{E}[X] = 9 + 10^{-6}$ and $\mathrm{E}[Y] = 8 + 9/10$

$\mathrm{Q}_\tau(X) \stackrel{\text{def}}{=} \inf\{\alpha \in \mathbb{R} : \mathbb{P}(X \leq \alpha) \geq \tau\}$

$$X \begin{cases} \equiv_{\mathrm{Q}_\tau} Y \text{ for } \tau \leq 1/10 \\ \preceq_{\mathrm{Q}_\tau} Y \text{ for } 1/10 < \tau \leq 1 - 10^{-6} \\ \succeq_{\mathrm{Q}_\tau} Y \text{ for } \tau > 1 - 10^{-6} \end{cases}$$

A route towards a (dynamic) quantile model

Instead of classical preferences, look at an agent maximizing her stream of the future quantile utilities.
For a given $\tau \in (0,1)$, Euler equation reads:

$$Q_\tau \left[u'(R)g(v)\ (1+R)\right] = 1, \tag{6}$$

where $v = F(R)$,

$G(\cdot) : [0,1] \to [0,1]$ probability weighting fct and

$g(\cdot) = G^{'}(\cdot)$.

Can we relate $g(\cdot)$ to sentiment?

Probability weighting function and sentiment

Prelec (1998) weighting function:

$$G(v) = G(\alpha, \beta; v) = \exp\{-(-\beta \log v)^{\alpha}\} \tag{7}$$

α, β parameters govern the shape of $G(\cdot)$.

Link sentiment S_t to β_t:

$$\beta_t = \beta(S_t, \rho) = \exp\{-\rho(S_t^{-1} - 1)\} - 1 \tag{8}$$

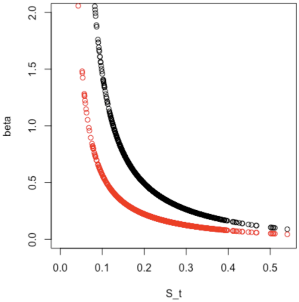


Figure: β versus S_t for $\rho = -0.1$ and $\rho = -0.05$

Fix $\alpha = 1$ to impose monotonicity and compute $v_t = (\operatorname{rank} R_t)/n$

$$
\begin{aligned}
G(v_t, S_t) &= v_t^{\beta(S_t,\rho)} = v_t^{\exp\{-\rho(S_t^{-1}-1)\}-1} \\
G(v_t, S_t) &= \exp\{(\exp\{-\rho(S_t^{-1}-1)\}-1)\log v_t\}
\end{aligned} \tag{9}
$$

A dynamic quantile model with sentiments

Equation (6) is beneficial, since it can be log-linearized as for a general random variable W, $\mathrm{Q}_\tau[\log(W)] = \log(\mathrm{Q}_\tau[W])$.
Hence

$$\mathrm{Q}_\tau\left[u'(R_t)g(v_t,S_t)\ (1+R_{t+1})\right] = 1 \tag{10}$$

considering power utility function:

$$\mathrm{Q}_\tau\left[-\gamma\log(R_t)+\log\{g(v_t,S_t)\}+\log(1+R_{t+1})\right] = 0. \tag{11}$$

One can estimate the parameter driving $g(v_t,S_t)$ with nonlinear quantile regression.

How to estimate sentiment S_t?

Data

- Panel of 100 most liquid constituents of S&P 500 stocks
- Sentiment variables: distilled from Nasdaq articles

Nasdaq Articles

- Terms of Service permit web scraping
- Currently $>$ 580k articles between October 2009 and January 2017
- Data available at RDC

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

The Power of Words: Textual Analytics

- Sentiment analysis
  - Lexica projection : positive, neutral and negative
  - Machine learning : text classification

Unsupervised Projection

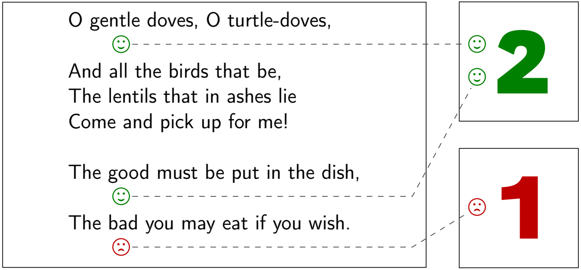


Figure: Example of Text Numerisization

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

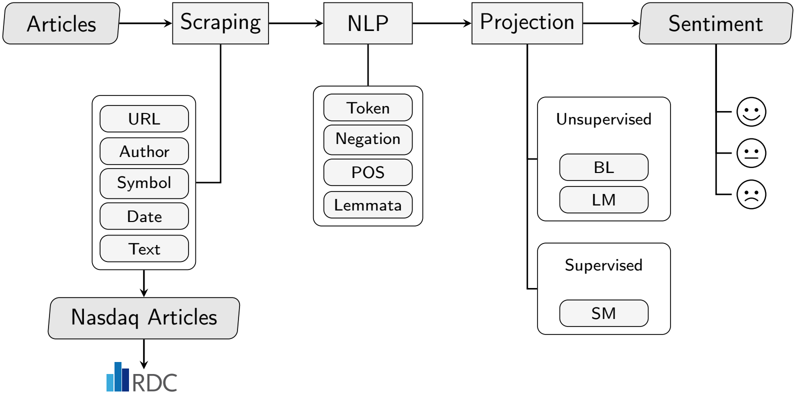
Examples

Supervised Projection

We use supervised projection (Zhang et al., JBES, 2016)

- 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: to incorporate human knowledge
  - Example: "profit" with different semantic orientations
    - Neutral in "profit was 1 million"
    - Positive in "profit increased from last year"

How to gather Sentiment Variables?



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:

$$Pos_i = n_i^{-1} \sum_{j=1}^{J} \mathbf{I}\left(W_j \in L\right) w_j \tag{12}$$

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

Equivalent calculation of negative sentiment Neg_i

Sentence-level Polarity

$$Pol_i = \begin{cases} 1, & \text{if } Pos_i > Neg_i \\ 0, & \text{if } Pos_i = Neg_i \\ -1, & \text{if } Pos_i < Neg_i \end{cases} \tag{13}$$

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:

$$n^{-1} \sum_{i=1}^{n} \underbrace{L\{y_i, s(X)\}}_{\text{Loss Function}} + \lambda \underbrace{R(\beta)}_{\text{Regularization Term}} \tag{14}$$

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

$$B = \log \left\{ \frac{1 + n^{-1} \sum_{j=1}^{n} \mathbf{I}\left(Pol_j = 1\right)}{1 + n^{-1} \sum_{j=1}^{n} \mathbf{I}\left(Pol_j = -1\right)} \right\} \tag{15}$$

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

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

Calibration of probability weighting functions

Estimate ρ_τ using nonlinear quantile regressions.
Employ power utility $u(R) = R^{1-\gamma}/(1-\gamma)$.

$$\mathrm{Q}_\tau\left[-\gamma\log(R_t) + \log\{g(v_t, S_t)\} + \log(1 + R_{t+1})\right] = 0, \qquad (16)$$

recall

$$g(v, S) = G'(1, \beta; v) = \beta v^{\beta-1},$$
$$\beta = \beta(S, \rho) = \exp\{-\rho(S^{-1} - 1)\} - 1.$$

where

$$g(v_t, S_t) = (exp\{-\rho(1/S_t - 1)\} - 1)v_t^{exp\{-\rho(1/S_t-1)\}-2}.$$

Calibration of probability weighting functions

Expect ρ_τ to differ across τ since sentiment distorts beliefs of a τ-quantile preference maker.

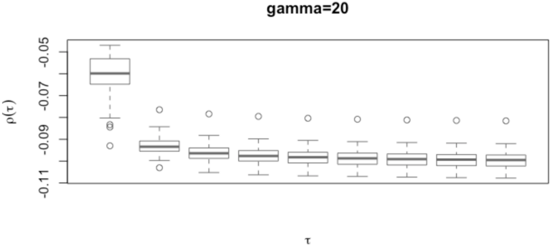
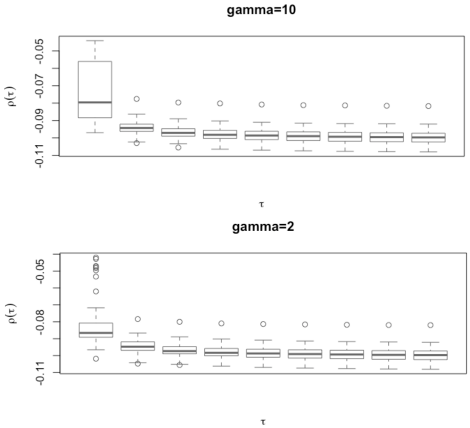


Figure: Variation over firms for $\gamma = 20$

Calibration of probability weighting functions



Message: ρ bigger for smaller τ

Calibration of probability weighting functions

Higher values of ρ_τ in the left tail indicate that large negative sentiment is connected to higher overweighting of the PK.

Empirical Results: Pricing tails with Sentiment

- We propose a dynamic quantile asset pricing model

$$\mathrm{Q}_{\tau}\left[\widetilde{M}_t \times (1 + R_{t+1}) + 1\right] = 0$$

- with $\widetilde{M}_t = \exp(-\alpha_\tau - \beta_{S,\tau} S_t - FF_t^\top \beta_{FF,\tau} - X_t^\top \beta_{X_t,\tau})$,

  FF=Fama French 5 factors

  X_t - control variables including idiosyncratic factors
- Factors are proxy for aggregate consumption

Empirical Results: Pricing tails with Sentiment

After log-linearization, we arrive to a simple linear model

$$\mathrm{Q}_\tau \left[\log(1+R_{t+1}) - \alpha_\tau - \beta_{S,\tau} S_t - FF_t^\top \beta_{FF,\tau} - X_t^\top \beta_{X_t,\tau} \right] = 0 \quad (17)$$

implying

$$\mathrm{Q}_\tau \left[\log(1+R_{t+1}) \right] = \alpha_\tau + \beta_{S,\tau} S_t + FF_t^\top \beta_{FF,\tau} + X_t^\top \beta_{X_t,\tau} \quad (18)$$

with *FF* Fama-French Factors

Empirical Results: Sentiment as factor

- Aggregate market sentiment as possible risk factor.
- Control also for firm-specific sentiment and volatility
- Negative sentiment captures "fear", related to VIX (Da et al., 2015)
- Following high investor sentiment, aggregate returns are low (Baker and Wurgler, 2007)
- Overly optimistic beliefs about future cash flows is not justified by fundamentals.

A dynamic quantile model with sentiment

Linear asset pricing model Fama-French Factors

$$Q_\tau(r_{i,t+1}) = \alpha_{i,\tau} + \beta_{1,\tau} B_{i,t} + \beta_{2,\tau}\sigma_{i,t} + \beta_{3,\tau}|BN_t| + FF_t^\top \beta_{FF,\tau} \quad (19)$$

with $\sigma_{i,t}$ Garman & Klass (1980) range-based volatility

$|BN_t|$ proxy for S_t (hence β_S from (18) is here β_3)
$B_{i,t}$ proxy for idiosyncratic sentiment
$\sigma_{i,t}$ proxy for volatility

$B_{i,t}$, $\sigma_{i,t}$ control variables, contained in the matrix X in (18).

Eq (20) tests if sentiment prices quantiles of the excess asset returns.

- Coefficients capture marginal effects of pricing factors
- Coefficients varying across τ imply marginal effect
- Coefficients constant over τ: EU works?

A dynamic quantile model with sentiment

Linear asset pricing model Fama-French Factors

$$Q_\tau(r_{i,t+1}) = \alpha_{i,\tau} + \beta_{1,\tau} B_{i,t} + \beta_{2,\tau}\sigma_{i,t} + \beta_{3,\tau}|BN_t| + FF_t^\top \beta_{FF,\tau} \quad (20)$$

with $\sigma_{i,t}$ - Garman & Klass (1980) range-based volatility.

(20) tests if sentiment prices quantiles of the excess asset returns.

- Coefficients capture marginal effects of pricing factors
- Coefficients varying across τ imply marginal effect
- Coefficients constant over τ: EU works?

Results

Estimate (20) via QR

- Panel of 100 most liquid constituents of S&P 500 stocks
- 10 main sectors Details
- Check sentiments across τ

Results: Panel of 100 stocks

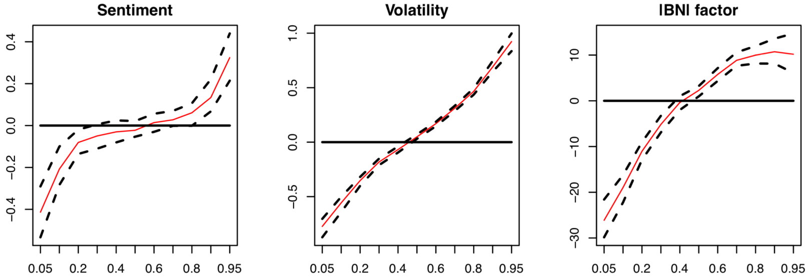


Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0,1)$

Full estimates of eq. (20) Further Graphics

Results: Panel of 100 stocks

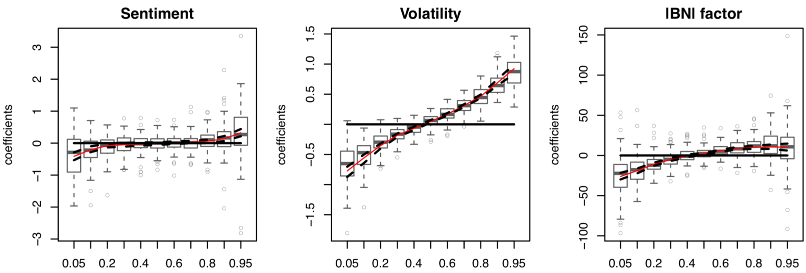


Figure: Estimates for $\beta_{i,\tau}$ together with box plots showing individual estimates with univariate individual I=1,...,100 QR estimates

Empirical Results

- Tails are strongly influenced
- Sentiment and volatility effects similarly
- $\beta_\tau \neq 0$ for most of the τs
- Asymmetric impact of market sentiment
- Holds even after control for firm specific sentiment
- Increase in negative bullishness has positive effect on right tail, and negative effect on left tail
- Contrary to literature, factors explain daily data in quantiles

Results: Sectors

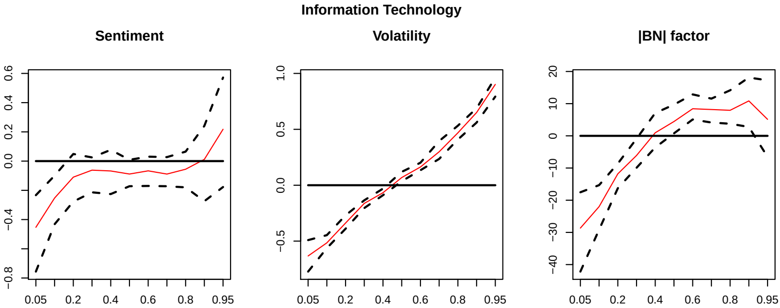


Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Further Graphics

Results: Sectors

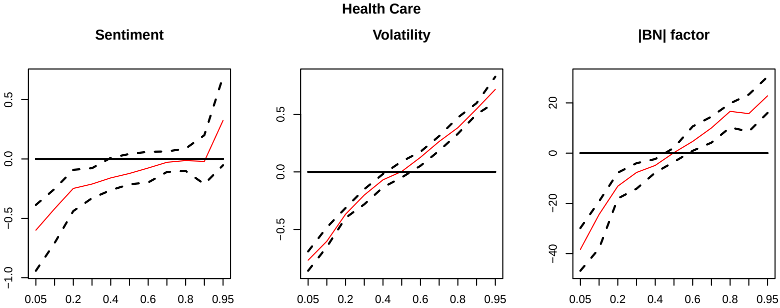


Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20) Further Graphics

Results: Sectors

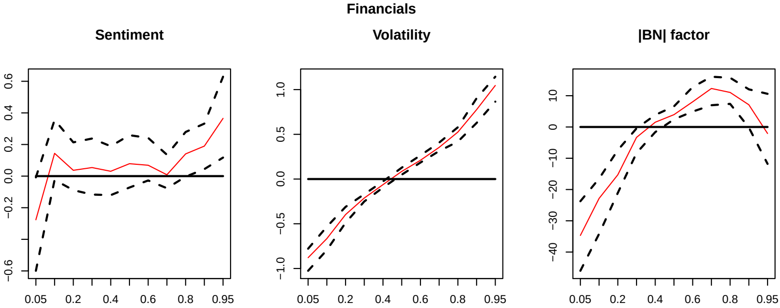


Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20) Further Graphics

Results: Sectors

Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0,1)$

Full estimates of eq. (20) Further Graphics

Results: Sectors

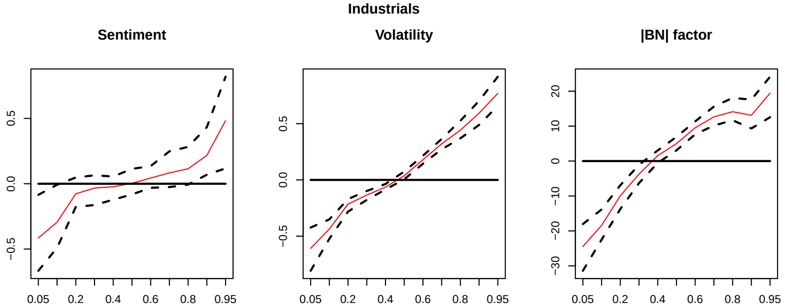


Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20) Further Graphics

Results: Sectors

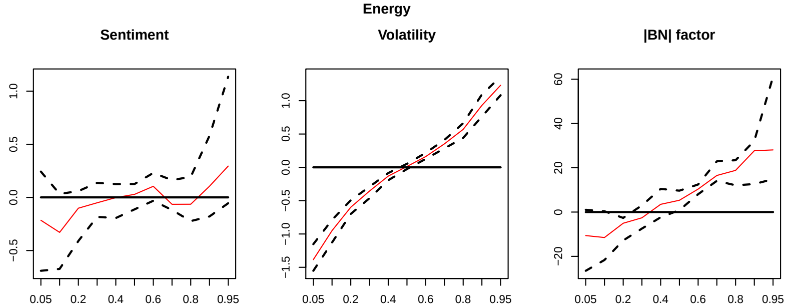


Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0,1)$

Full estimates of eq. (20) Further Graphics

Results: Sectors

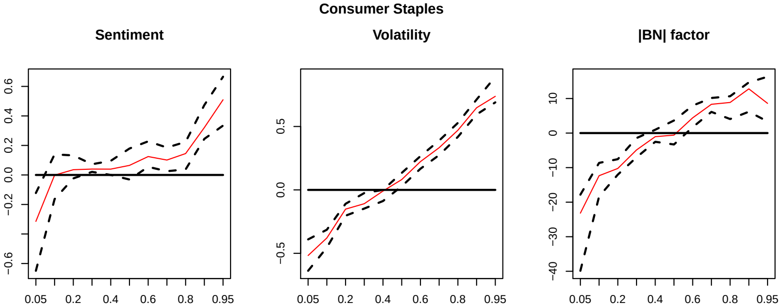


Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20) Further Graphics

Results: Sectors

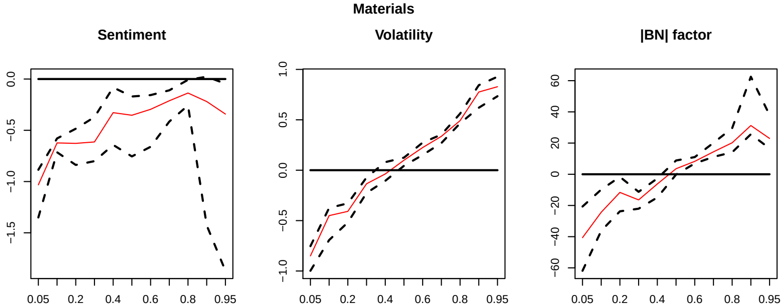


Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20) Further Graphics

Results: Sectors

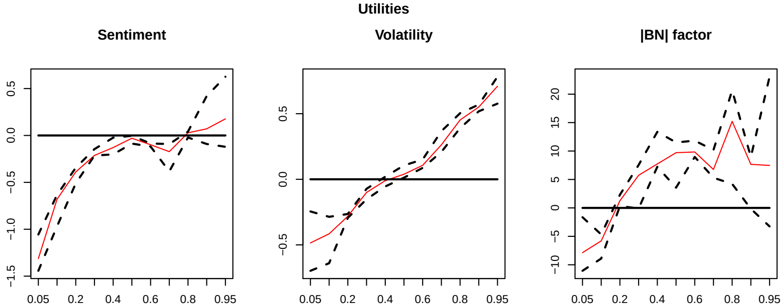


Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0,1)$

Full estimates of eq. (20) Further Graphics

Results: Sectors

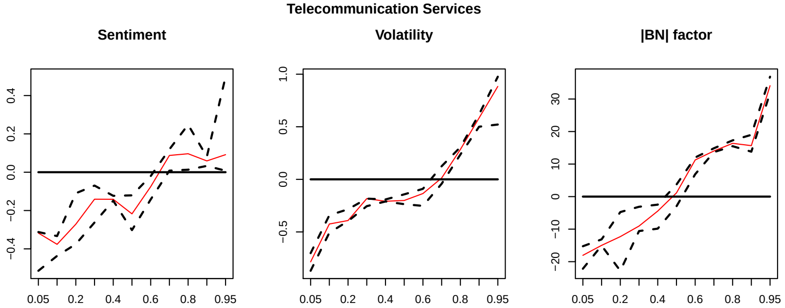


Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0,1)$

Full estimates of eq. (20) Further Graphics

Summary

- Tales of sentiment driven tails
- Dynamic quantile model for asset pricing with sentiment
- Investor sentiment distilled from public news with cross-section of future return's quantiles.

Tales of sentiment driven tails

Jozef Baruník
Cathy Yi-Hsuan Chen
Wolfgang Karl Härdle

Institute of Economic Studies
Charles University in Prague
Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin

http://ies.fsv.cuni.cz
http://lvb.wiwi.hu-berlin.de

Bibliography

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

Baker, M., and J. Wurgler.
Investor sentiment and the cross-section of stock returns
Journal of Finance, 2006

de Castro, L. I. and A. F. Galvao
Dynamic quantile models of rational behavior
2017

Da, Z., Engelberg J. and Gao, P.
The Sum of All FEARS Investor Sentiment and Asset Prices
Review of Financial Studies, 2015

De Long, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann.
Noise trader risk in financial markets.
Journal of Political Economy, 1990

Dumas, B., Kurshev, A., Uppal, R.
Equilibrium Portfolio Strategies in the Presence of Sentiment Risk and Excess Volatility
Journal of Finance, 2009

Fama, E. and K. French.
A Five-Factor Asset Pricing Model
J. Financial Econom., 2015

Huang, D., Jun Tu, J., Jiang, F., and Zhou, G.
Investor Sentiment Aligned: A Powerful Predictor of Stock Returns
Journal of Finance, 2014

Keynes, J. M.
The general theory of employment, interest and money.
London: Macmillan, 1936

Koenker, R.
Quantile regression for longitudinal data.
Journal of Multivariate Analysis, 2004

Manski, C.F.
Ordinal utility models of decision making under uncertainty.
Theory and Decision, 1988

Polkovnichenko, V. and Zhao, F.
Probability weighting functions implied in options prices
Journal of Financial Economics, 2013

Prelec, D.
The probability weighting function
Econometrica, 1998

Rostek, M.
Quantile maximization in decision theory.
The Review of Economic Studies, 2010

Stambaugh, R.F., Yu, J.F., Yuan, Y.
The short of it: Investor sentiment and anomalies.
Journal of Financial Economics, 2012

Tversky, A. and Kahneman, D.
Advances in prospect theory: Cumulative representation of uncertainty
Journal of Risk and uncertainty, 1992

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

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

Back

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

$$L\{y, s(X)\} = \log(2)^{-1} \log[1 + \exp\{-s(X)y\}] \tag{21}$$

- Hinge: Support Vector Machines

$$L\{y, s(X)\} = \max\{0,\, 1 - s(X)y\} \tag{22}$$

Back

Regularization Term

- L2 norm

$$R(\beta) = 2^{-1} \sum_{i=1}^{p} \beta_i^2 \tag{23}$$

- L1 norm

$$R(\beta) = \sum_{i=1}^{p} |\beta_i| \tag{24}$$

Back

RLM Example

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

$$y = (1, -1) \quad (25)$$

$$X = \begin{array}{c} \\ \textit{the} \\ \textit{profit} \\ \textit{of} \\ \textit{Apple} \\ \textit{increased} \\ \textit{company} \\ \textit{decreased} \end{array} \begin{array}{c} \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array} \quad (26)$$

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

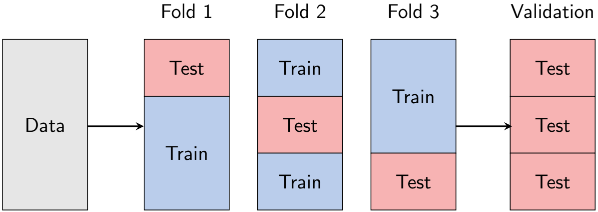


Figure: 3-fold Cross Validation

Back

Oversampling

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

$$L(\theta) = \sum_{i=1}^{n} L_i(\theta) \tag{27}$$

- Iteratively update

$$\theta_i = \theta_{i-1} - \eta \, \frac{\partial L_i(\theta)}{\partial \theta} \tag{28}$$

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 $x_1, \ldots, x_n$ 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

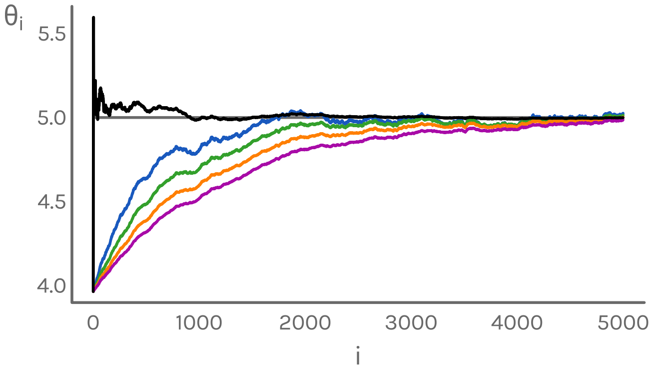


Figure: Estimate Mean via SGD, $x_t \sim \mathrm{N}(5, 1)$

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

Back

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: Confusion Matrix - Supervised Learning with Oversampling

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: Sector Abbreviations

back

Fama-French 5 factors

$FF1$ - the Mkt factor: excess return on the market index

$FF2$ - the SMB factor: (Small Minus Big) the average return on the nine small-stock portfolios minus that on the nine big-stock portfolios.

$FF3$ - the HML factor: (High Minus Low) the average return on the two value-stock portfolios minus that on the two growth-stock portfolios

Back

Fama-French 5 factors cont.

$FF4$ - the RMW factor: (Robust Minus Weak) the average return on the two robust operating profitability portfolios minus that on the two weak operating profitability portfolios

$FF5$ - the CMA factor: (Conservative Minus Aggressive) the average return on the two conservative investment portfolios minus that on the two aggressive investment portfolios

Back

Garman & Klass range-based volatility

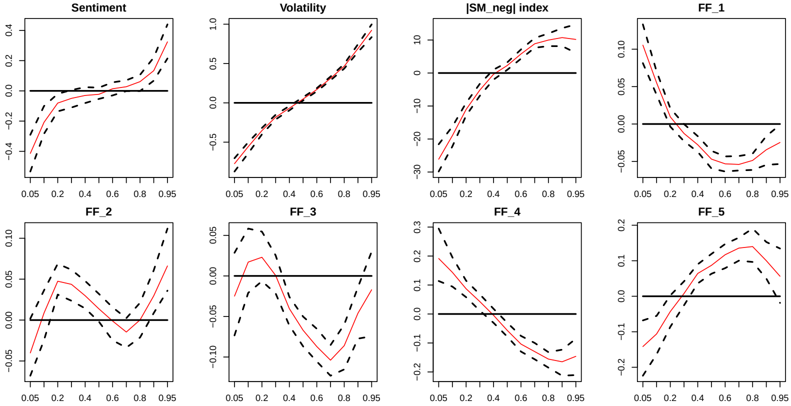
$$\sigma_{i,t} = 0.511(u-d)^2 - 0.019\{c(u+d) - 2ud\} - 0.838c^2 \quad (29)$$

$$\begin{aligned} \text{with } u &= \log(P^H_{i,t}) - \log(P^L_{i,t}) \\ d &= \log(P^L_{i,t}) - \log(P^O_{i,t}) \\ c &= \log(P^C_{i,t}) - \log(P^O_{i,t}), \end{aligned}$$

where the $P^H_{i,t}, P^L_{i,t}, P^O_{i,t}, P^C_{i,t}$ are the daily highest, lowest, opening and closing stock prices.

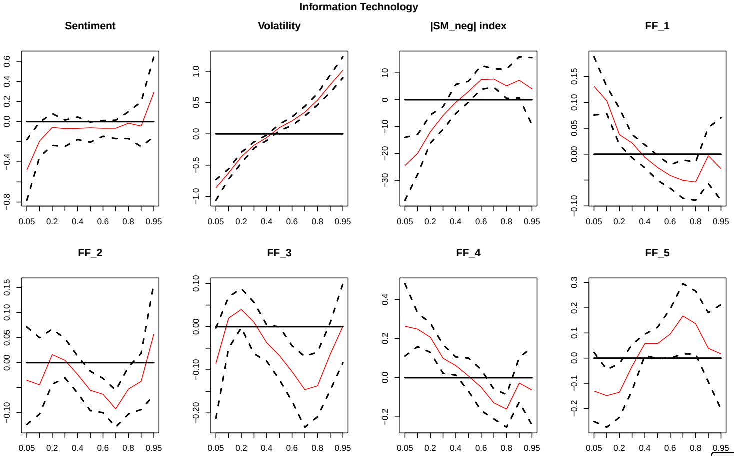
Back

Results: Panel of 100 stocks



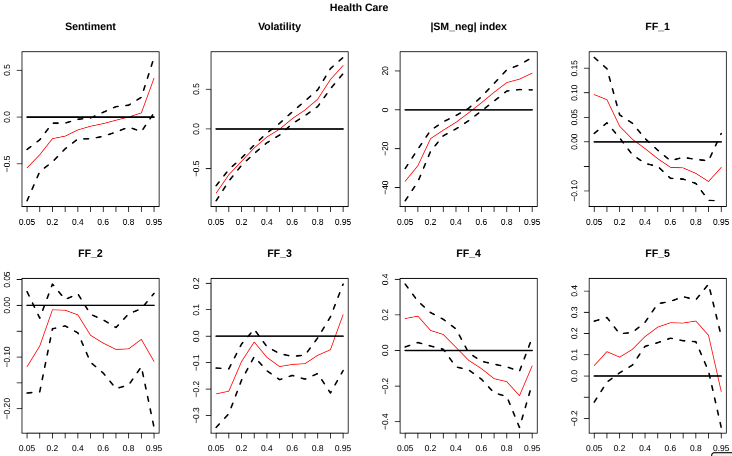
Back

Results: Sectors



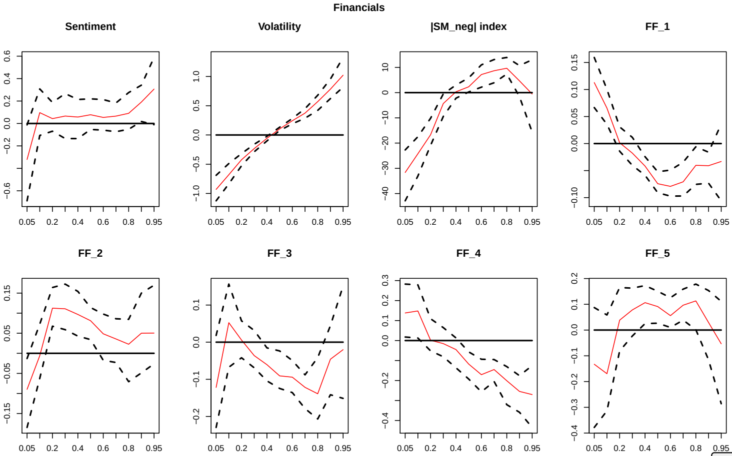
Back

Results: Sectors



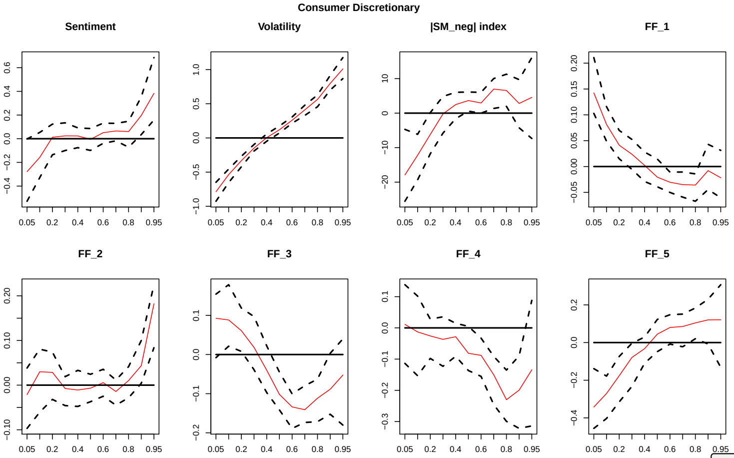
Back

Results: Sectors



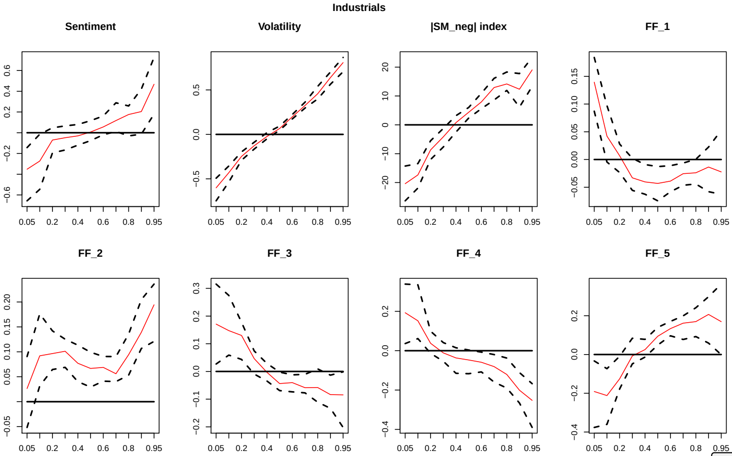
Back

Results: Sectors



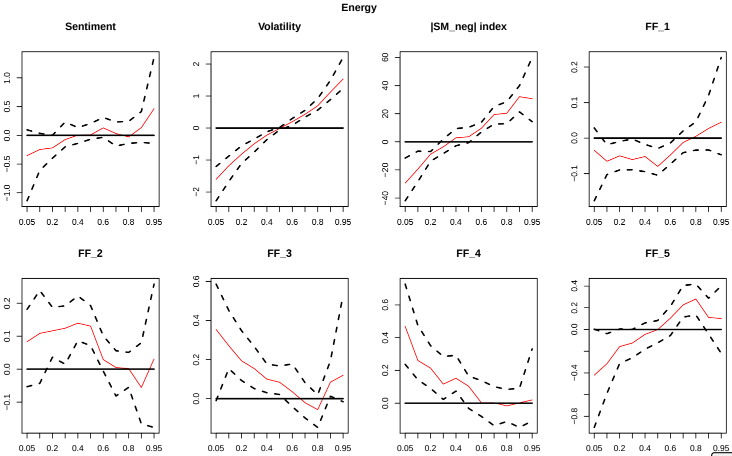
Back

Results: Sectors



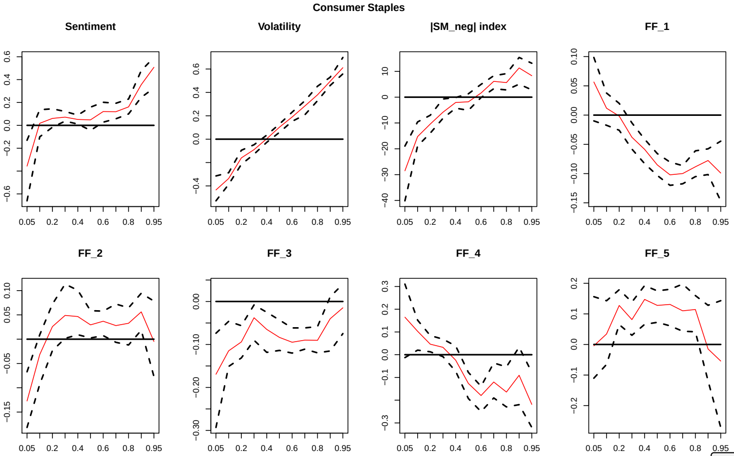
Back

Results: Sectors



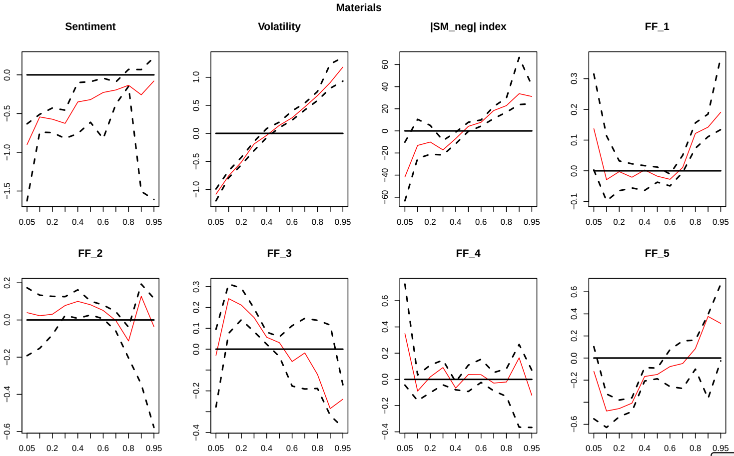
Back

Results: Sectors



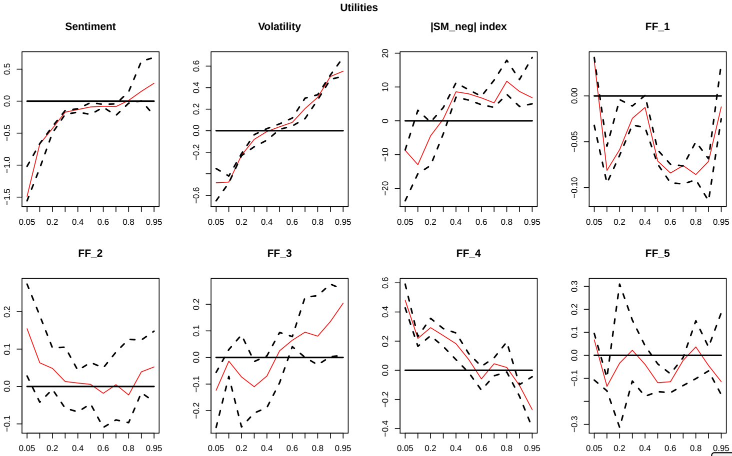
Back

Results: Sectors



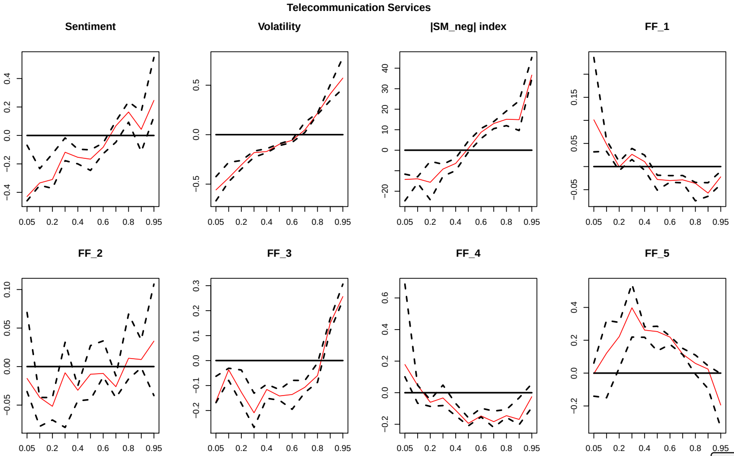
Back

Results: Sectors



Back

Results: Sectors



Back