Questioning The News about Economic Growth: Sparse Forecasting Using Thousands of News-based Sentiment Values

Fan Jia

Central University of Finance and Economics

December 7, 2019

Contents

Introduction

1. The importance of forecasting; the role of the quantitative information and qualitative Information
2. In practice, quantitative information is the dominant type of data used in prediction.
3. Innovation: complements the quantitative information with sentiment values of news text
4. An overall introduction of the methodology
5. Advantages of the proposed methodology
6. The introduction of contributions and approaches for dealing with high dimensionality of data
7. The overall introduction of the experiment
8. Extra work: R package released
9. The structure of this paper

Methodology

1. Data preparation
2. Aggregating sentiment into a prediction
3. Forecast precision and attribution

Methodology-Data preparation

The variable being predicted is the h-period logarithmic change in the variable Y_t,

$$y_t^h \equiv 100 \times (\ln Y_{t+h} - \ln Y_t)$$

where $t = 1, 2, ..., T$ is a time index; and y_t^h to be covariance stationary.
y_t^h: the monthly logarithmic growth in industrial production of the US.
$E(y_T^h | I_T)$: the expected value of y_T^h given the information at time T

Methodology-Aggregating sentiment into a prediction

- Step 1: Classify texts(N_t) by topic then use expert opinion to choose a subset of topics

  Topic-makers: Provided by the publishers or Extracted from the texts (LDA, keywordsbased identification, SVM, etc.)
- Step 2: Compute the sentiment for each text($s_{n,t,l}$) n of corpus t using L methods.(Bag-of-words)

$$\mathbf{S}_{t,l} \equiv (s_{1,t,l}, \ldots, s_{N_t,t,l})'$$

- Step 3: Obtain K topic-based sentiments for each corpus n and method l.

$$W_t s_{t,l} = W_{(K \times N_t)} s_{(N_t \times 1)}$$

- Step 4: Obtain time series aggregated values for each topic k and method l.

$$\mathbf{V}_{t,l(K\times(\tau+1))} \equiv \begin{bmatrix} | & & | \\ \mathbf{W}_{t-\tau}\mathbf{s}_{t-\tau,l} & \cdots & \mathbf{W}_t\mathbf{s}_{t,l} \\ | & & | \end{bmatrix}$$

Methodology-Aggregating sentiment into a prediction

- Step 4: Obtain time series aggregated values for each topic k and method l.

$$\mathbf{V}_{t,l(K\times(\tau+1))} \equiv \begin{bmatrix} | & & | \\ \mathbf{W}_{t-\tau}\mathbf{s}_{t-\tau,l} & \cdots & \mathbf{W}_{t}\mathbf{s}_{t,l} \\ | & & | \end{bmatrix}$$

$$\mathbf{v}_{t(LK\times(\tau+1))} \equiv \begin{bmatrix} \mathbf{v}_{t,1} \\ \vdots \\ \mathbf{v}_{t,L} \end{bmatrix}$$

$$\mathbf{s}_{t(LKB\times 1)} \equiv \mathrm{vec}\left(\mathbf{V}_t\mathbf{B}\right)$$

where $\mathrm{vec}\,(\cdot)$ is the vectorization operator (stacks the columns of a matrix into a vector one on top of another).

Methodology-Aggregating sentiment into a prediction

- Beta weighting requires two parameters $a > 0$ and $b > 0$.

$$c(i; a, b) \equiv \frac{f\left(\frac{i}{\tau}; a, b\right)}{\sum_{i=1}^{\tau} f\left(\frac{i}{\tau}; a, b\right)}$$

where $f(x; a, b) \equiv \dfrac{x^{a-1}(1-x)^{b-1}\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ is the Beta density

Given a grid $\{a_i, b_i\}_{i=1}^{B}$, the $(\tau + 1) \times B$ aggregation matrix is given by:

$$\mathbf{B} \equiv \begin{bmatrix} c(1; a_1, b_1) & \cdots & c(1; a_B, b_B) \\ \vdots & & \vdots \\ c(\frac{i}{\tau}; a_1, b_1) & & c(\frac{i}{\tau}; a_B, b_B) \\ \vdots & \cdots & \vdots \\ c(0; a_1, b_1) & & c(0; a_B, b_B) \end{bmatrix}$$

Methodology-Aggregating sentiment into a prediction

- Step 5: Calibrate to optimize the forecast precision.

$$y_t^h = \alpha + \gamma' \mathbf{x}_t + \beta' \mathbf{s}_t + \varepsilon_t, \quad t = 1, \ldots, T$$

where x_t is a $M \times 1$ vector of (nontextual sentiment) variables available at time t, $\beta \equiv (\beta_1, \ldots, \beta_P)'$ is a vector of parameters associated with the P textual sentiment indices (P = LKB).
Typically, x_t includes y_s, where y_s is the dependent variable up to time t, that is $s \leq t$. In practice, in economics we often have $s < t$ due to the release lag faced by economic indicators.
A penalized least squares criterion to estimate the regression (Elastic Net Regularization)

Methodology-Aggregating sentiment into a prediction

- Step 5: Calibrate to optimize the forecast precision.

$$y_t^h = \alpha + \gamma' \mathbf{x}_t + \beta' \mathbf{s}_t + \varepsilon_t, \quad t = 1, \ldots, T$$

For ease of presentation, $\mathbf{z}_t \equiv (\mathbf{x}_t', \mathbf{s}_t')'$ and $\boldsymbol{\theta} \equiv \left(\gamma', \boldsymbol{\beta}'\right)'$ both of size $(M+P) \times 1$.

$$\min_{\tilde{\alpha},\hat{\theta}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \left[y_t^h - \left(\tilde{\alpha} + \tilde{\theta} \tilde{\mathbf{z}}_t \right) \right]^2 + \lambda_1 \left[\lambda_2 |\tilde{\theta}\,\|_1 + (1-\lambda_2) \| \tilde{\theta} \|_2^2 \right] \right\}$$

where $\lambda_1 \geq 0$ is the parameter that sets the level of regularization and $0 \leq \lambda_2 \leq 1$ is the weight between the two types of penalties.(when $\lambda_1 = 0$, Lasso; when $\lambda_2 = 0$, Ridge regularization).
The variable $\widetilde{\mathbf{z}}_t$ is the standardized version of $\mathbf{Z}_t$
Follow Zou, Hastie, and Tibshirani (2007) and minimize the BIC-like criterion, where BIC stands for Bayesian information criterion.

Methodology-Forecast precision and attribution

- Step 6: Forecasting.
- Step 7: Forecast precision evaluation.

the root mean squared forecast error (RMSFE) and the mean absolute forecast error (MAFE).

$$\mathrm{RMSFE}_i^h \equiv \sqrt{\frac{1}{T_F}\sum_{t=T+1}^{T+T_F}\left(e_{i,t}^h\right)^2}, \quad \mathrm{MAFE}_i^h \equiv \frac{1}{T_F}\sum_{t=T+1}^{T+T_F}\left|e_{i,t}^h\right|$$

where T is the size of the estimation sample and T_F is the number of out-of-sample observations.

Methodology-Forecast precision and attribution

- Step 8: Attribution.

$$\widehat{y}_T^h = \widehat{\alpha} + \widehat{\gamma}' \mathbf{X}_T + \sum_{t=(T-\tau)}^{T} \sum_{n=1}^{N_t} \sum_{k=1}^{K} \sum_{k=1}^{K} \sum_{b=1}^{B} \widehat{\beta}' \boldsymbol{e}_{l,k,b} \cdot W_{t,k,n} B_{T-t,b} \cdot s_{n,t,l}$$

the weight $\omega_{n,t,l}$ that is attributed to the sentiment $s_{n,t,l}$ is equal to:

$$\omega_{n,t,l} = \sum_{k=1}^{K} \sum_{b=1}^{B} \widehat{\beta}' e_{l,k,b} \cdot W_{t,k,n} B_{T-t,b}$$

$$\widehat{y}_T^h = \widehat{\alpha} + \widehat{\gamma}' \mathbf{x}_T + \sum_{t=(T-\tau)}^{T} \sum_{n=1}^{N_t} \sum_{l=1}^{L} \omega_{n,t,l} \cdot s_{n,t,l}$$

Considering feasibility, obtain the attribution of topic $g(1 \leq g \leq K)$ by fixing k=g

$$a_g \equiv \sum^{T} \sum^{N_t} \sum^{L} \sum^{B} \widehat{\beta}' e_{l,g,b} \cdot W_{t,g,n} B_{T-t,b} \cdot s_{n,t,l}$$

Application-Data and descriptive statistics

- Quantitative data: the industrial production.

$$y_t^h \equiv 100 \times (\ln IP_{t+h} - \ln IP_t)$$

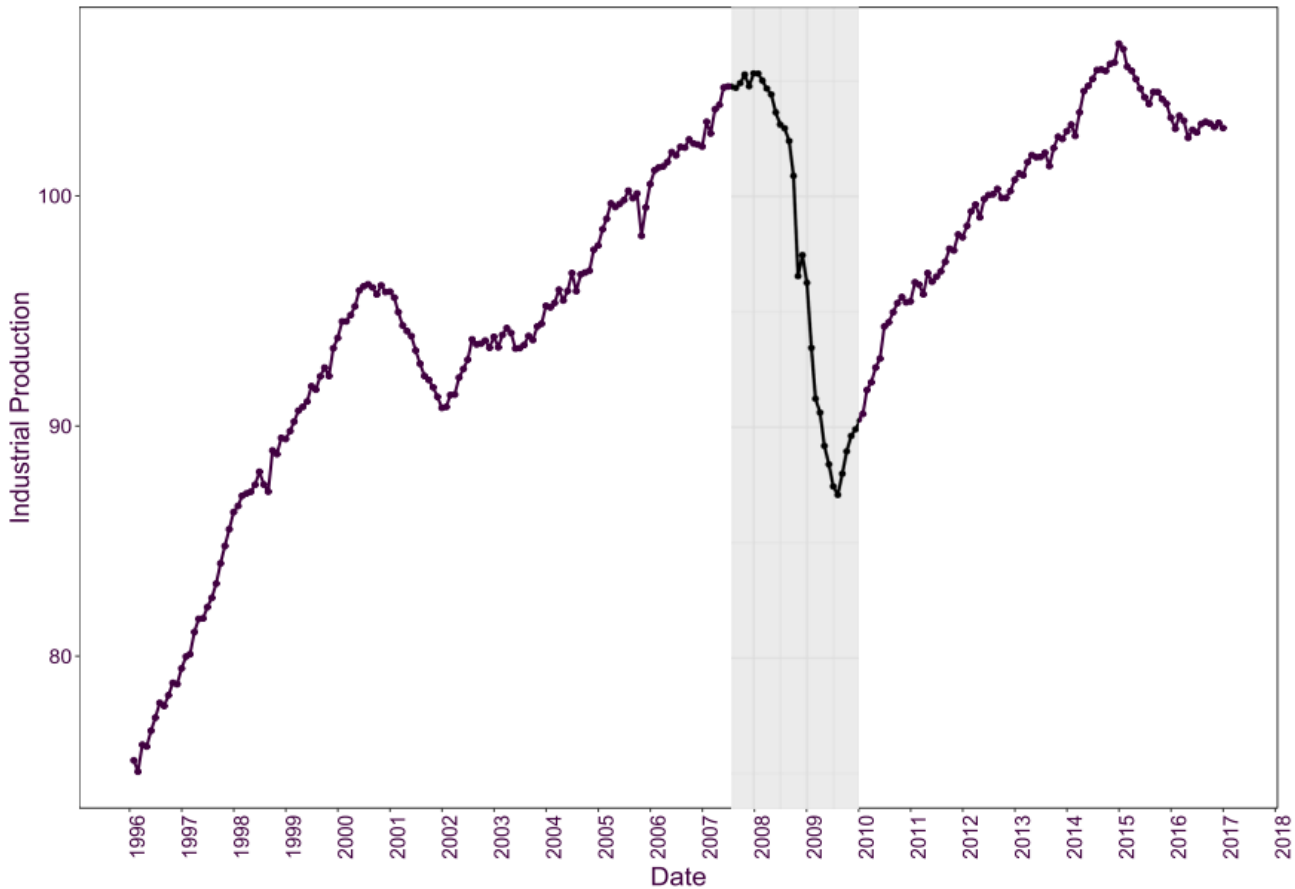


Fig. 2. US industrial production. Notes: The figure presents the US industrial production from January 1996 to December 2016 (192 monthly observations). The gray zone indicates the crisis period, which spans the period July 2007 to December 2009 (30 months).

Application-Data and descriptive statistics

- Quantitative data
  - Macroeconomic Variables: 105 to 128 variables from the FRED-MD historical vintage databases for every month from August 1999 to December 2016.
  - Financial Indicators: 16 financial metrics such as dividend ratios, long/short term yields, stock variances, etc. Goyal and Welch (2008)
  - Survey Data: Chicago Board of Exchange's forward-looking volatility index (VIX); The list of variables the media-attention EPU index; six survey-based Conference Board indices (CB).

Application-Data and descriptive statistics

- Qualitative data
  - Corpus
    All English articles from "Major US Newspapers" in the LexisNexis database with reference to the US from January 1, 1994, to December 31, 2016.(the topic filter and at least 200 words), including 338,408 articles and 44 topics over six clusters
  - Sentiment calculation
    Specialized financial dictionaries for the analysis of financial and economic discourses(7 lexicons)

$$s_{n,t,l} \equiv \frac{N^{+}_{n,t,l} - N^{-}_{n,t,l}}{N^{+}_{n,t,l} + N^{-}_{n,t,l} + N^{0}_{n,t,l}}$$

Application-Data and descriptive statistics

- Qualitative data
  - Sentiment calculation

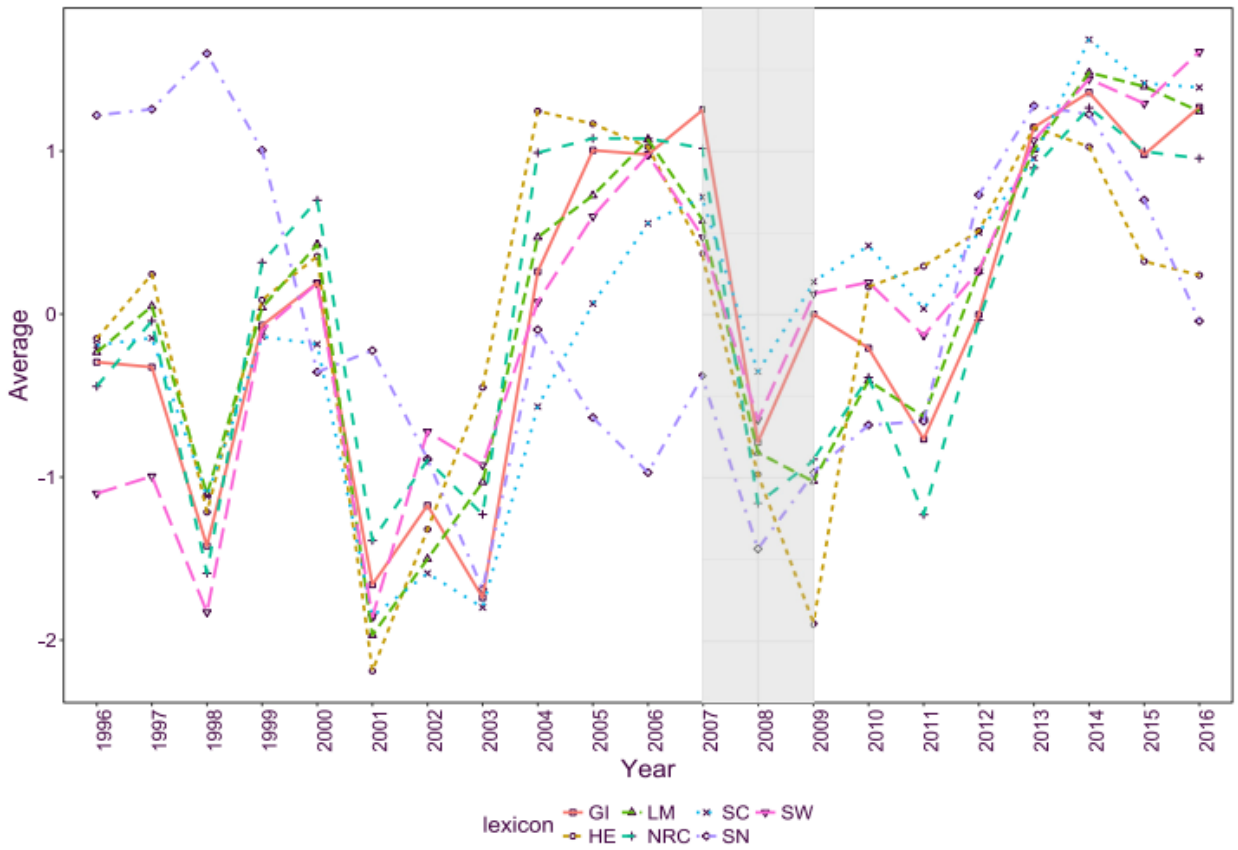


Fig. 3. Yearly lexicon-based averages of the individual news articles' sentiments. Notes: The figure presents the seven lexicon-based yearly averages of the individual news articles' sentiment for the period from 1994 to 2016. Sentiment values are standardized for readability purposes. The gray zone indicates the 2007–2009 crisis period.

Application-Data and descriptive statistics

- Aggregation of sentiment

$$\tau = 180 \text{ and } P = LKB = 7 \times 44 \times 16 = 4928$$

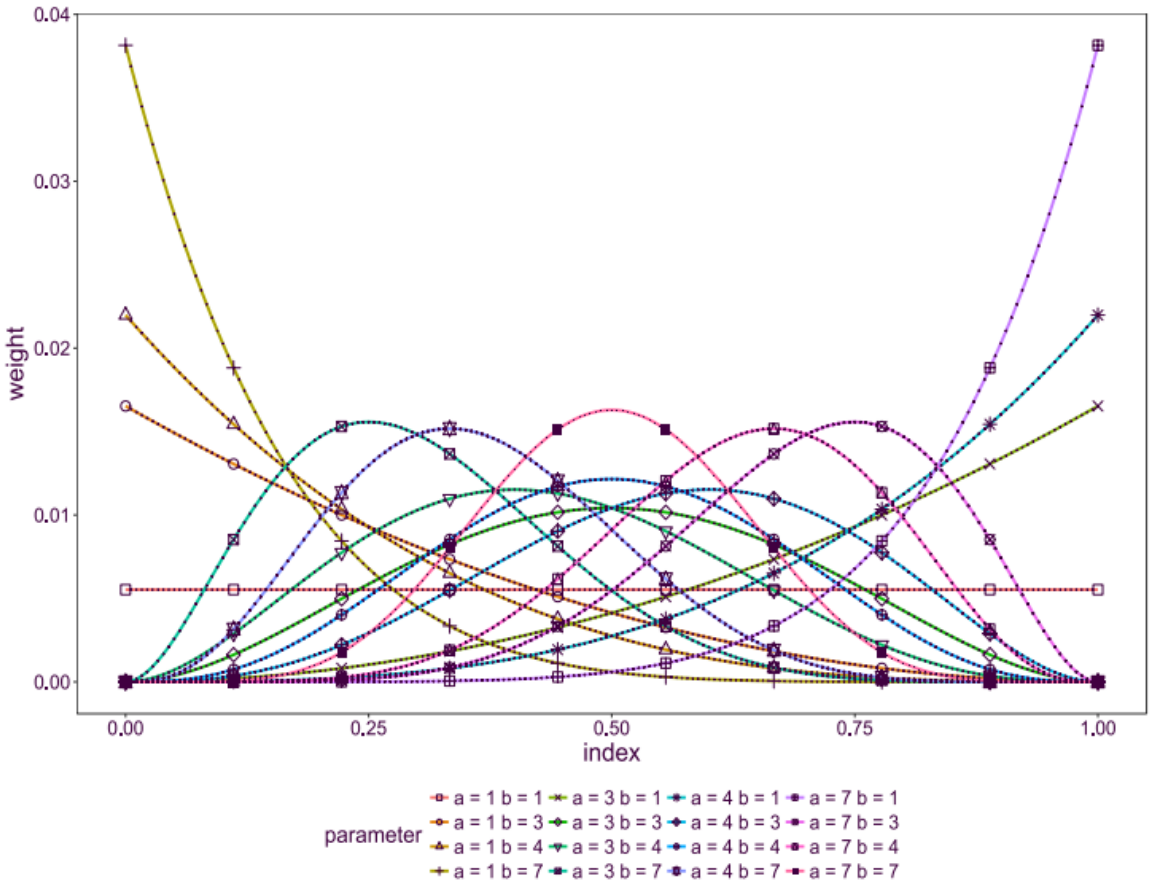
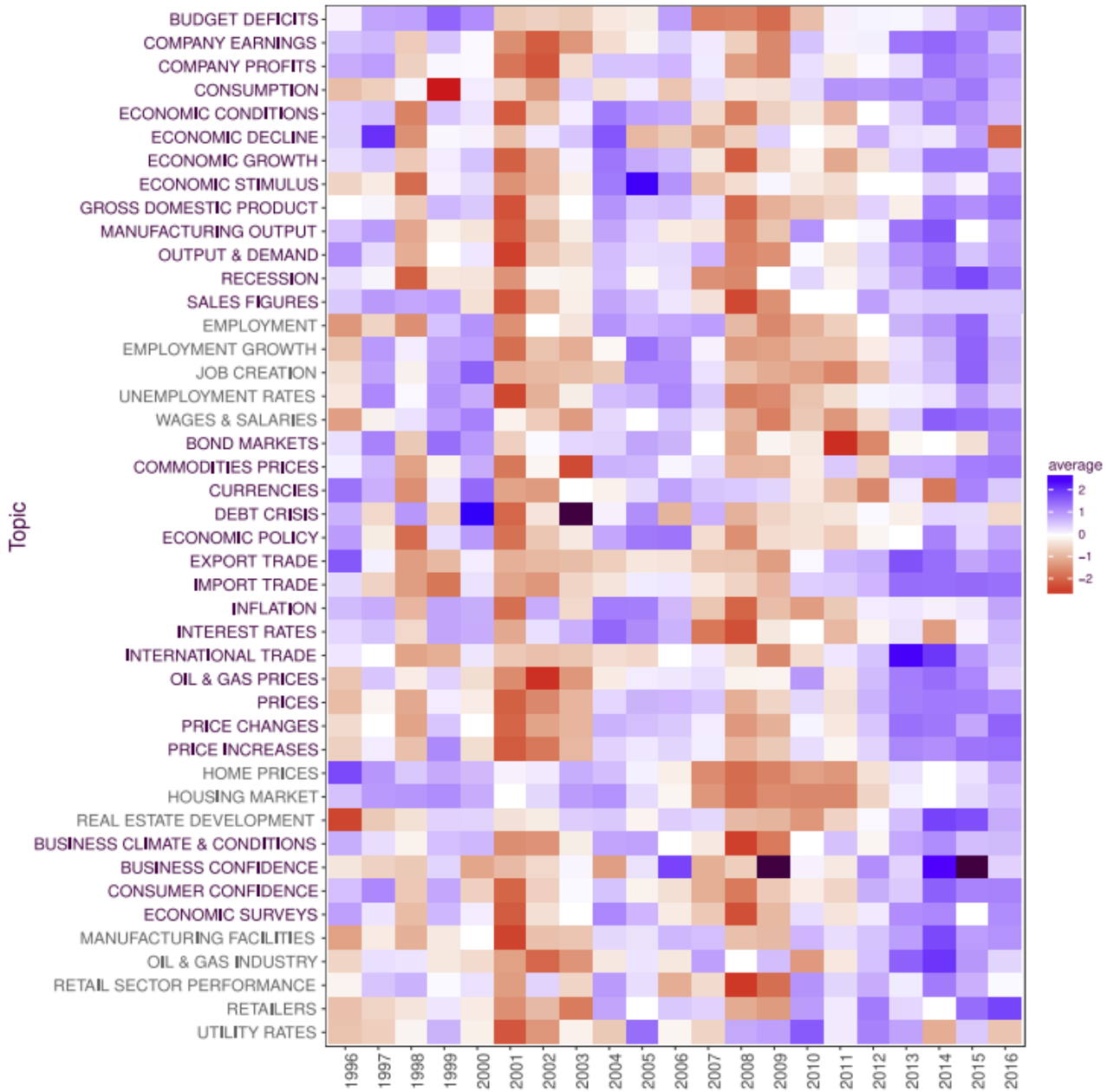


Fig. 4. Beta weights. Notes: The figure presents the time-aggregation weights of the Beta function for the grid $\{1, 3, 4, 7\} \times \{1, 3, 4, 7\}$ for a total of 16 weighting schemes.

Application-Main results



Application-Main results

- Aggregation of sentiment

$$
\begin{aligned}
\mathcal{M}_{1a}: \quad & y_t^h = \alpha_0^h + \alpha_1^h y_{t-h}^h + (\boldsymbol{\gamma}^h)' \mathbf{x}_t + \varepsilon_t^h \\
\mathcal{M}_{1b}: \quad & y_t^h = \alpha_0^h + \alpha_1^h y_{t-h}^h + (\boldsymbol{\gamma}^h)' \mathbf{x}_t + (\boldsymbol{\beta}^h)' \mathbf{s}_t + \varepsilon_t^h \\
& \text{and:} \\
\mathcal{M}_{2a}: \quad & y_t^h = \alpha_0^h + \alpha_1^h y_{t-h}^h + (\boldsymbol{\gamma}^h)' \mathbf{f}_t + \varepsilon_t^h \\
\mathcal{M}_{2b}: \quad & y_t^h = \alpha_0^h + \alpha_1^h y_{t-h}^h + (\boldsymbol{\gamma}^h)' \mathbf{f}_t + (\boldsymbol{\beta}^h)' \mathbf{s}_t + \varepsilon_t^h
\end{aligned}
$$

where f_t are factors extracted from x_t using the IC_{p1} criterion. All models are estimated using the elastic net.

Application-Main results

- Model's forecasting performance comparison

Table 2
Forecasting results.

		RMSFE				MAFE			
Period	h	$\mathcal{M}_{1a}$	$\mathcal{M}_{1b}$	$\mathcal{M}_{2a}$	$\mathcal{M}_{2b}$	$\mathcal{M}_{1a}$	$\mathcal{M}_{1b}$	$\mathcal{M}_{2a}$	$\mathcal{M}_{2b}$
Full sample	1	0.68	0.70	0.64	0.70	0.49	0.49	0.45	0.49
	3	1.52	1.54	1.59	1.52	0.96	1.01	1.02	1.01
	6	4.86	3.93	5.01	3.14	2.36	2.35	2.85	2.14
	9	7.01	4.95	8.36	4.58	3.71	3.28	4.89	3.19
	12	6.39	5.19	8.69	5.14	4.25	3.41	6.03	3.32
Pre-crisis	1	0.55	0.57	0.56	0.56	0.43	0.42	0.43	0.44
	3	0.99	0.93	1.21	0.93	0.72	0.70	0.87	0.70
	6	1.67	1.65	2.62	1.62	1.31	1.36	1.80	1.32
	9	2.41	2.42	4.67	2.53	1.96	1.93	3.00	1.98
	12	3.27	2.00	6.07	1.90	2.72	1.67	3.73	1.57
Crisis	1	1.19	1.27	1.08	1.27	0.81	0.87	0.69	0.88
	3	3.20	3.19	3.17	3.04	2.46	2.52	2.31	2.29
	6	11.30	8.54	10.64	6.20	7.63	6.44	7.45	4.99
	9	8.58	7.94	9.92	7.94	6.67	6.20	7.67	6.20
	12	10.43	10.14	9.42	10.12	8.34	7.84	7.49	7.70
Post-crisis	1	0.53	0.50	0.49	0.50	0.42	0.40	0.40	0.41
	3	0.78	0.93	0.89	1.03	0.62	0.74	0.70	0.82
	6	1.72	2.26	2.86	2.32	1.32	1.68	2.05	1.77
	9	8.47	4.93	9.72	4.07	3.93	3.22	5.27	3.00
	12	6.02	3.85	9.81	3.78	3.80	2.98	7.01	2.90

Application-Main results

- Attribution

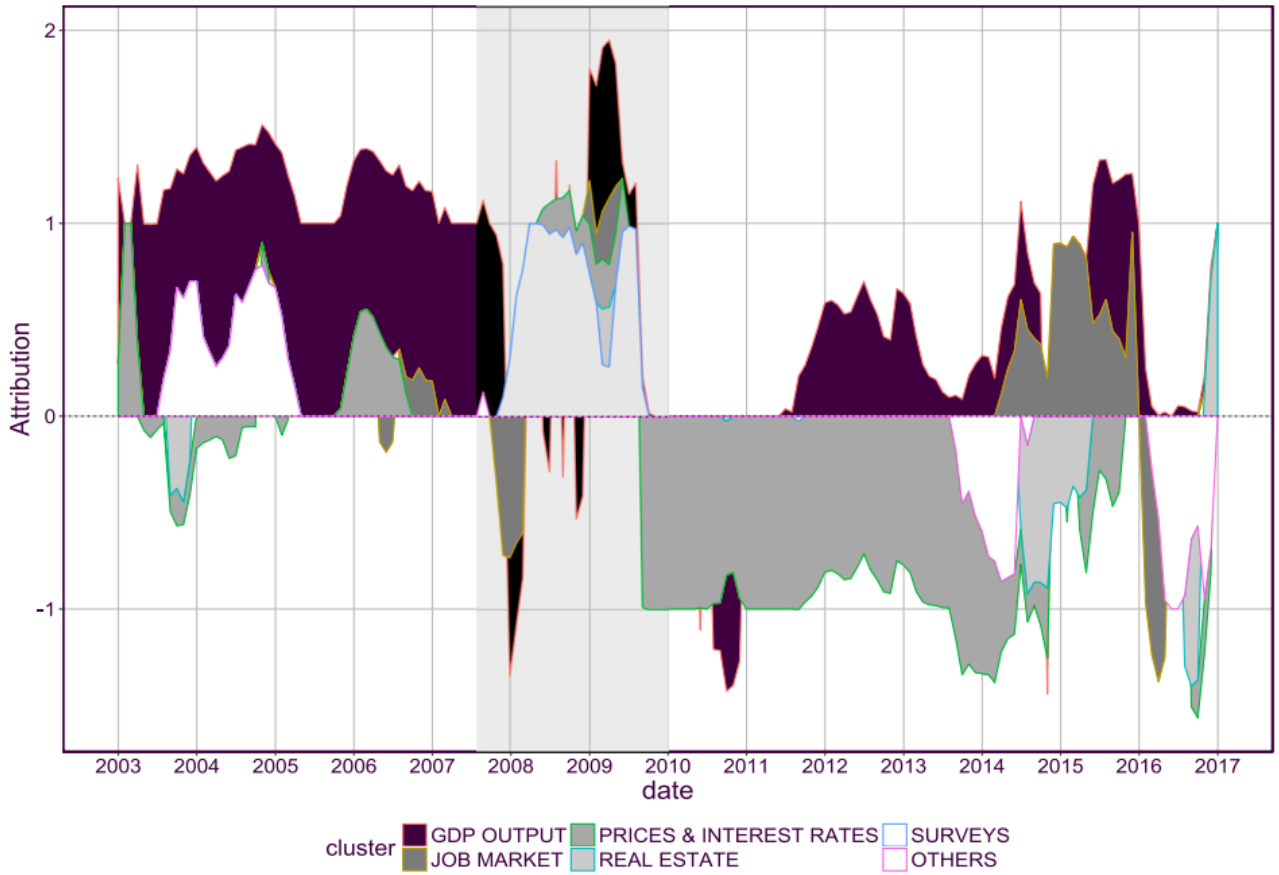


Fig. 6. Forecast attribution. Notes: The figure presents the cluster attribution of model $\mathcal{M}_{1b}$ for the out-of-sample forecasts of the twelve-month US industrial production log-growth. The period ranges from January 2003 to December 2016 (180 monthly observations). The attribution vector for a given date is scaled by dividing each element of the attribution vector by the L^2-norm of the attribution vector for that date. The gray zone indicates the July 2007 to December 2009 crisis period. A positive (negative) value indicates that the topic contributes positively (negatively) to the forecast, and therefore increases (decreases) the forecast of the US industrial production log-growth.

Application-Main results

- Importance of the optimization of each dimension

Table 3
Robustness results: aggregation of dimensions.

		RMSFE					MAFE				
	h	$\mathcal{M}$	LEX	TOPIC	TIME	ALL	$\mathcal{M}$	LEX	TOPIC	TIME	ALL
$\mathcal{M}_{1b}$	1	0.70	0.69	0.68	0.68	0.64	0.49	0.48	0.48	0.49	0.46
	3	1.54	1.50	1.41	1.52	1.58	1.01	0.98	0.93	0.96	0.99
	6	3.93	4.51	4.52	4.86	5.24	2.35	2.42	2.32	2.36	2.55
	9	4.95	5.91	5.57	7.01	8.37	3.28	3.43	3.28	3.71	4.17
	12	5.19	5.85	6.11	6.39	8.24	3.41	4.01	4.09	4.25	5.02
$\mathcal{M}_{2b}$	1	0.70	0.69	0.68	0.65	0.68	0.49	0.49	0.48	0.46	0.48
	3	1.52	1.53	1.50	1.39	1.31	1.01	1.06	1.07	0.95	0.92
	6	3.14	3.72	3.23	3.62	3.34	2.14	2.39	2.17	2.25	2.20
	9	4.58	5.65	5.36	6.99	6.23	3.19	3.74	3.42	4.16	4.06
	12	5.14	6.79	7.14	8.18	7.82	3.32	4.81	5.04	5.30	5.34

Notes: The table presents the forecasting results when the various dimensions (lexicon, topic, and time) are aggregated. We compare the results of the extended models $\mathcal{M}_{1b}$ and $\mathcal{M}_{2b}$ with those of four alternative approaches in which we (with equal weights) aggregate: (i) the lexicon-dimension (denoted LEX), (ii) the topic-dimension (denoted TOPIC), (iii) the time-dimension (denoted TIME), and (iv) all dimensions (denoted ALL). A light (dark) gray cell indicates that the extended model ($\mathcal{M}_{1b}$ or $\mathcal{M}_{2b}$) is superior (inferior) at the 5% significance level according to the Diebold and Mariano (1995) test statistic. See Table 2 for details.

Conclusion

- News releases may have predictive value for the future economic activity, and the various methods of calculating sentiment
- A framework that optimizes sentiment aggregation for the prediction of economic growth
- Test the predictive power of text-based sentiment indices by forecasting the growth in US industrial production
- The proposed optimized text-based sentiment analysis can improve the forecasting performance for predicting the nine-month and annual growth rates significantly