Expectations, disagreement and news

Philipp Adämmer[1], Joscha Beckmann[2], Rainer Schüssler[3]

[1] Helmut Schmidt University Hamburg, Germany
[2] Ruhr-University Bochum, Germany
[3] University of Rostock, Germany

LATINR
Conferencia Latinoamericana sobre el uso de R en Investigación + Desarrollo

Abstract

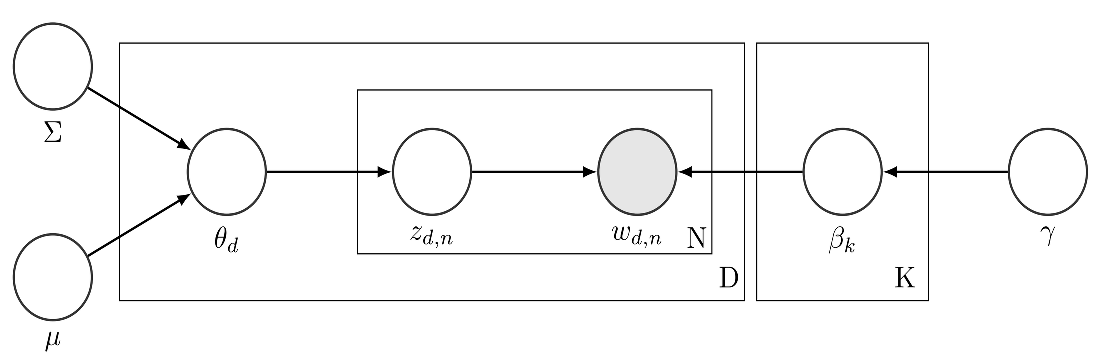
An increasing amount of research focuses on the effects of news and uncertainty on macroeconomic aggregates. Although it is widely agreed that uncertainty exhibits various transmission channels with regard to the real economy and financial markets, little is known about the effects of economic news on macroeconomic and financial expectations. Recent advances in natural language processing have made it feasible to quantify vast amounts of written texts without relying on pre-determined keywords or manual compilations. We thus combine a correlated topic model and a dictionary based sentiment analysis to extract economic topics from approx. 500,000 U.S. newspaper articles. The results are used to investigate which type of news is correlated with professional economic forecasts and whether this relationship is varying over time. We use a flexible version of dynamic model averaging for the econometric analysis, which allows the combination of a large set of dynamic logistic regression models, differing with respect to the included explanatory variables and the degree of time variation in the parameters. The model's weight within the combination is based on the data support for each individual model, that is, its likelihood. The newspaper articles are obtained from *LexisNexis Group* and the survey data from *Consensus Economics*.

Text Analysis I

Topic model and sentiment analysis

Probabilistic Topic Models

Probabilistic topic models enable the automatic detection of themes in large collections of written documents without additional document information, such as topic labels, classifications or annotations. The most prominent and widely used algorithm is latent Dirichlet allocation by Blei et al. (2003), which has been extended to the correlated topic model (CTM), allowing for correlations between topic proportions (Blei and Lafferty, 2007).



Graphical Model of CTM.

- A collection of documents (corpus) consists of K topics and D documents. Each document d consists of N words.
- Shaded nodes are observed, unshaded nodes are hidden random variables.
- Plates indicate replicated variables.
- $\beta_k \sim$ Dirichlet (γ): Topics (topic level)
- $\theta_d \sim$ Logistic Normal (Σ, μ): Topic proportions (document level),
- $z_{d,n} \sim$ Multinomial (θ_d): Topic assignment (word level),
- $\omega_{d,n} \sim$ Multinomial $(\beta_{z_{d,n}})$: Observed word (word level).

Newspaper Analysis

- We analyze appr. 500,000 *New York Times* and *Washington Post* articles which at least once use the word *econom*.
- We use *dplyr* by Wickham et al. (2018) and *quanteda* by Benoit (2018) to prepare the text documents.
- We run a CTM with 100 topics by using the *stm-package* by Roberts et al. (2018). The package uses a *partially collapsed variational EM algorithm*, developed by Roberts et al. (2016).
- We estimate for each document d a topic proportion $\hat{\theta}_d$ and choose those topics with economic content.

Sentiment Analysis

- We follow Dybowski and Adämmer (2018) and capture the tone of each article with the dictionary by Young and Soroka (2012), implemented in *quanteda*:

$$\phi_d = \frac{(w_{d,n}^{pos} + w_{d,n}^{neg/neg} - w_{d,n}^{neg} - w_{d,n}^{neg/pos})}{w_{d,n}^{total}}.$$

$w_{n,d}^{pos}$ $(w_{d,n}^{neg})$ = positive (negative) word counts, $w_{d,n}^{neg/neg}$ $(w_{d,n}^{neg/pos})$ = negative (positive) word preceded by a negation, and $w_{d,n}^{total}$ = total number of words within each document.

Text Analysis II

Document bootstrap and topic/sentiment results

Refining the Sentiment Analysis

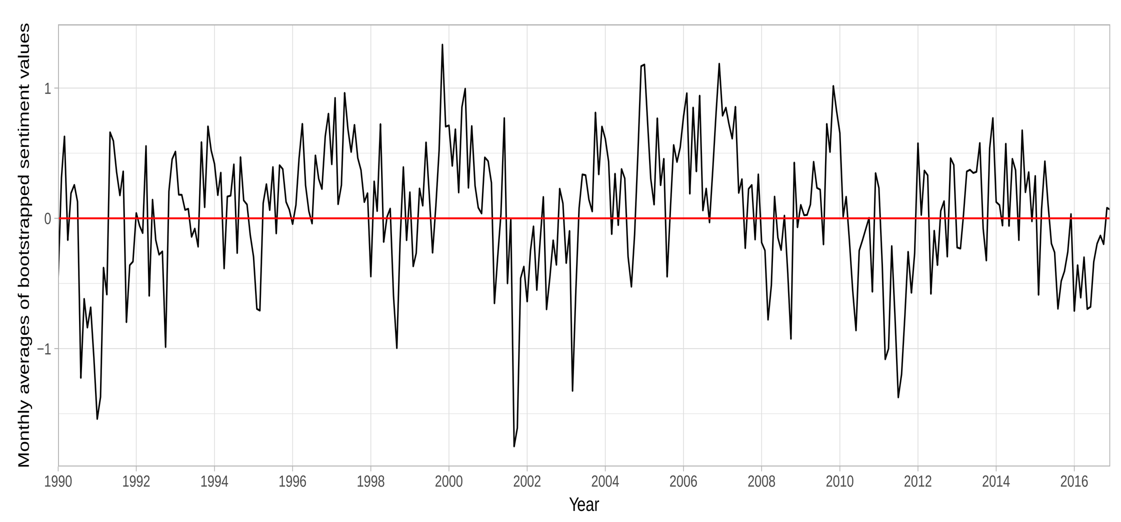
- We use *quanteda* to resample (bootstrap) each newspaper article 1000 times by assuming that sentences are independent but words within a sentence dependent. We estimate the standard deviation of ϕ_d for each document:

$$\sigma_{\phi_d} = \left(\frac{1}{999}\sum_{i=1}^{1000}(\phi_i - \bar{\phi})^2\right)^{1/2},$$

where ϕ_i denotes the sentiment value for the i^{th} bootstrapped document.
- We divide the average sentiment value by its estimated (bootstrap) standard deviation:

$$\phi_{d_B} = \frac{\phi_d}{\sigma_{\phi_d}}.$$

Results: Sentiment Analysis



Monthly averages of ϕ_{d_B}.

Combining Topics with Sentiment Analysis

- We combine the topic proportions with the sentiment value of each document by multiplication:

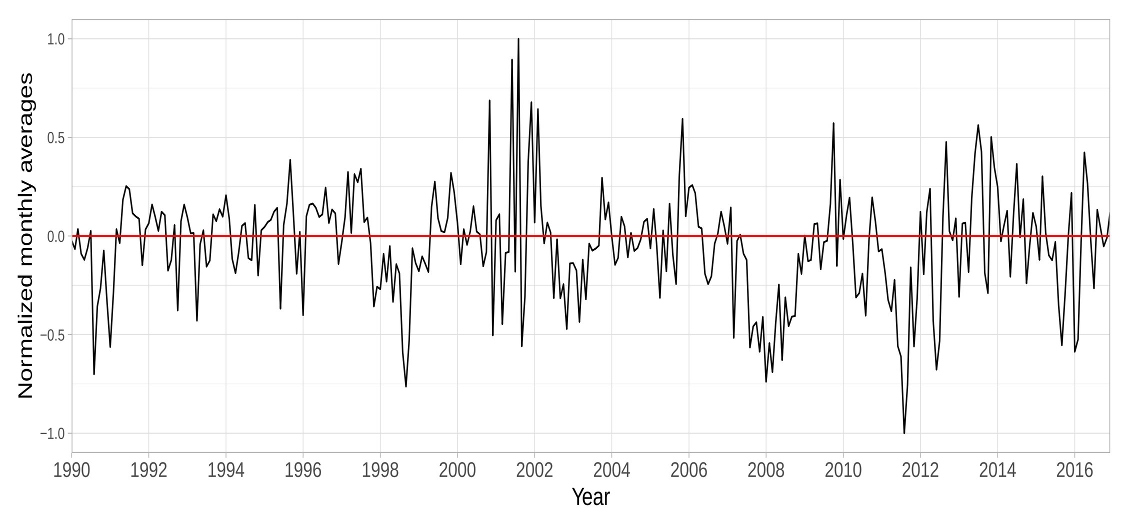
$$\xi_{d,k} = \phi_{d_B}\hat{\theta}_{d,k}.$$

- We calculate monthly averages for the econometric analysis:

$$\bar{\xi}_{m,k} = \frac{1}{D_m}\sum_{d_m=1}^{D_m}\xi_{m,k,d_m},$$

where D_m denotes the number of articles in month m.

Results: Topic and Sentiment



Positively/negatively monthly news about central banking ($\bar{\xi}_{m,64}$).

Econometric Analysis

The impact of news on economic forecasts

Bayesian Dynamic Logistic Regression

- Consider a binary response y_t (i.e., the direction of forecast expectations) and a set of predictors $x_t = (x_{1,t}, x_{2,t}, ..., x_{k,t})$.
- For an arbitrary point in time t we have:

$$y_t \sim \text{Binary}(p_t), \tag{1}$$

with

$$\text{logit}\,(p_t) = x_t^{'}\theta_t, \tag{2}$$

where θ_t is a k-dimensional vector of coefficients (states).
- Prediction equation:

$$\theta_t = \theta_{t-1} + w_t, \quad w_t \sim N(0, W_t). \tag{3}$$

Estimation is done via the Kalman filter.
- Given a set of past outcomes $Y^{t-1} = y_1, ..., y_{t-1}$, we obtain the prediction equation:

$$\theta_t|Y^{t-1} \sim N(\hat{\theta}_{t-1}, R_t), \tag{4}$$

where

$$R_t = \hat{\Sigma}_{t-1}/\delta. \tag{5}$$

- Time-varying coefficients are specified by (5), with the forgetting factor δ (typically slightly less than one).

Dynamic Model Averaging

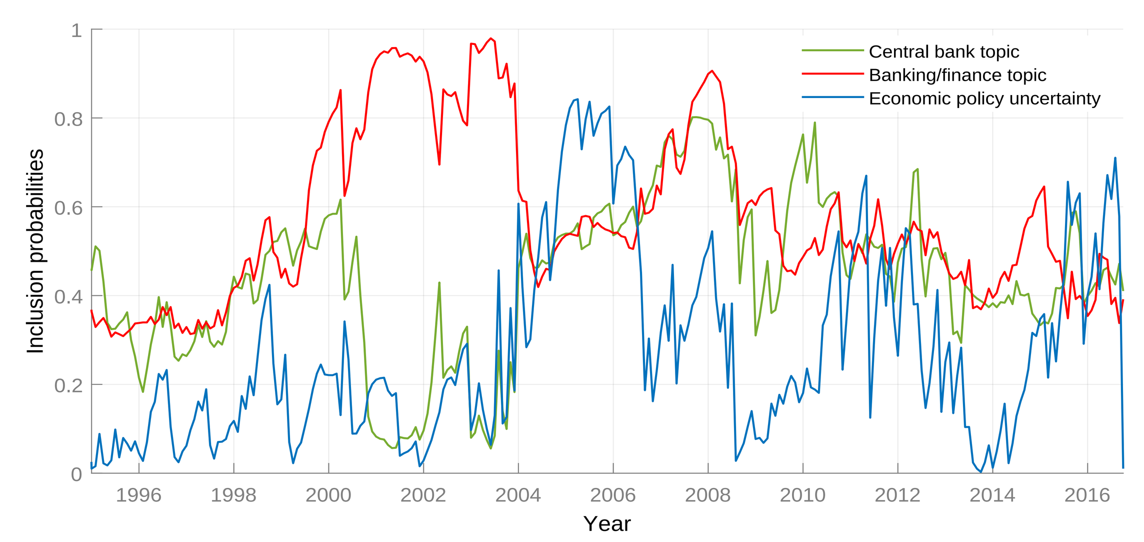
- A model is referred to as M_j ,with $j = 1, ..., J$.
- Let $p(M_i|Y^{t-1})$ be the updated model weight for model i at time $t-1$, and $\mathcal{P}(M_i|Y^{t-1})$ the prior weight for time t.
- The difference between the updated model weight of period $t-1$, $p(M_i|Y^{t-1})$, and the predictive model weight for period t is controlled by the forgetting/discount factor α :

$$\mathcal{P}(M_i|Y^{t-1}) = \frac{p(M_i|Y^{t-1})^{\alpha}}{\sum_{j=1}^{J} p\,(M_j|Y^{t-1})^{\alpha}}. \tag{6}$$

- To update model weights we use Bayes' rule

$$p(M_i|Y^t) = \frac{f(y_t|Y^{t-1}, M_i)\mathcal{P}(M_i|Y^{t-1})}{\sum_{j=1}^{J} f(y_t|Y^{t-1}, M_j)\mathcal{P}(M_j|Y^{t-1})}. \tag{7}$$

Results: Econometric Analysis



Inclusion probabilities for explaining US CPI 12m expectations.

References

Benoit, K. (2018). quanteda: Quantitative analysis of textual data. r package version 1.2.0.

Blei, D. M. and Lafferty, J. D. (2007). A Correlated Topic Model of Science. *The Annals of Applied Statistics*, 1(1):17–35.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Dybowski, T. and Adämmer, P. (2018). The economic effects of us presidential tax communication: Evidence from a correlated topic model. *European Journal of Political Economy*.

McCormick, T. H., Raftery, A. E., Madigan, D., and Burd, R. S. (2012). Dynamic logistic regression and dynamic model averaging for binary classification. *Biometrics*, 68(1):23–30.

Roberts, M., Stewart, B., and Tingley, D. (2018). stm: R package for structural topic models. r package version 1.3.3. *Journal of Statistical Software*.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Wickham, H., Francois, R., Henry, L., and Müller, K. (2018). dplyr: A grammar of data manipulation. r package version 0.7.5.

Young, L. and Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.

Contact: Dr. Philipp Adämmer · Holstenhofweg 85, 22043 Hamburg, Germany · adaemmer@hsu-hh.de