Does unusual news forecast market stress?

Consortium for Systemic Risk Analytics 2015

Paul Glasserman
Harry Mamaysky

May 26, 2015

Introduction

- Automated processing of natural language is opening a previously unavailable window into market behavior
  - Unavailable to *both* practitioners and academics!
- Prior work has documented fascinating relationships between short-term price responses and news sentiment
- We provide evidence that:
  1. Not only sentiment matters, but *unusualness* does as well
  2. News effects in markets happen over months, not days

An example

- Two phrases from news articles from September 2008:

  "the collapse of Lehman"
  "problem accessing the internet"

- Both contain negative words
- Intuitively, the first one probably has more meaningful content
- According to our definition of unusualness, the first is one of the most unusual negative phrases in September 2008, and the second one of the least unusual
- How to come up with a probabilistic model for language?

n-grams

- Consider word sequence

$$w_1^N \equiv w_1 w_2 \cdots w_N$$

- Using chain rule, we can write

$$P(w_1^N) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)\cdots P(w_N|w_1^{N-1}) \qquad (1)$$

- For an n-gram model we assume these conditional probabilities only depend on the prior $n-1$ words
  - For a 4-gram model, we assume

$$P(\text{industry}|\ldots \text{ deal comes amid sweeping changes in the US cable}) = P(\text{industry}|\text{the US cable})$$

- Dropping the first few terms in (1), we can then write

$$P(w_1^N) \approx \prod_{k=n}^{N} P(w_k|w_{k-n+1}^{k-1})$$

for an n-gram grammar

Likelihood of a corpus

- We work with a 4-gram model
- Given a training corpus, we can estimate the 4-gram probability of word k via

$$m(w_k|w_{k-3}w_{k-2}w_{k-1}) = \frac{C(w_{k-3}w_{k-2}w_{k-1}w_k)}{C(w_{k-3}w_{k-2}w_{k-1})}$$

  where C is the count operator
- Given a new corpus with N words, we can compute its per word log probability score, i.e. $1/N \log P(w_1^N)$ via

$$\frac{1}{N}\sum_{k=4}^{N} \log m(w_k|w_{k-3}^{k-1}) = \sum_{W \in \text{All 4-grams}} p(W) \log m(w_4|w_1^3)$$

  where p is the in-sample probability of the 4-gram $W = w_1w_2w_3w_4$
- As $N \to \infty$, this quantity converges to the *cross-entropy* of m (model) with respect to the true word sequence probabilities

Key concepts

Entropy

- Entropy for a given collection of 4-grams is given by

$$\mathbf{p} \cdot \log \mathbf{m}$$

- **m** comes from training corpus
- **p** comes from any subset of current data
- So we can speak about: *negative* entropy, *positive* entropy, *Citigroup* entropy, etc.

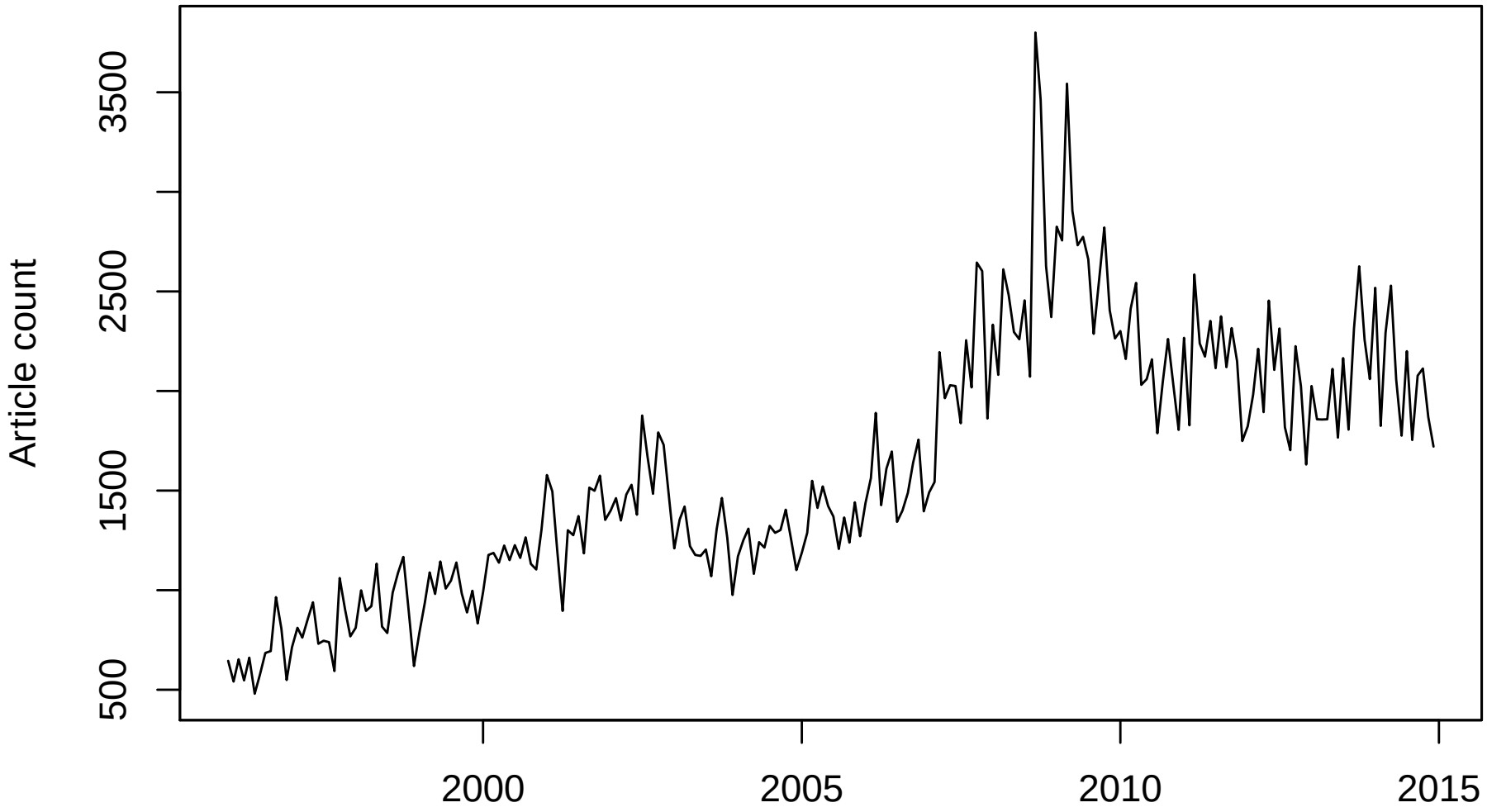
Sentiment

- Likewise we can classify 4-grams by sentiment, for the entire sample or subsets
- For example:
  - % of all 4-grams containing only negative words
  - % of all 4-grams that come from articles that mention JP Morgan that contain negative words
  - etc.

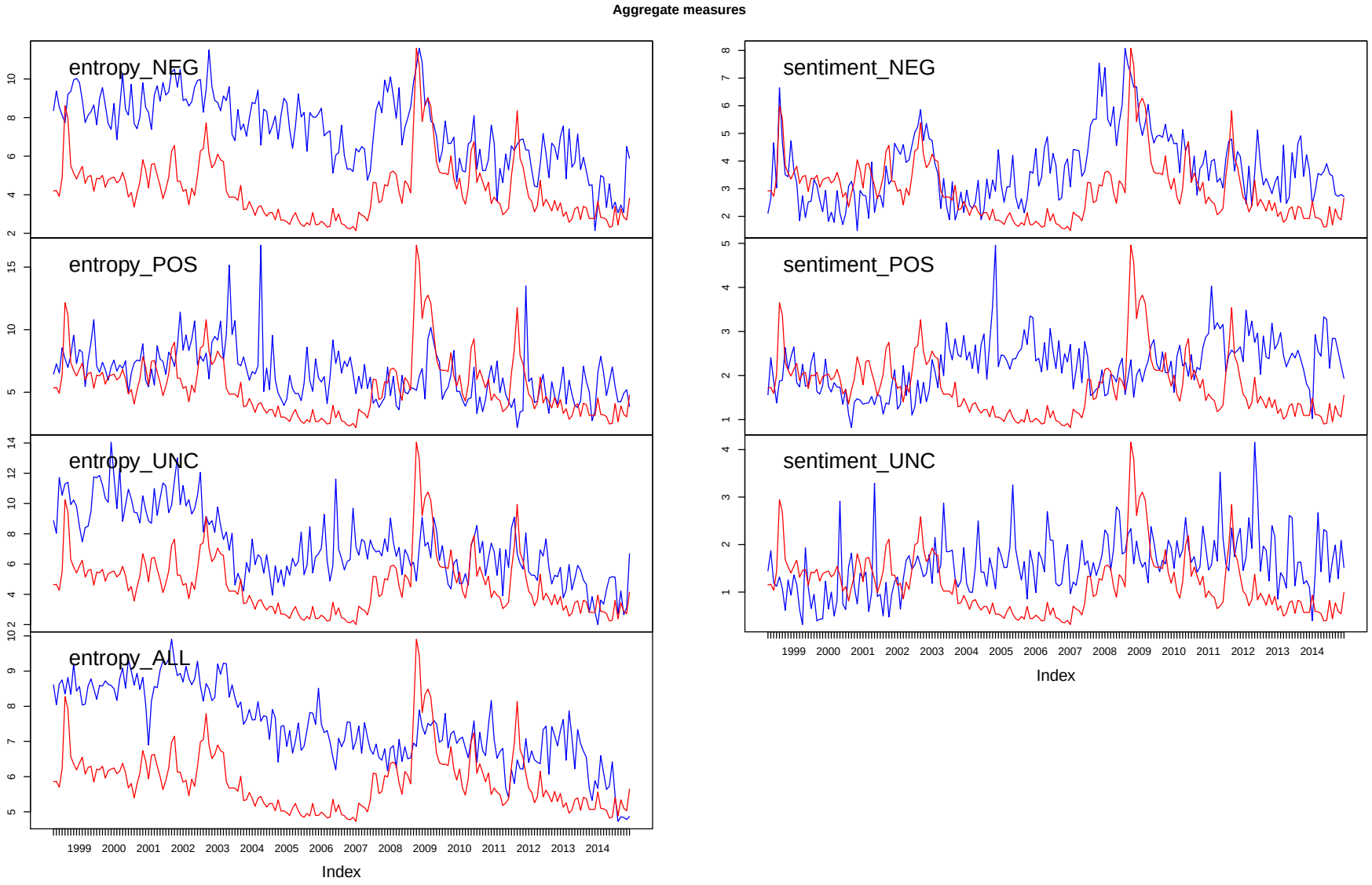
Data

- All articles from Thomson-Reuters archive for 50 largest global financial firms (as of February 6, 2015)
- Data from 1996 to 2014 for 228 year-months
- 367,331 articles
- Average of 1,611 articles per month
- Sentiment dictionary from Loughran-McDonald (`http://www3.nd.edu/~mcdonald/Word_Lists.html`)

Article count per month



Entropy and sentiment



NOTE: Scaled VIX shown in red

Single-name tests

- For each single name i, run time series regression:

$$iVol^i_{1mo}(t) = c + \sum_{l=1}^{6} s^i_l \; sentNEG^i(t-l) + \sum_{l=1}^{6} e^i_l \; entNEG^i(t-l) + \epsilon^i(t)$$

- These measures come from subset of 4-grams from articles mentioning company i
- Report cross-sectional averages of coefficients

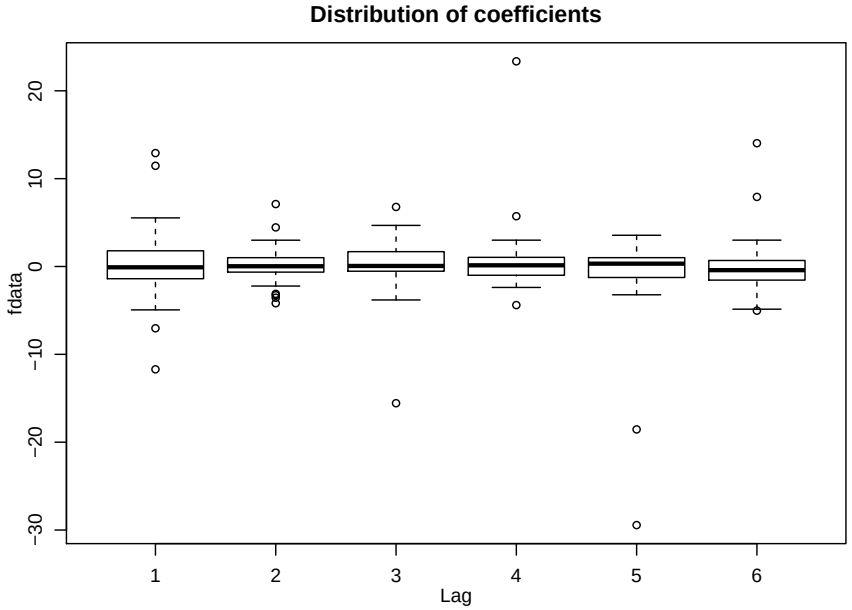
$$s_l \equiv \frac{1}{N} \sum_{i=1}^{N} s^i_l$$

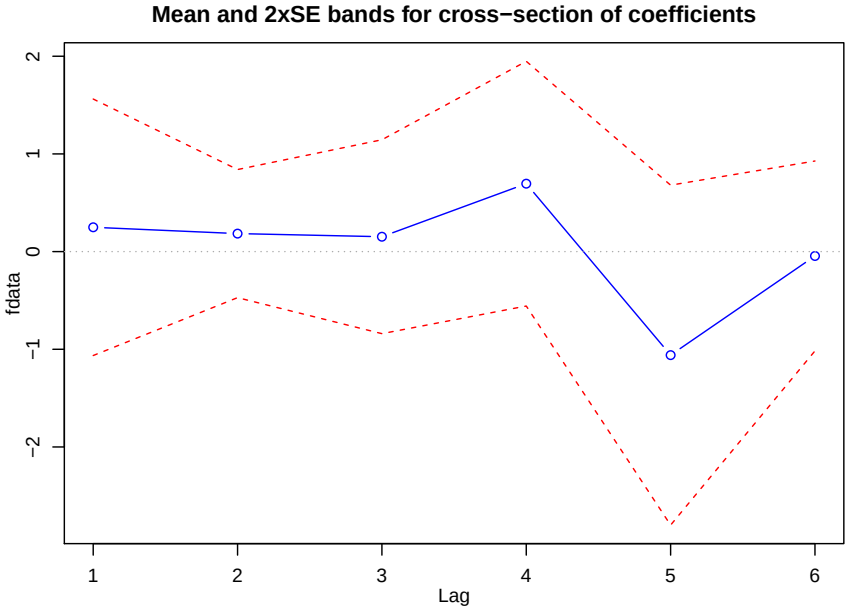
and

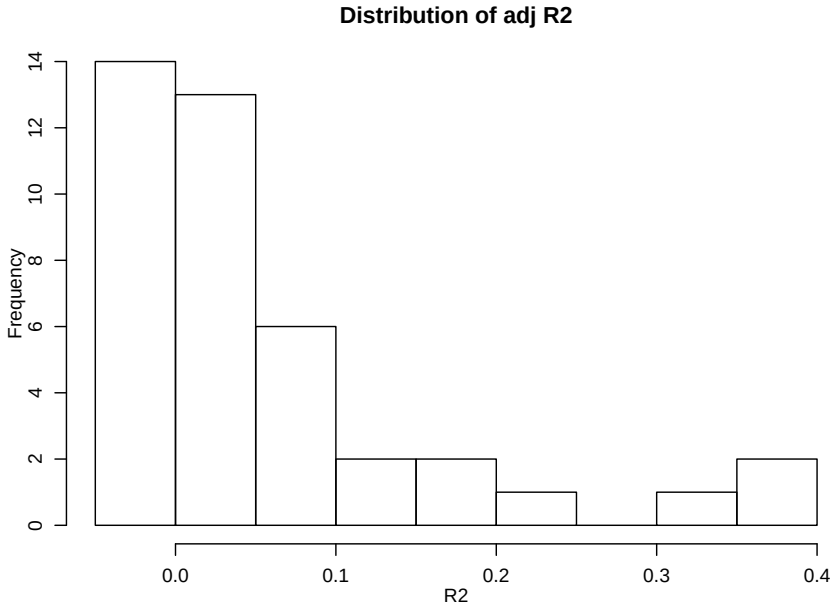
$$c_l \equiv \frac{1}{N} \sum_{i=1}^{N} c^i_l$$

Single-name control

Regression summary for article percent of total

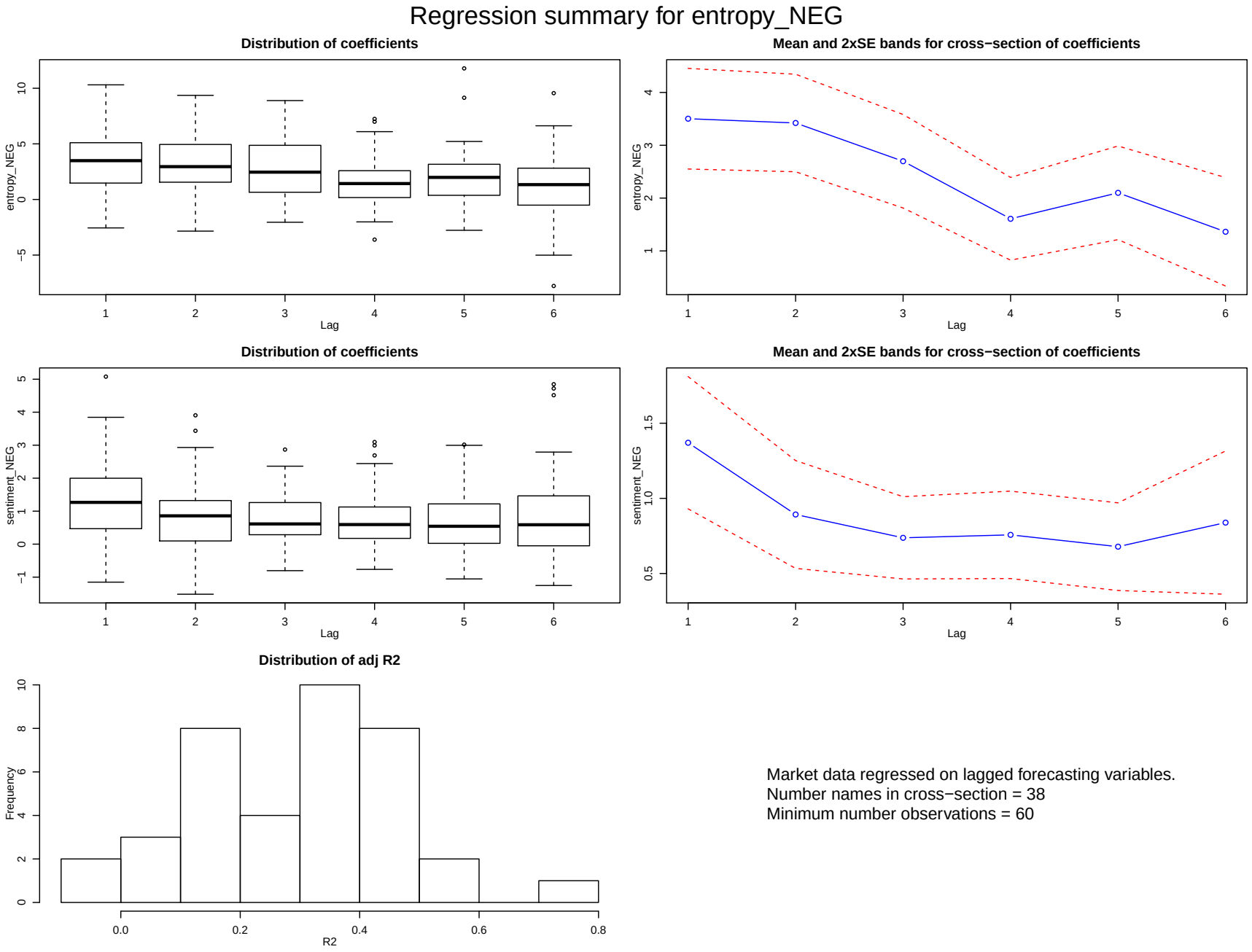






Market data regressed on lagged forecasting variables.
Number names in cross−section = 41
Minimum number observations = 60

Single-name results



Market data regressed on lagged forecasting variables.
Number names in cross−section = 38
Minimum number observations = 60

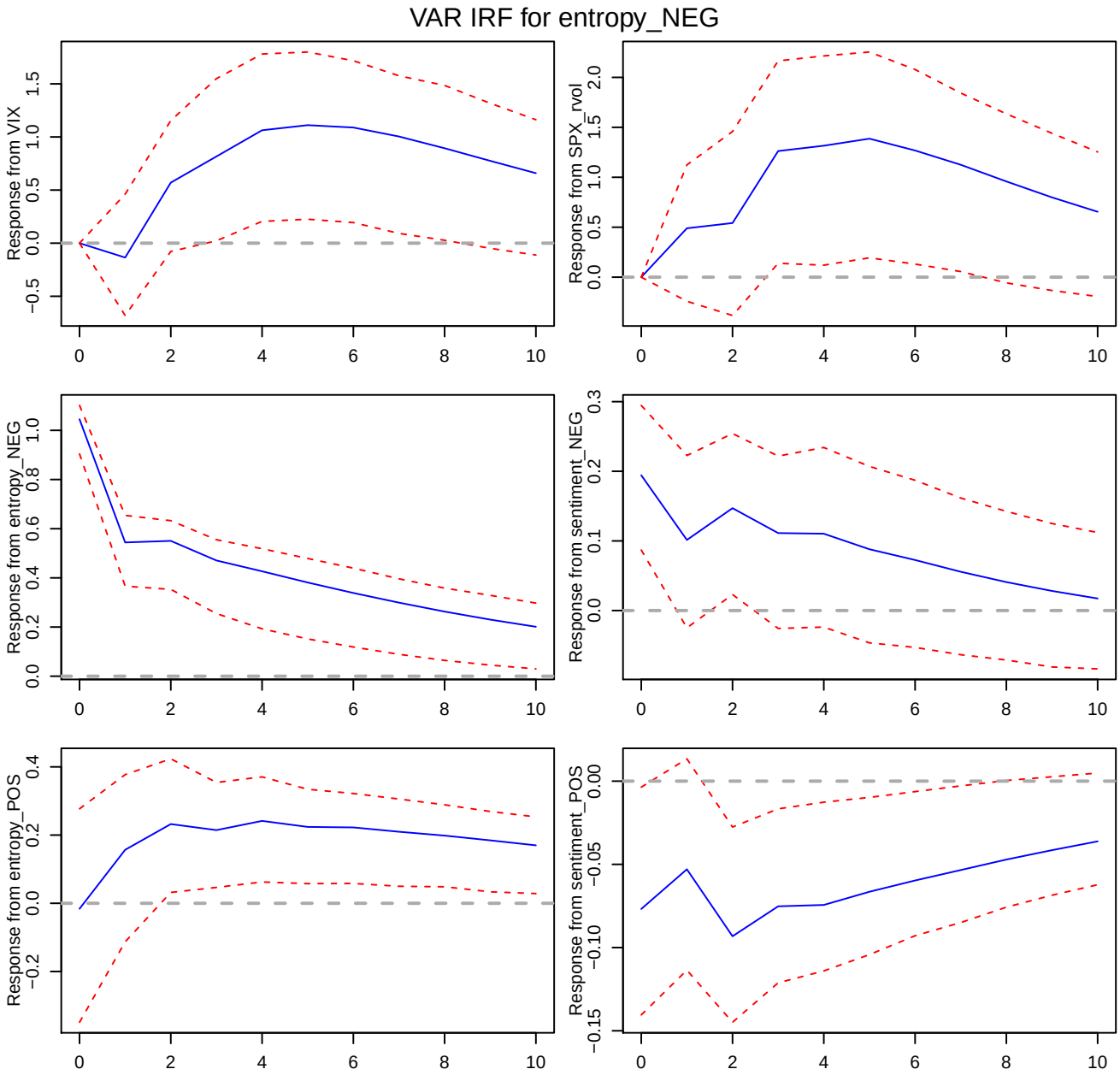
Aggregate level impulse response

- We estimate a VAR(2) model for the following vector:

$$\begin{bmatrix} VIX \\ rVol_{1mo}^{SPX} \\ entNEG \\ sentNEG \\ entPOS \\ sentPOS \end{bmatrix}$$

- Report orhogonalized impulse response
- Given the above ordering, the *entNEG* response is the part of *sentNEG* that is due to negative entropy
  - **p** comes from 4-grams with only negative words

Impulse response to negative entropy shock



Main results

- Today's news predict future implied and realized volatility, and therefore future market stress
- Forecasting horizon is in months, not days
- A surpsising amount of future volatility can be explained with today's news
- Sentiment and unusualness both matter
- Lots more to be done!

Thank you!