Measuring Uncertainty through Word Vector Representations

J. Daniel Aromí

IIEP UBA-Conicet

November 2017

Table of Contents

Introduction

Uncertainty: Assessment of the (in)ability to anticipate future scenarios (subjective/unobserved/multidimensional).

Relevant in macroeconomic/financial contexts:

- Delays in consumption and investment decisions (Bernanke 1983, Dixit and Pindick 1994)
- Risk aversive/precautionary behavior (Ilut and Schneider 2011, Basu et al. 2017).
- More severe consequences of contractual imperfections (Christiano et al. 2014, Gilchrist et al. 2014).

Measuring uncertainty

Motivation:

- Interpretation of economic events (expansions, recessions, crises).
- Decision making (risk management, policy making, speculation).

Existent proxies:

- Estimation of conditional volatility (Jurado 2015)
- Asset markets (VIX, realized volatility).
- Surveys: households and professionals (Michigan, SPF).
- Text analysis (Baker et al. 2016)

This contribution:

A metric based on a novel unsupervised learning technique and a large corpus (WSJ 1990-2017).

Methodology

Two steps:

1. **Word vector representations:** Linear structure of meaning based on term co-occurrence matrix. GloVe algorithm (Pennington et al. 2014).

2. **Computation of uncertainty index:** Words related to uncertainty are identified through word vector representations. Then, their relative frequency is calculated.

Data: The Wall Street Journal 1900-2017.

Word vector representation

Following Pennington et al. (2014):

- Let W denote a dictionary.
- Let X_{ij} denote the number of times word i occurs in the context of word j.
- Word vector representations $\{v_i\}_{i \in W}$ solve:

$$\min_{\{v_i\}_{i \in W}} \sum_i \sum_j f(X_{ij}) \left[v_i' v_j + b_i + b_j - log(X_{ij}) \right]^2$$

where $f(X_{ij})$ is a weighting function and b_i is word i's bias.

Comments:
Efficient factorization of semantic (and syntactic) information.
E.g. BUNDESBANK-GERMANY+US=FED, GM-CAR+AIRPLANE=BOEING.
Distance between vectors as word similarity.

Parameters: vector dimensionality, size of context window, weighting function.

Uncertainty Index

Second stage:

- Identify a set of **uncertainty related words** $U \subset W$ using the distance between word vector representations.
  Three seed words: uncertainty, uncertain, uncertainties.
  Closest 500 words (similar results under alternative cutoffs).

- Compute the frequency of words in U:

$$UI_t = \frac{\sum_{i \in U} c_{it}}{\sum_{i \in W} c_{it}}$$

  where c_{it} is the number of times word i is observed in period t.

Data

Content published by The Wall Street Journal.
For each article, the corpus includes: title, lead and first lines.
Extracted from: http://pqasb.pqarchiver.com/djreprints/

Two datasets:

Corpus	Number of articles	Number of tokens
Training (1900-1989)	3,233,481	134,797,611
Test (1990-2017)	1,241,706	98,979,322

Training dataset dictionary: 28296 words (after applying 100-minimum-counts filter).

Evaluation of the Uncertainty Index

1. **Qualitative analysis:**
   - Do **word vector representations** identify uncertainty related terms?
   - Does the index correlate with recessions, crises?
   - Which is the association with other uncertainty proxies?
2. **Forecasting expected volatility (VIX):**
   - Does it contain information (beyond that provided by lags of VIX)?
   - Do simpler text based indicators (e.g. Baker et al. EMU) lead to similar results?

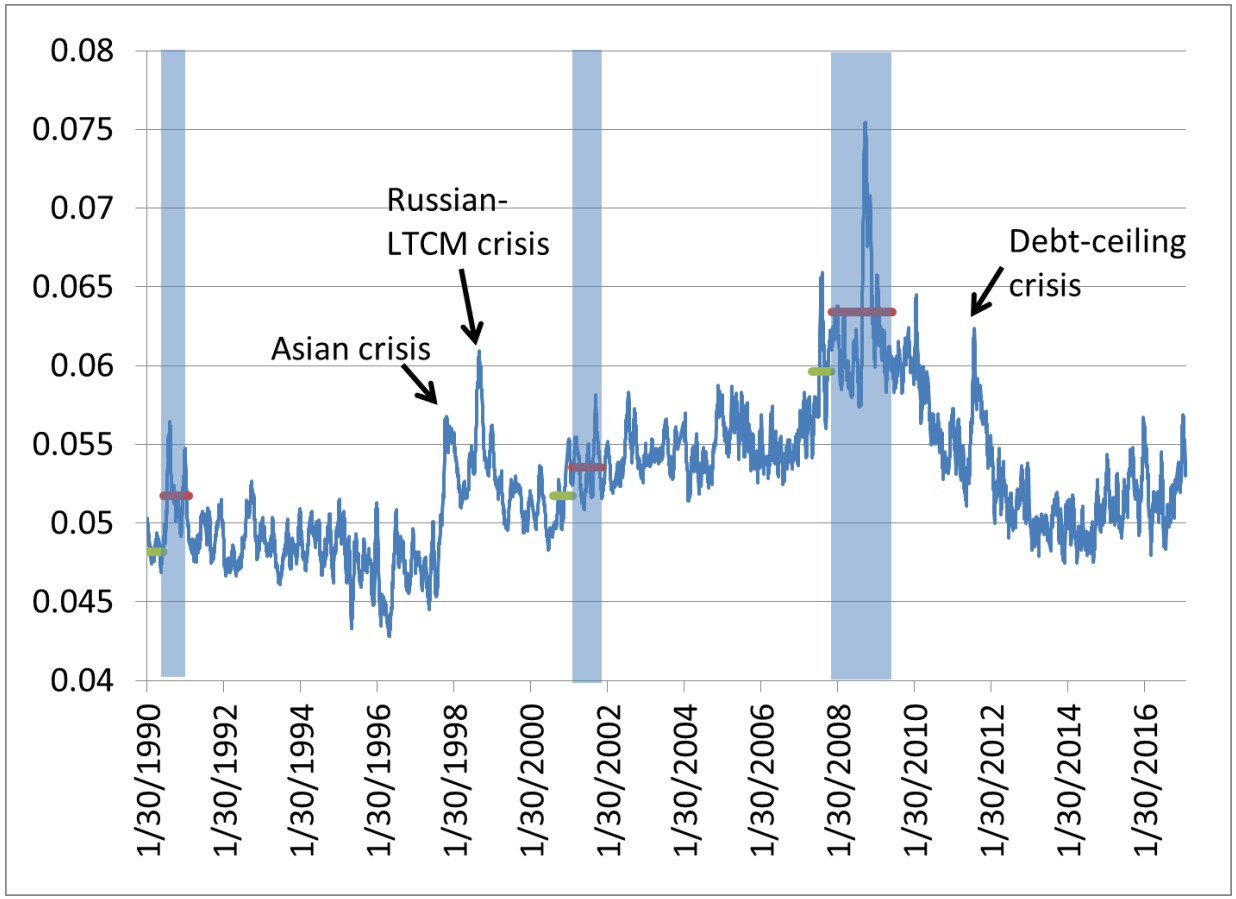
Uncertainty Related Words

50 Closest Words:

uncertainty	uncertainties	uncertain	unsettled	disturbed
situation	confusion	nervousness	clouded	feeling
confused	apprehension	view	outcome	outlook
fears	uneasiness	sentiment	developments	doubt
owing	unsettlement	complications	optimism	anxiety
reflecting	surrounding	disturbing	nervous	future
effects	coupled	political	considerations	restricted
prospects	clarification	pessimistic	regarding	worries
unfavorable	adverse	impact	conditions	troubles
pessimism	fate	belief	extremely	crisis

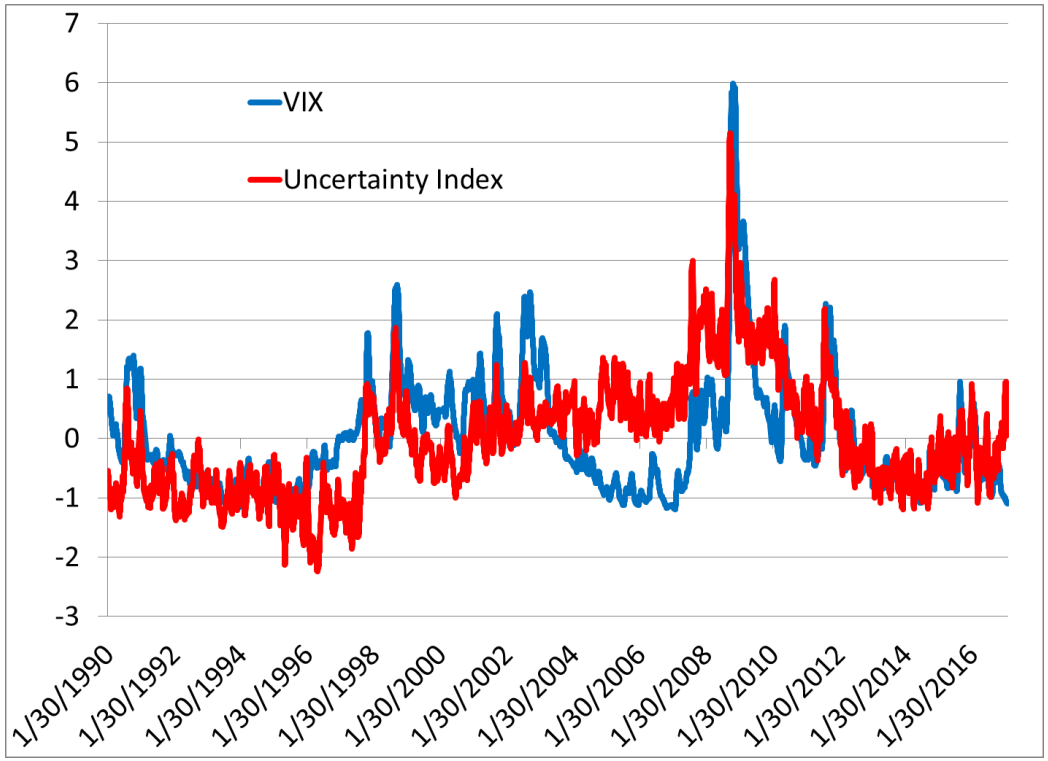
Parameters: 100 dimensions, +/-5 steps, 100-min-count vocabulary filter, 100 threshold in weighting function.

Uncertainty Index



Note: Smoothed index (20-day moving windows).

Uncertainty Index and VIX



VIX: CBOE's implicit volatility index (implied volatility over the next 20 business days). Standardized values.

Uncertainty Index and VIX

Descriptive Statistics

	Mean	St.Dev.	Min	Max
VIX	19.603	7.911	9.310	80.860
UI	0.053	0.006	0.034	0.095
EMU	3.683	1.057	1.569	7.502

Correlations

	VIX	UI	EMU
VIX	1	0.421	0.384
UI		1	0.170
EMU (Baket et al. 2016)			1

EMU: Equity Market Uncertainty Index proposed by Baker and coauthors in: www.policyuncertainty.com

Parsimonious forecast models:

- Heterogeneous Autoregressive Model

$$VIX_{t+h-1} = \alpha + \beta_1 VIX_{t-1} + \beta_5 VIX_{[t-5,t-1]} + \beta_{20} VIX_{[t-20,t-1]} + u_t$$

- Heterogeneous Autoregressive Model + Proxy

$$VIX_{t+h-1} = \alpha + \beta_1 VIX_{t-1} + \beta_5 VIX_{[t-5,t-1]} + \beta_{20} VIX_{[t-20,t-1]} + \beta_{Pr} Proxy_{[t-5,t-1]} + u_t$$

Proxies:

- Uncertainty Index (UI): GloVe, 500 words.
- Equity Market Uncertainty (EMU): Baker et al.
- 3 words index (3w): uncertainty, uncertain and uncertainties.

VIX forecasts (in sample)

Forecast horizon $h = 20$

	HAR	HAR+UI	HAR+EMU	HAR+3w
c	3.16***	4.21***	3.40***	3.05***
	(0.77)	(0.88)	(1.06)	(0.81)
β_{t-1}	0.50	0.49***	0.50***	0.49***
	(0.09)	(0.08)	(0.09)	(0.09)
$\beta_{[t-5,t-1]}$	0.30**	0.23**	0.31**	0.32**
	(0.13)	(0.12)	(0.13)	(0.13)
$\beta_{[t-20,t-1]}$	0.04	0.06	0.04	0.04
	(0.17)	(0.15)	(0.16)	(0.17)
β_{Pr}	-	0.72**	-0.09	-0.21
	-	(0.36)	(0.26)	(0.16)

Newey-West standard errors in parenthesis.

VIX forecasts(out of sample)

Models trained with 1000 day windows.
Forecast horizon $h = 20$

A. Mean squared errors

	HAR A	HAR+UI B	Ratio B/A	DM text p-value
5-day	3.10	3.07	0.99	0.262
20-day	5.12	4.94	0.97	0.024
40-day	6.70	6.40	0.96	0.064

B. Mean absolute errors

	HAR A	HAR+UI B	Ratio B/A	DM text p-value
5-day	2.07	2.05	0.99	0.336
20-day	3.34	3.18	0.95	0.001
40-day	4.46	4.15	0.93	0.001

Diebold-Mariano tests adjusted by forecast horizon.

Concluding remarks

Application of NLP tools in macroeconomic/financial contexts:

- Suggestive correlations with relevant events.
- Anticipation of expected volatility.
- Contrast with the performance of simple text analysis techniques.

Work ahead:

- Beyond expected volatility (e.g. macro surprises Scotti 2015).
- More specific measures of uncertainty (monetary policy, corporate investment).
- UI as data in structural analyses.