Seeing the Forest for the Trees: Using hLDA Models to Evaluate Communication in Banco Central do Brasil

Angelo M. Fasolo Flávia M. Graminho Saulo B. Bastos

Banco Central do Brasil

BIS, March, 8th, 2021

Disclaimer:

This work should not be reported as representing the views of the Banco Central do Brasil. The views expressed in this presentation are those of the authors and do not necessarily reflect those of the Banco Central do Brasil.

Introduction and motivation

- There is a widespread consensus that Central Bank communication influences the expectations of economic agents and that increasing transparency enhances the effectiveness of monetary policy.
- Literature on Central Bank communication $\rightarrow$ measures the content of communication to learn signals about monetary policy decisions and estimates its impact on the economy.
- Traditional algorithms, however, usually create **independent** "bags of words", without considering the degree of abstraction of words for a given topic or offering insights about the structure of the document. Critical parameters of the algorithms are predefined instead of being endogenously estimated from data.

In this paper

- Estimate hierarchical LDA (hLDA) model to extract the content and measure the tone of communication in Banco Central do Brasil's monetary policy decisions.
- hLDA model: organize the "bags of words" in a "tree of topics", providing the basis for indexes measuring the perception of BCB's monetary policy committee (Copom) on different aspects of the economic situation.
- Evaluate the coherence of BCB's communication between statements and minutes of the Copom meetings.
- Computational linguistics: using feature selection techniques before the hLDA model's estimation.

Literature

- Transparency and monetary policy: Swanson (2006), Neuenkirch (2012), Jitmaneeroj, Lamla and Wood (2019).
- Text analysis and monetary policy: Boukus and Rosenberg (2006), Bailey and Schonhardt-Bailey (2008), Lucca and Trebbi (2009), Hendry and Madeley (2010), Hendry (2012), Acosta et al (2015), Hansen and McMahon (2016).
- Tone indices: Labondance and Hubert (2017), Shapiro and Wilson (2019).
- In Brazil: Carvalho, Cordeiro and Vargas (2013), Cabral and Guimaraes (2015), Chague, De Losso, Giovannetti and Manoel (2015), Montes, Oliveira, Curi and Nicolay (2016), Garcia-Herrero, Girardin, and Dos Santos (2017).

Main results

- "Tree of topics" estimated from hLDA model provides information on the structure of documents about four topics: inflation, economic activity, monetary policy, and financial markets and international economics.
- Situation indexes derived from the model are highly correlated with inflation, economic activity, and uncertainty.
- Model is capable of handling structural breaks in communication.
- BCB's communication is usually coherent, with minutes presenting detailed information first offered in statements.

Outline

From LDA to hLDA model

- hLDA (Griffiths et ali, 2004): Unsupervised Bayesian nonparametric model.
- Topics are organized according to a hierarchy $\rightarrow$ Easier to interpret.
- In LDA (Blei, Ng and Jordan, 2003), topics are not organized.
- Additional step in Gibbs-Sampler, drawing from the posterior also the shape of the tree of topics ("Nested Chinese Restaurant Process").
- Drawbacks:
  - Several local maxima in the likelihood function of the model;
  - Significantly larger number of iterations for convergence.

Indexes of economic situation

- Indexes of Economic Situation: measuring the tone of communication.

- Building indexes in the context of hLDA models requires:
  - Given estimated hLDA model, associate leaves of the tree with the subject;
  - For a given set of leaves and the subject, create a dictionary of words characterizing the sentiment with respect to the subject;
  - Using the model, locate all paragraphs associated with the subject and, inside each paragraph, keywords defined in the dictionary;
  - Establish a metric comparing the frequency of positive and negative words found in each document.

- Formally:

$$\text{SubjSit}_m = \frac{\sum_{(m,s)} a_{(m,s)} \left(\text{Pos}_{(m,s)} - \text{Neg}_{(m,s)} \right)}{\sum_{(m,s)} a_{(m,s)} \text{Tot}_{(m,s)}}$$

Indexes of economic situation

- Reverse of the sentiment – Shapiro and Wilson (2019)
- Example: Situation on Inflation (Copom meeting 202, October 2016)

> *"5. Returning to the domestic economy, recent inflation figures came in more* **favorable** *than expected, partly due to the* **reversal** *of food price* **increases**. *These results contributed to a* **decrease** *in expectations for 2016 IPCA inflation measured by the Focus survey, which stood at around 7.0%. As for 2017, IPCA inflation expectations reported in the same survey have* **declined** *to around 5.0% and remain* **above** *the inflation target of 4.5%. Expectations for 2018 and more distant horizons are already around this level."*

Example: economic activity

Textual data and feature selection

- **Dataset:**
  - Minutes (main dataset): description of motivation for monetary policy decision, published one week after the meeting – sample July 1999 to May 2020.
  - Statements: announcement of the monetary policy decision.
  - Structural break: increase in content of statements after July 2016 meeting.

- **Preprocessing of dataset:** remove all non-alphanumeric characters; text in lower case; tokenization and evaluation of compound words (e.g. *"produto interno bruto"*, GDP, in Portuguese); remove common stopwords (months, days of the week, Brazilian States and Capitals, and the name of Copom's members).

- **Feature selection** (Baeza-Yates and Ribeiro Neto, 2008): reduce the dimensionality of the vocabulary by the exclusion of stopwords and by removing words based on its frequency in the dataset; choice: remove words not used in at least three sentences across all documents.

The structural break of 2016

Figure 1: Statistics of minutes and statements

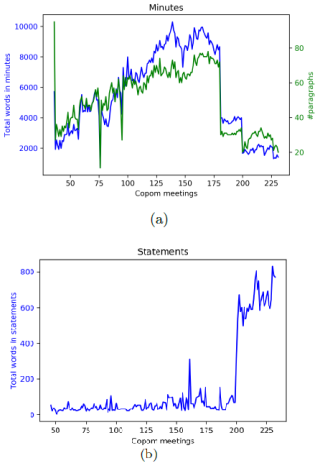
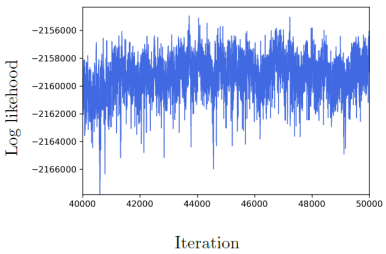


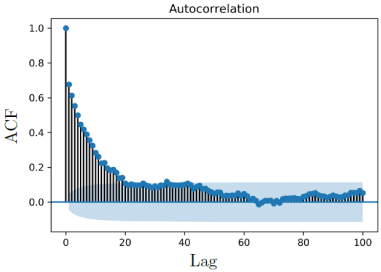
Table 1: Statistics of minutes and statements of Banco Central do Brasil

		Copom meeting			
		36 (Jun/1999) – 81 (Feb/2003)	82 (Mar/2003) – 180 (Jan/2014)	181 (Feb/2014) – 199 (Jun/2016)	200 (Jul/2016) – 231 (Jun/2020)
Minutes	#words	3727.9 ± 1195.3	7887.9 ± 1593.6	3851.4 ± 151.8	1831.5 ± 265.9
	#paragraph	43.1 ± 11.5	64.1 ± 8.3	31.0 ± 1.2	27.4 ± 3.9
	#sentence	138.6 ± 38.6	255.7 ± 46.3	119.9 ± 5.1	68.4 ± 10.6
	$\frac{\#words}{\#paragraph}$	86.3 ± 53.1	122.9 ± 71.9	124.2 ± 58.1	66.6 ± 39.9
	$\frac{\#words}{\#sentence}$	26.8 ± 13.8	30.8 ± 13.6	32.1 ± 12.2	26.7 ± 11.7
Statements	#words	32.9 ± 13.2	56.0 ± 38.8	60.7 ± 35.7	630.5 ± 96.0
	#sentence	1.6 ± 0.7	1.7 ± 1.2	2.0 ± 1.1	24.1 ± 4.4

hLDA Estimation

- Estimation of the model: Metropolis-Hastings and Montecarlo step inside Gibbs-Sampler to combine the estimation of hyperparameters and parameters of the model.
- Estimation of hyperparameters: Draw of Metropolis-Hastings (as in Blei et al (2010)) and Montecarlo procedure (as in Escobar and West (1995)).
- Gibbs-Sampler: after initialization, 50,000 draws, discard first 40,000 as burn-in.





The model tree

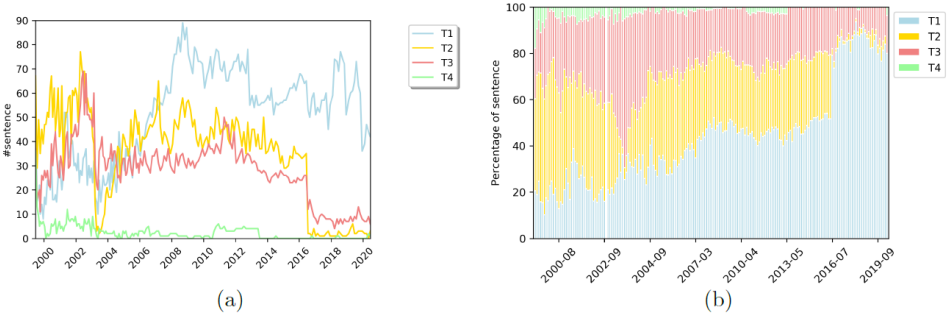
Why not LDA?

Why Feature Selection?

hLDA topics over time

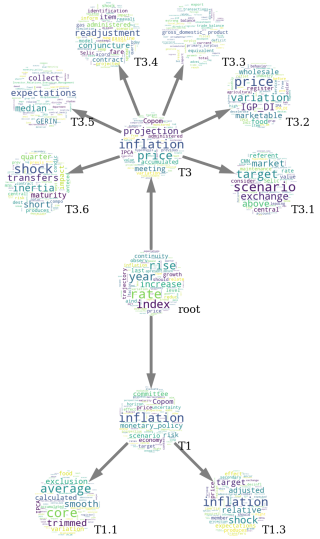
- Significant change in composition of topics after structural break of 2016

Figure 4: Evolution of topics of the hLDA model

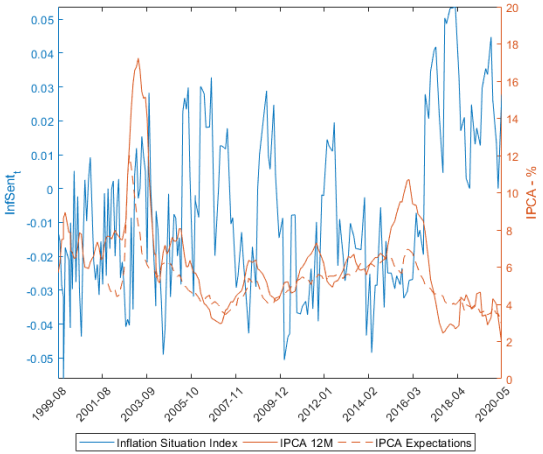


Inflation situation

- Normative analysis: "low inflation" is good.
- Dictionary:
  - Positive words: *"adequate"*, *"low"*, *"benign"*, *"contraction"*, *"decrease"*...
  - Negative words: *"above"*, *"high"*, *"acceleration"*, *"increase"*, *"elevation"*...
- Expected negative correlation with inflation and inflation expectations.
- Leaves added from Topic 1 ("Monetary Policy") discussing the evolution of core inflation and Copom's expectations about inflation.



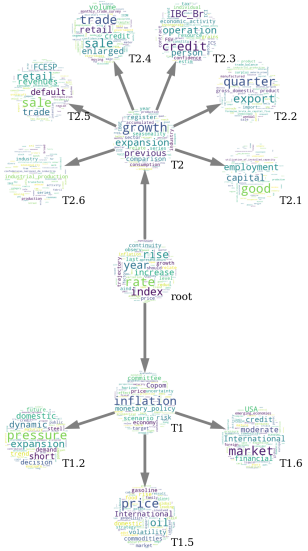
Inflation situation



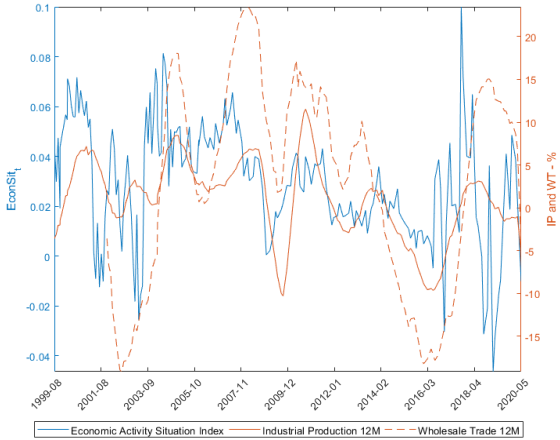
- Corr(IPCA): -0.326 (Full sample), -0.556 (After July 2016).
- Corr(Inflation expectations): -0.451 (Full sample), -0.274 (After July 2016).

Economic activity situation

- Normative analysis: in general, "high economic activity" is good.
- Dictionary:
  - Positive words: *"adequate"*, *"high"*, *"benign"*, *"increase"*, *"elevation"*...
  - Negative words: *"below"*, *"low"*, *"retreat"*, *"deteriorate"*, *"decrease"*...
- Expected positive correlation with industrial production and retail sales.
- Leaves added from Topic 1 ("Monetary Policy").
- Leaves about labor market removed from Topic 2: avoid problems with dictionary and "Unemployment".



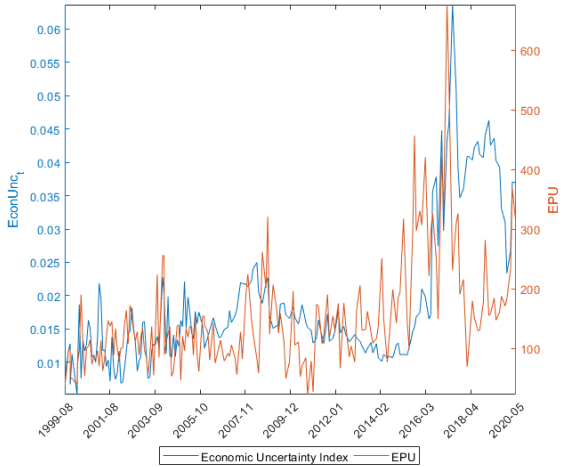
Economic activity situation



- Corr(Industrial production): 0.361.
- Corr(Wholesale trade): 0.225.

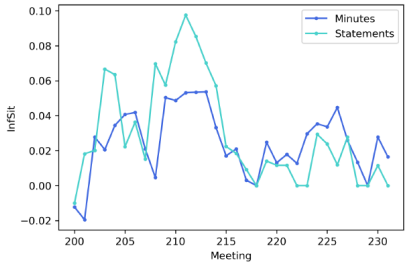
Economic uncertainty index

- Dictionary: based on word list of Loughran and McDonald (2011).
- Use of the whole tree: smoother series and different sources of uncertainty.
- Corr(EPU): 0.484.

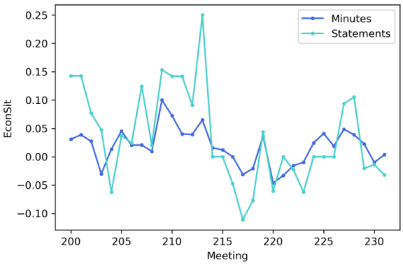


Statements and minutes

- The hLDA model and the dictionaries are used to build situation indexes based on statements of Copom meeting after July 2016.
- The temporal structure of documents allows for a simultaneous analysis both in terms of a given Copom meeting and across consecutive meetings.
- Indexes from the statements are usually more volatile compared to those of the minutes partly due to the smaller overall number of words used in statements.



(a) Inflation situation



(b) Economy situation

Statements and minutes

- Test: use system of simultaneous equations (SUR estimation).

$$\text{Ind}_t = \alpha_0 + \alpha_1 \widehat{\text{Ind}}_t + \alpha_2 \text{Ind}_{t-1} + \alpha_3 \text{Weight}_t + \alpha_4 \text{X}_t + \epsilon_{1,t}$$
$$\widehat{\text{Ind}}_t = \beta_0 + \beta_1 \text{Ind}_{t-1} + \beta_2 \widehat{\text{Ind}}_{t-1} + \beta_3 \widehat{\text{Weight}}_t + \beta_4 \hat{\text{X}}_t + \epsilon_{2,t}$$

- Main hypothesis 1: for a given Copom meeting the sentiment expressed in minutes is consistent with the sentiment expressed in the statements ($\alpha_1 > 0$).

- Main hypothesis 2: the sentiment in the previous Copom meeting influences the sentiment in the current meeting ($\alpha_2 \neq 0$ or $\beta_1 \neq 0$ or $\beta_2 \neq 0$).

- Main hypothesis 3: what other factors might influence the sentiment in a given document ($\alpha_3 \neq 0$ or $\alpha_4 \neq 0$ or $\beta_3 \neq 0$ or $\beta_4 \neq 0$).

Statements and minutes

Table 4: Coherence of communication – SUR estimation

	InfSit_t	EconSit_t	InfSit_t δSwap	EconSit_t δSwap	InfSit_t δER	EconSit_t δER	InfSit_t EconUnc	EconSit_t EconUnc
α_0	0.024**	0.007	0.030**	0.007	0.025**	0.007	0.042**	0.067**
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.014)	(0.025)
α_1	0.430**	0.288**	0.513**	0.277**	0.431**	0.309**	0.417**	0.279**
	(0.097)	(0.0558)	(0.097)	(0.056)	(0.097)	(0.055)	(0.094)	(0.051)
α_2	0.154	0.071	0.054	0.125	0.158	0.023	0.184	0.038
	(0.147)	(0.140)	(0.148)	(0.145)	(0.146)	(0.139)	(0.143)	(0.129)
α_3	-0.060*	0.005	-0.082**	-0.016	-0.064**	0.020	-0.060**	-0.064
	(0.030)	(0.067)	(0.031)	(0.068)	(0.032)	(0.066)	(0.029)	(0.067)
α_4			2.380*	-3.009	-0.001	-0.004	-0.480	-1.302**
			(1.373)	(2.452)	(0.002)	(0.003)	(0.290)	(0.518)
β_0	0.016	0.033	-0.008	0.033	0.016	0.031	0.005	0.072
	(0.016)	(0.027)	(0.018)	(0.026)	(0.016)	(0.027)	(0.021)	(0.071)
β_1	0.041	0.969	0.310	0.986*	0.057	1.044*	0.001	0.944
	(0.266)	(0.580)	(0.280)	(0.574)	(0.272)	(0.588)	(0.268)	(0.577)
β_2	0.723**	0.159	0.690**	0.169	0.712**	0.158	0.713**	0.171
	(0.155)	(0.226)	(0.146)	(0.224)	(0.160)	(0.225)	(0.154)	(0.225)
β_3	-0.027	-0.311	0.017	-0.419	-0.026	-0.278	-0.035	-0.341
	(0.045)	(0.283)	(0.047)	(0.309)	(0.046)	(0.286)	(0.046)	(0.284)
β_4			-1.789**	-2.378	0.000	-0.001	0.282	-0.692
			(0.791)	(2.965)	(0.001)	(0.002)	(0.343)	(1.196)
Wald: (H0)								
$\alpha_1 = \beta_1 = 0$	19.64**	29.50**	27.97**	27.34**	19.65**	34.35**	20.05**	32.95**
$\alpha_2 = \beta_2 = 0$	22.06**	0.76	22.45**	1.31	20.25**	0.53	22.19**	0.66
$\alpha_3 = \beta_3 = 0$	5.02*	1.21	7.11**	1.89	5.14*	1.03	5.75*	2.31
N	30	30	30	30	30	30	30	30
$R^2(\alpha)$	0.468	0.579	0.489	0.599	0.473	0.605	0.505	0.650
$R^2(\beta)$	0.541	0.301	0.574	0.315	0.541	0.309	0.548	0.309

Note: Standard-deviation in parenthesis. (**) significant at 5%, (*) significant at 10%.

Statements and minutes

Table 4: Coherence of communication – SUR estimation

	InfSit_t	EconSit_t	InfSit_t δSwap	EconSit_t δSwap	InfSit_t δER	EconSit_t δER	InfSit_t EconUnc	EconSit_t EconUnc
α_0	0.024**	0.007	0.030**	0.007	0.025**	0.007	0.042**	0.067**
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.014)	(0.025)
α_1	0.430**	0.288**	0.513**	0.277**	0.431**	0.309**	0.417**	0.279**
	(0.097)	(0.0558)	(0.097)	(0.056)	(0.097)	(0.055)	(0.094)	(0.051)
α_2	0.154	0.071	0.054	0.125	0.158	0.023	0.184	0.038
	(0.147)	(0.140)	(0.148)	(0.145)	(0.146)	(0.139)	(0.143)	(0.129)
α_3	-0.060*	0.005	-0.082**	-0.016	-0.064**	0.020	-0.060**	-0.064
	(0.030)	(0.067)	(0.031)	(0.068)	(0.032)	(0.066)	(0.029)	(0.067)
α_4			2.380*	-3.009	-0.001	-0.004	-0.480	-1.302**
			(1.373)	(2.452)	(0.002)	(0.003)	(0.290)	(0.518)
β_0	0.016	0.033	-0.008	0.033	0.016	0.031	0.005	0.072
	(0.016)	(0.027)	(0.018)	(0.026)	(0.016)	(0.027)	(0.021)	(0.071)
β_1	0.041	0.969	0.310	0.986*	0.057	1.044*	0.001	0.944
	(0.266)	(0.580)	(0.280)	(0.574)	(0.272)	(0.588)	(0.268)	(0.577)
β_2	0.723**	0.159	0.690**	0.169	0.712**	0.158	0.713**	0.171
	(0.155)	(0.226)	(0.146)	(0.224)	(0.160)	(0.225)	(0.154)	(0.225)
β_3	-0.027	-0.311	0.017	-0.419	-0.026	-0.278	-0.035	-0.341
	(0.045)	(0.283)	(0.047)	(0.309)	(0.046)	(0.286)	(0.046)	(0.284)
β_4			-1.789**	-2.378	0.000	-0.001	0.282	-0.692
			(0.791)	(2.965)	(0.001)	(0.002)	(0.343)	(1.196)
Wald: (H0)								
$\alpha_1 = \beta_1 = 0$	19.64**	29.50**	27.97**	27.34**	19.65**	34.35**	20.05**	32.95**
$\alpha_2 = \beta_2 = 0$	22.06**	0.76	22.45**	1.31	20.25**	0.53	22.19**	0.66
$\alpha_3 = \beta_3 = 0$	5.02*	1.21	7.11**	1.89	5.14*	1.03	5.75*	2.31
N	30	30	30	30	30	30	30	30
$R^2(\alpha)$	0.468	0.579	0.489	0.599	0.473	0.605	0.505	0.650
$R^2(\beta)$	0.541	0.301	0.574	0.315	0.541	0.309	0.548	0.309

Note: Standard-deviation in parenthesis. (**) significant at 5%, (*) significant at 10%.

Internal consistency: minutes contain information from statements. But not one-to-one mapping!

Statements and minutes

Table 4: Coherence of communication – SUR estimation

	InfSit_t	EconSit_t	InfSit_t δSwap	EconSit_t δSwap	InfSit_t δER	EconSit_t δER	InfSit_t EconUnc	EconSit_t EconUnc
α_0	0.024**	0.007	0.030**	0.007	0.025**	0.007	0.042**	0.067**
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.014)	(0.025)
α_1	0.430**	0.288**	0.513**	0.277**	0.431**	0.309**	0.417**	0.279**
	(0.097)	(0.0558)	(0.097)	(0.056)	(0.097)	(0.055)	(0.094)	(0.051)
α_2	0.154	0.071	0.054	0.125	0.158	0.023	0.184	0.038
	(0.147)	(0.140)	(0.148)	(0.145)	(0.146)	(0.139)	(0.143)	(0.129)
α_3	-0.060*	0.005	-0.082**	-0.016	-0.064**	0.020	-0.060**	-0.064
	(0.030)	(0.067)	(0.031)	(0.068)	(0.032)	(0.066)	(0.029)	(0.067)
α_4			2.380*	-3.009	-0.001	-0.004	-0.480	-1.302**
			(1.373)	(2.452)	(0.002)	(0.003)	(0.290)	(0.518)
β_0	0.016	0.033	-0.008	0.033	0.016	0.031	0.005	0.072
	(0.016)	(0.027)	(0.018)	(0.026)	(0.016)	(0.027)	(0.021)	(0.071)
β_1	0.041	0.969	0.310	0.986*	0.057	1.044*	0.001	0.944
	(0.266)	(0.580)	(0.280)	(0.574)	(0.272)	(0.588)	(0.268)	(0.577)
β_2	0.723**	0.159	0.690**	0.169	0.712**	0.158	0.713**	0.171
	(0.155)	(0.226)	(0.146)	(0.224)	(0.160)	(0.225)	(0.154)	(0.225)
β_3	-0.027	-0.311	0.017	-0.419	-0.026	-0.278	-0.035	-0.341
	(0.045)	(0.283)	(0.047)	(0.309)	(0.046)	(0.286)	(0.046)	(0.284)
β_4			-1.789**	-2.378	0.000	-0.001	0.282	-0.692
			(0.791)	(2.965)	(0.001)	(0.002)	(0.343)	(1.196)
Wald: (H0)								
$\alpha_1 = \beta_1 = 0$	19.64**	29.50**	27.97**	27.34**	19.65**	34.35**	20.05**	32.95**
$\alpha_2 = \beta_2 = 0$	22.06**	0.76	22.45**	1.31	20.25**	0.53	22.19**	0.66
$\alpha_3 = \beta_3 = 0$	5.02*	1.21	7.11**	1.89	5.14*	1.03	5.75*	2.31
N	30	30	30	30	30	30	30	30
$R^2(\alpha)$	0.468	0.579	0.489	0.599	0.473	0.605	0.505	0.650
$R^2(\beta)$	0.541	0.301	0.574	0.315	0.541	0.309	0.548	0.309

Note: Standard-deviation in parenthesis. (**) significant at 5%, (*) significant at 10%.

Consistency across meetings: statements from previous meeting explains current statement on inflation.

Statements and minutes

Table 4: Coherence of communication – SUR estimation

	InfSit_t	EconSit_t	InfSit_t δSwap	EconSit_t δSwap	InfSit_t δER	EconSit_t δER	InfSit_t EconUnc	EconSit_t EconUnc
α_0	0.024**	0.007	0.030**	0.007	0.025**	0.007	0.042**	0.067**
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.014)	(0.025)
α_1	0.430**	0.288**	0.513**	0.277**	0.431**	0.309**	0.417**	0.279**
	(0.097)	(0.0558)	(0.097)	(0.056)	(0.097)	(0.055)	(0.094)	(0.051)
α_2	0.154	0.071	0.054	0.125	0.158	0.023	0.184	0.038
	(0.147)	(0.140)	(0.148)	(0.145)	(0.146)	(0.139)	(0.143)	(0.129)
α_3	-0.060*	0.005	-0.082**	-0.016	-0.064**	0.020	-0.060**	-0.064
	(0.030)	(0.067)	(0.031)	(0.068)	(0.032)	(0.066)	(0.029)	(0.067)
α_4			2.380*	-3.009	-0.001	-0.004	-0.480	-1.302**
			(1.373)	(2.452)	(0.002)	(0.003)	(0.290)	(0.518)
β_0	0.016	0.033	-0.008	0.033	0.016	0.031	0.005	0.072
	(0.016)	(0.027)	(0.018)	(0.026)	(0.016)	(0.027)	(0.021)	(0.071)
β_1	0.041	0.969	0.310	0.986*	0.057	1.044*	0.001	0.944
	(0.266)	(0.580)	(0.280)	(0.574)	(0.272)	(0.588)	(0.268)	(0.577)
β_2	0.723**	0.159	0.690**	0.169	0.712**	0.158	0.713**	0.171
	(0.155)	(0.226)	(0.146)	(0.224)	(0.160)	(0.225)	(0.154)	(0.225)
β_3	-0.027	-0.311	0.017	-0.419	-0.026	-0.278	-0.035	-0.341
	(0.045)	(0.283)	(0.047)	(0.309)	(0.046)	(0.286)	(0.046)	(0.284)
β_4			-1.789**	-2.378	0.000	-0.001	0.282	-0.692
			(0.791)	(2.965)	(0.001)	(0.002)	(0.343)	(1.196)
Wald: (H0)								
$\alpha_1 = \beta_1 = 0$	19.64**	29.50**	27.97**	27.34**	19.65**	34.35**	20.05**	32.95**
$\alpha_2 = \beta_2 = 0$	22.06**	0.76	22.45**	1.31	20.25**	0.53	22.19**	0.66
$\alpha_3 = \beta_3 = 0$	5.02*	1.21	7.11**	1.89	5.14*	1.03	5.75*	2.31
N	30	30	30	30	30	30	30	30
$R^2(\alpha)$	0.468	0.579	0.489	0.599	0.473	0.605	0.505	0.650
$R^2(\beta)$	0.541	0.301	0.574	0.315	0.541	0.309	0.548	0.309

Note: Standard-deviation in parenthesis. (**) significant at 5%, (*) significant at 10%.

Other factors: large share of sentences about inflation in minutes when there are bad news ($\alpha_3 < 0$).

Conclusion

- Improvement from LDA to hLDA: interpreting topics.
- hLDA Model captures the "structural break" of communication at Banco Central do Brasil in 2016.
- Indexes of economic situation seem to reflect the current state of the economy.
- Coherence in communication: minutes closely linked to statements; some effect from previous meetings.

THANK YOU!!!

Indexes of economic situation

- Reverse of the sentiment – Shapiro and Wilson (2019)
- Example: Situation on economic activity (Copom meeting 186, October 2014)

"5. (...) The PMI of the industrial sector, on its turn, indicates in September a **reversion** *of the* **expansion** *seen in August."*

Back

Quantitative data

- Used to evaluate the behavior of the economy with respect to changes in indexes of economic situation.
- Inflation: monthly and 12-month accumulated IPCA (Extended Consumer Price Index, official measure of the inflation target in Brazil).
- Inflation expectations: Focus survey, collected by Banco Central do Brasil; smoothed cumulative inflation for the next 12 months.
- Real variables: industrial production and retail sales.

Estimation

- Description of the Gibbs-Sampler:

  Defining additional hyperparameters η, m, and π and $Z \sim Discrete(\theta)$ as the distribution setting $Z = i$ with probability θ_i, documents in a corpus are assumed drawn from the following process in the hLDA model:

  - For each level $k \in T$ in the infinite tree,
    - Draw a topic $\beta_k \sim \text{Dirichlet}(\eta)$.
  - For each document $d \in \{1, 2, \ldots, D\}$,
    - Draw $c_d \sim nCRP(\gamma)$.
    - Draw a distribution over levels in the tree $\theta_d | \{m, \pi\} \sim GEM(m, \pi)$.
    - For each word,
    - Choose level $Z_{d,n} | \theta_d \sim Discrete(\theta_d)$.
    - Choose word $W_{d,n} | \{z_{d,n}, c_d, \beta\} \sim Discrete(\beta_{c_d}[z_{d,n}])$, which is parametrized by the topic in position $z_{d,n}$ on the path c_d.

Why not LDA?

- Simple LDA model with same number of topics of the third level of the hLDA tree.

Table 2: LDA model cloud from Copom minutes – Relevant words

Topic 0	**growth**	quarter	*price*	*projection*	*accumulated*
Topic 1	**growth**	sale	commercial	wholesale	*consumption*
Topic 2	*price*	*inflation*	*scenario*	*risk*	*Copom*
Topic 3	*price*	**index**	*inflation*	**increase**	variation
Topic 4	**rate**	**growth**	employment	**year**	**index**
Topic 5	*inflation*	*Copom*	*scenario*	*Monetary Policy*	**rate**
Topic 6	**year**	**increase**	employment	*expansion*	*industry*
Topic 7	*price*	*inflation*	*accumulated*	expected	variation
Topic 8	*price*	*inflation*	**index**	variation	**year**
Topic 9	**rise**	*price*	**rate**	*inflation*	*Copom*
Topic 10	Utilizationof InstalledCapacity	good	**rate**	*industry*	*previous*
Topic 11	**rate**	*Copom*	*scenario*	*economy*	*committee*
Topic 12	*Monetary Policy*	effect	*price*	import	should
Topic 13	**growth**	**year**	**increase**	**index**	production
Topic 14	*inflation*	*projection*	*scenario*	*price*	**rate**
Topic 15	good	producer	*consumption*	*inflation*	capital
Topic 16	*inflation*	trajectory	**rate**	*price*	**increase**
Topic 17	variation	*inflation*	average	core	*price*
Topic 18	*economy*	**index**	**growth**	*Monetary Policy*	*Economic Activity*
Topic 19	*inflation*	*price*	*meeting*	*Copom*	*scenario*
Topic 20	**year**	adjustment	*inflation*	**continuity**	**increase**
# Root	8	5	6	4	6
# Second level	10	9	9	12	9

Back

Why feature selection?

- Compare the root of the tree with two estimations: first, removing feature selection from data, but keeping estimated parameters constant; second, removing feature selection from data and estimating again the model.

Table: hLDA model root without feature selection

Baseline Model	No Feature Selection	No Feature Selection: New Parameters
rate	of (*de*)	that (*que*)
rise	the (*a*)	of (*do*)
year	at (*em*)	price
index	of (*do*)	of (*da*)
increase	of (*da*)	with (*com*)
continuity	and (*e*)	at (*no*)
last	the (*o*)	for (*pelo*)
growth	at (*no*)	international (*internacional*)

- Results without feature selection are consistent with Blei et al. (2010) with the roots including high-frequency but low-meaning words.

Back