A fuzzy topic modeling approach to legal corpora

Antonio Calcagnì and Arjuna Tuzzi

University of Padova

SDS Statistica e Data Science | Palermo, April 11-12, 2024

Introduction

Budget laws are fundamental tools for governments to address economic and financial aspects of countries. They drive **fiscal policies** and guide government spending.

The way they are composed often poses difficulties for administrative employees in **recognizing expenditure chapters** (e.g., education, healthcare, defense).

The classification can be improved if tagging with external interpretable texts is available (e.g., other laws or clauses).

A snippet of the budget law 178/2020

```
Stato.
  2. Al fine di dare attuazione a interventi in  materia  di  riforma
del  sistema  fiscale,  nello  stato  di  previsione  del   Ministero
dell'economia e delle finanze e' istituito un Fondo con una dotazione
di 8.000 milioni di euro per l'anno 2022 e di 7.000 milioni  di  euro
annui a decorrere dall'anno 2023, di cui una quota  non  inferiore  a
5.000 milioni di euro e non superiore  a  6.000  milioni  di  euro  a
decorrere  dall'anno  2022  e'  destinata  all'assegno  universale  e
servizi alla  famiglia.  I  predetti  interventi  sono  disposti  con
appositi provvedimenti normativi, a valere sulle risorse del Fondo di
cui al primo periodo.
  3. Al Fondo di cui al comma 2 sono destinate altresi'  a  decorrere
```

A snippet of the budget law 234/2021

2. Resta fermo quanto previsto dall'articolo 1, comma 6-bis, del decreto-legge 22 ottobre 2016, n. 193, convertito, con modificazioni, dalla legge 1° dicembre 2016, n. 225.
3. Sono riversate od acquisite all'entrata del bilancio dello

Introduction

The **dimension** of legal clauses and the use of highly **specialized language** [4, 6] present additional **challenges** when attempting to learn stable categories from the text.

Idea: Extract topics from budget laws by removing incoherent background topics that can make noisy the recognition of expenditure chapters.

Fuzzy topic modeling

Fuzzy topic models is an alternative approach to topic analysis which promises boosted performance in terms of classification accuracy, document clustering, and redundancy issues [7]. Several simulation studies have established those results when compared to standard approaches (e.g., LDA) [8, 9].

Fuzzy topic modeling

Based on the Latent Semantic Analysis rationale, the fuzzy topic model (**fLSA**) combines

- a dimensionality reduction technique (e.g., **SVD**) to alleviate sparsity and high-dimensionality of word tokens
- a fuzzy clustering method (e.g., **fuzzy c-means**) to extract K topics across the set of documents

to get document-topic $\mathbf{P}_{D|T_{(n\times K)}}$ and the word-topic $\mathbf{P}_{W|T_{(J\times K)}}$ probability matrices.

Fuzzy topic modeling
fLSA algorithm

<u>INPUT</u>:	(1)	K	(number of topics)
	(2)	$\mathbf{X}_{n \times J}$	(possibly weighted DTM)
	(3)	$\mathbf{p}_{D_{n \times 1}}$	(document probability vector)

INPUT:	(1)	K	(number of topics)
	(2)	$\mathbf{X}_{n\times J}$	(possibly weighted DTM)
	(3)	$\mathbf{p}_{D_{n\times 1}}$	(document probability vector)
DO:	(4)	$\mathbf{X}_{n\times J} \stackrel{\sim}{=} \mathbf{U}_{n\times Q}\boldsymbol{\Sigma}_{Q\times Q}\mathbf{V}^T_{Q\times J}$	(truncated-SVD, usually $Q=2$)
	(5)	$\widehat{\boldsymbol{\Xi}}_{n\times K} \leftarrow \min_{\boldsymbol{\Xi},\mathbf{c}} \; \mathcal{J}(\boldsymbol{\Xi}_{n\times K}, \mathbf{c}_K; \mathbf{U})$	(fuzzy c-means)
	(6)	$\widehat{\mathbf{P}}_{D\vert T_{(n\times K)}} \propto \; \widehat{\boldsymbol{\Xi}}_{n\times K} \circ \mathbf{p}_{D_{n\times 1}}\mathbf{1}^T_K$	(document-topic probability matrix)
	(7)	$\widehat{\mathbf{P}}_{W\vert T_{(J\times K)}} \propto \; \mathbf{X}^T_{n\times J}\,\widehat{\mathbf{P}}_{D\vert T_{n\times K}}$	(word-topic probability matrix)

Fuzzy topic modeling
fLSA algorithm

INPUT:	(1)	K	(number of topics)
	(2)	$\mathbf{X}_{n\times J}$	(possibly weighted DTM)
	(3)	$\mathbf{p}_{D_{n\times 1}}$	(document probability vector)
DO:	(4)	$\mathbf{X}_{n\times J} \stackrel{\sim}{=} \mathbf{U}_{n\times Q}\boldsymbol{\Sigma}_{Q\times Q}\mathbf{V}^{T}_{Q\times J}$	(truncated-SVD, usually $Q=2$)
	(5)	$\widehat{\boldsymbol{\Xi}}_{n\times K} \leftarrow \min_{\boldsymbol{\Xi},\mathbf{c}} \ \mathcal{J}(\boldsymbol{\Xi}_{n\times K}, \mathbf{c}_K; \mathbf{U})$	(fuzzy c-means)
	(6)	$\widehat{\mathbf{P}}_{D\vert T_{(n\times K)}} \propto \ \widehat{\boldsymbol{\Xi}}_{n\times K} \circ \mathbf{p}_{D_{n\times 1}} \mathbf{1}^{T}_{K}$	(document-topic probability matrix)
	(7)	$\widehat{\mathbf{P}}_{W\vert T_{(J\times K)}} \propto \ \mathbf{X}^{T}_{n\times J}\ \widehat{\mathbf{P}}_{D\vert T_{n\times K}}$	(word-topic probability matrix)
OUTPUT:	(9)	$\widehat{\mathbf{P}}_{D\vert T_{(n\times K)}}$	
	(10)	$\widehat{\mathbf{P}}_{W\vert T_{(n\times K)}}$	

Fuzzy topic modeling
fLSA algorithm

Depending on the problem being faced, the fLSA algorithm can be easily **generalized** by modifying the steps

(4) $\mathbf{X}_{n\times J} \overset{\sim}{=} \mathbf{U}_{n\times Q}\Sigma_{Q\times Q}\mathbf{V}^{T}_{Q\times J}$ (truncated-SVD)

(5) $\widehat{\mathbf{\Xi}}_{n\times K} \leftarrow \min_{\mathbf{\Xi},\mathbf{c}} \ \mathcal{J}(\mathbf{\Xi}_{n\times K}, \mathbf{c}_K; \mathbf{U})$ (fuzzy c-means)

to include other dimensionality reduction techniques (e.g., **NNMF**, **ISOMAP**) as well as different fuzzy clustering (e.g., fuzzy **k-medoids**, fuzzy-**SOM**).

Fuzzy topic modeling

fLSA algorithm

Depending on the problem being faced, the fLSA algorithm can be easily generalized by modifying the steps

$$(4) \qquad \mathbf{X}_{n\times J} \overset{\sim}{=} \mathbf{U}_{n\times Q}\Sigma_{Q\times Q}\mathbf{V}^{T}_{Q\times J} \qquad \text{(truncated-SVD)}$$

$$(5) \qquad \widehat{\mathbf{\Xi}}_{n\times K} \leftarrow \min_{\mathbf{\Xi},\mathbf{c}} \ \mathcal{J}(\mathbf{\Xi}_{n\times K}, \mathbf{c}_K; \mathbf{U}) \qquad \text{(fuzzy c-means)}$$

to include other dimensionality reduction techniques (e.g., NNMF, ISOMAP) as well as different fuzzy clustering (e.g., fuzzy k-medoids, fuzzy-SOM).

Although simple and flexible enough, fLSA **misses a consistent stochastic framework** to model $\mathbf{P}_{D|T}$ and $\mathbf{P}_{W|T}$.

Raw corpus

- $n = 2179$ clauses/documents
- $N = 235819$ word-tokens wt, $J = 11284$ word-types
- $TTR = 4.78\%$, hapax $= 37\%$

Preprocessing

- Basic steps lowercase conversion, punctuation, marks, symbols
- Multiword expressions
  - predefined list .g., *partita_iva*
  - 5-to-2 grams with highest PMI .g., *arma_carabinieri*
- Modal verbs e.g., *essere*, *avere*, *dovere*

Final corpus

- $n = 2179$ clauses/documents
- $N = 69903$ word-tokens $\cong$ 30% of the raw corpus
  $J = 2760$ word-types
- $TTR = 3.95\%$
- Document-Term-Matrix with TF-IDF schema

Methods

Repeated **10-fold Cross Validation**:

- $B = 25$ random repetitions for each fold
- Two commonly used coherence metrics:
  UMass and **UCI** implemented by [11]
- Number of topics $K \in \{2, 3, \ldots, 30\}$

Methods

Repeated **10-fold Cross Validation**:

- $B = 25$ random repetitions for each fold
- Two commonly used coherence metrics:
  UMass and **UCI** implemented by [11]
- Number of topics $K \in \{2, 3, \ldots, 30\}$
- Two techniques contrasted:
  - **fLSA**[7]
    using fuzzy k-means [5] and fast truncated-SVD algorithm [1]
  - **LDA** Latent Dirichlet Allocation **LDA** [2]
    using Gibbs sampler burn-in: 500; iterations: 5000

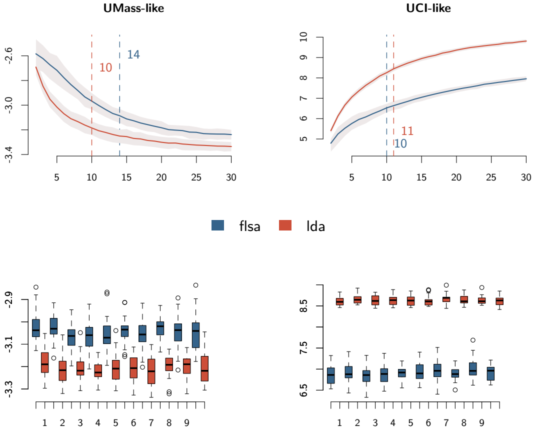
Methods

Repeated **10-fold Cross Validation**:

- $B = 25$ random repetitions for each fold
- Two commonly used coherence metrics:
  UMass and **UCI** implemented by [11]
- Number of topics $K \in \{2, 3, \ldots, 30\}$
- Two techniques contrasted: fLSA vs LDA
- **Training**: compute $\widehat{\mathbf{P}}_{W|T}$, detect the top 30 FREX words
- **Test**: use the top 30 FREX words, get the feature co-occurrence matrix, compute the coherence metrics

Results

Topics coherence



Notes:

Elbow of curves computed via Kneedle algorithm [10]

Gray areas represent the $q_{0.75} - q_{0.25}$ tolerance intervals

UMass and UCI metrics computed using the top 30 topic words

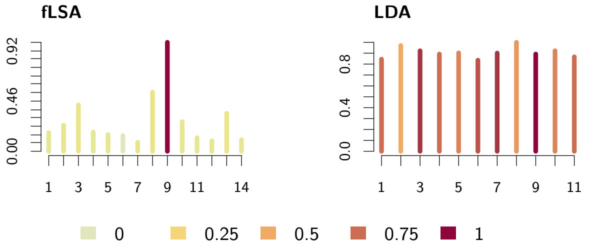
Retained number of topics:
14 fLSA, 11 LSA

Average coherence at the elbow

	fLSA	LDA
UMass	-3.21	-3.73
UCI	-1.43	-0.80

Results

Marginal topic distributions



<u>Notes</u>:

Vertical bars: Topic occurrence across documents normalized
Gradient color: Exclusive term ratio values closed to one → topic with exclusive words

Results

Marginal topic distributions

If compared to LDA, fLSA produces a simplified solution with topics being somewhat redundant. Thus, they can be further clustered into a less number of topics.

Indeed, the **Hellinger distance** of each topic from an overall underlying topic is lower for fLSA 0.271 than LDA 0.648.

Conversely, the average **topics overlap** is higher for fLSA 0.524 than LDA 0.272.

Topic contents fLSA

topic 9 `risorse regioni`
regione, produttive, farmaceut, servizi_sociali, fondo_solidarieta

topic 3 `risorse città`
umano, risorse, sicurezza, citta_metropol, spesa_compless

topic 2 `rilancio economia post-covid19`
rispetto_limite_spesa, pensione, nido, sicurezza, dl_104_2020

topic 7 `programmazione fiscale`
l_244_2020, tributi, spesa_complessiva, costi_fissi, dlgs_175_2016

topic 8 `politiche sociali`
assolvimento, enti_locali, politiche_sociali, educazione, mobilita_sostenibili

Results
Topic contents LDA

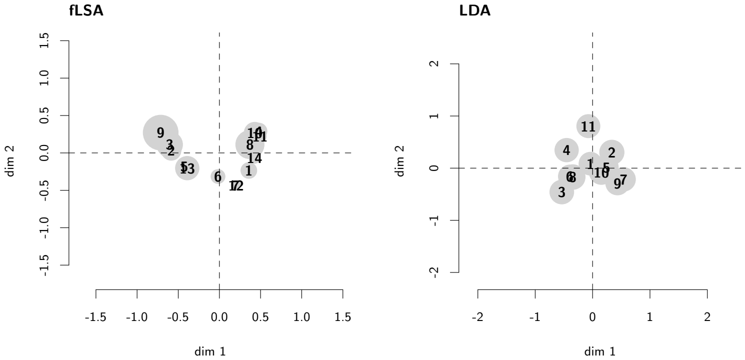
topic 9 `lavoro`
banca_italia, contratto, licenziamenti, ripartizione, l_26_2019

topic 3 `salari`
famiglia, cooperazione, amministrazioni, integr_salario, l_141_2019

topic 7 `consolidamento conti pubblici`
aziendale, lavorative, l_214_2011, lavoro_autonomo, medico

topic 6 `contratti`
contratti, contribuzione, ingegneri, alitalia, carriera

topic 1 `mille-proroghe`
libri, gestori, mercato, mezzogiorno, l_8_2020

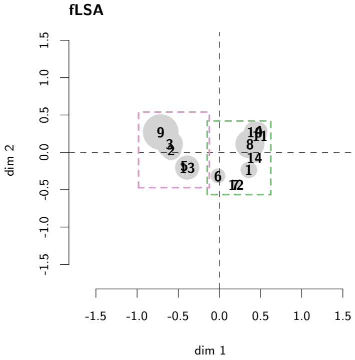


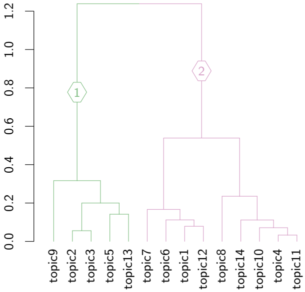
<u>Notes</u>:

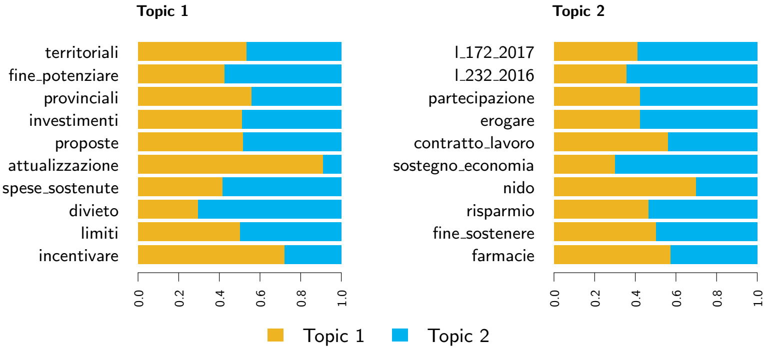
Intertopic cosine distance via MDS-based plot on dist($\widehat{\mathbf{P}}^{T}_{W\mid T}$).
The circle radius represents the marginal probability of the topic in log scale.

Results

Intertopic distance







Notes:

Highest FREX words and proportion of their occurrence colored horizontal bars.
Topic 1: `Future`; Topic 2: `Present situation`

Conclusions

- If used in fLSA, coherence measures suggest solutions with possibly redundant and overlapped topics

- Unlike LDA, fLSA tends to provide a simplified representation where only few topics need to retrieved to avoid redundancy
  $\rightarrow$ this could mask relevant facets of the corpus

Conclusions

- Further investigations are needed to compare fLSA with other techniques such as Correlated Topic Model (CTM)

Conclusions

- Further investigations are needed to compare fLSA with other techniques such as Correlated Topic Model (CTM)

- Further studies are also needed to go beyond the current implemented naive linear aggregator
$$\mathbf{P}_{D|T} \propto \ \boldsymbol{\Xi} \circ \mathbf{p}_D \mathbf{1}^T$$
used to integrate possibility $\boldsymbol{\Xi}$ with probability $\mathbf{p}_D$ (e.g., generalized Bayes rules [3])

[1] Baglama, J., Reichel, L., and Lewis, B. W.
irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices, 2022.
R package version 2.3.5.1.

[2] Blei, D. M., Ng, A. Y., and Jordan, M. I.
Latent dirichlet allocation.
Journal of machine Learning research 3, Jan (2003), 993–1022.

[3] Coletti, G., Gervasi, O., Tasso, S., and Vantaggi, B.
Generalized bayesian inference in a fuzzy context: From theory to a virtual reality application.
Computational Statistics & Data Analysis 56, 4 (2012), 967–980.

[4] Cortelazzo, M.
La lingua delle leggi italiane.
In *Il dovere costituzionale di farsi capire. A trent'anni dal Codice di stile*, M. E. Piemontese et al., Eds. Carocci, 2023, pp. 110–122.

[5] Ferraro, M., Giordani, P., and Serafini, A.
fclust: An r package for fuzzy clustering.
The R Journal 11 (2019).

[6] Garavelli, B. M.
Le parole e la giustizia.
Einaudi, 2001.

[7] Karami, A., Gangopadhyay, A., Zhou, B., and Kharrazi, H.
Fuzzy approach topic discovery in health and medical corpora.
International Journal of Fuzzy Systems 20 (2018), 1334–1345.

[8] Rijcken, E., Scheepers, F., Mosteiro, P., Zervanou, K., Spruit, M., and Kaymak, U.
A comparative study of fuzzy topic models and lda in terms of interpretability.
In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* (2021), IEEE, pp. 1–8.

[9] Rijcken, E., Zervanou, K., Spruit, M., Mosteiro, P., Scheepers, F., and Kaymak, U.
Exploring embedding spaces for more coherent topic modeling in electronic health records.
In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (2022), IEEE, pp. 2669–2674.

[10] Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B.
Finding a" kneedle" in a haystack: Detecting knee points in system behavior.
In *2011 31st international conference on distributed computing systems workshops* (2011), IEEE, pp. 166–171.

[11] Selivanov, D., Bickel, M., and Wang, Q.
text2vec: Modern Text Mining Framework for R, 2023.
R package version 0.6.4.

antonio.calcagni@unipd.it