LATENT BAYESIAN CLUSTERING FOR TOPIC MODELLING

Lorenzo Schiavon [1]

[1] Department of Economics, Ca' Foscari University of Venice, (e-mail: lorenzo.schiavon@unive.it)

ABSTRACT: The main objective in topic modelling is uncovering the underlying themes present in a corpus of text data. This process is generally constituted by two phases: (i) identifying the main words associated with each topic; (ii) grouping documents that contain similar sets of words together. In this work, we exploit recent advances in Bayesian factor models to represent the high-dimensional space of the observed words through a set of low-dimensional latent variables, and to jointly cluster the documents according to their distribution over such latent constructs. Groups and underlying constructs are interpreted as document topics and language concepts, respectively, with the number of such dimensions that is not required in advance. We apply the proposed approach to a data set of newspaper headlines.

KEYWORDS: Dirichlet process, infinite factor model,nonparametric Bayes, text data

1 Introduction

Nowadays, the digitalization is making available huge quantities of data, which require, on one hand, suitable automatized procedure to recognize, classify and organize such information and, on the other, allows the development and training of the algorithms to respond to such demand. In particular, it has become widespread in several field of studies and businesses the necessity of powerful tools that emulate human being capacity in extracting and summarizing the information expressed in text data. For instance, in political sciences, the use of automatic methods applied to large corpus of institutional reports and documents can represent a fast methodology to uncover and highlight the topics on which public organizations focus more. Indeed, topic modelling techniques aim to reveal the underlying semantic structures in large collections of documents and cluster them in topics. Such methodologies are based on vector space models that represent each document as a vector. In last decades, several techniques has been considered in topic modelling, including low rank decomposition of the document-term matrix, as non-negative matrix factorization, and probabilistic latent semantic analysis, as Latent Dirichlet Allocation.

Inspired by these approaches and exploiting recent advances in Bayesian nonparametric models, we propose a factor model able to jointly cluster the documents in topics and to recover a distinct set of latent concepts. A Bayesian nonparametric approach allows the number of topics can be inferred along with posterior distribution, as for the dimension of the latent semantic space. In addition, we exploit shrinkage prior to promote sparse structures on the low-rank matrices favouring parsimony and interpretation.

2 Latent mixture model in infinite factorization

We are interested in specifying a model able to provide a parsimonious representation of a document-term matrix along both matrix dimensions: on one hand clustering the documents in few groups, on the other reducing the term-space to a low-dimensional space of latent concepts. In view of this, we rely on the general class of latent factor mixture models (LAMB), proposed by Chandra *et al.* (2020), which is able to combine a dimensional reduction via factorization and a suitable use of Bayesian nonparametric framework to cluster the subjects. In particular, considering y_i the p-variate vector including a binary indication on the presence-absence of p terms in the document i, we adjust the LAMB specification by defining the probit model

$$y_i = \mathbb{1}(y_i^* > 0), \quad y_i^* \sim N_p(\Lambda\eta_i, I_p), \quad \eta_i \sim \sum_{k=1}^{\infty} \pi_k N_H(\mu_k, \Delta_k), \tag{1}$$

where Λ is a $p \times H$ matrix of factor loadings with $H \ll p$. The latent factor scores $\eta_i = (\eta_{i1}, \dots, \eta_{iH})^\top$ are modelled according to an infinite mixture of Gaussian distributions with $\{\pi_k\}_{k=1}^{\infty}$ following a stick-breaking representation

$$\pi_k = v_k \prod_{l<k}(1 - v_l), \qquad v_l \sim \text{Beta}(1, \alpha). \tag{2}$$

Then, clusters are determined by the membership of η_i to the posterior kernels.

Differently from Infinite Mixture of Factor Analyser (Murphy *et al.* , 2020) models, where observations are clustered over kernels with factorized covariance, LAMB defines the clustering over the low-dimensional space of latent constructs ensuring parsimony in the dimension of cluster-specific parameters. In addition, having a unique loadings matrix shared by the different clusters aids the interpretation of the latent factors as language concepts such that each of them explains the presence or absence of several terms belonging to the same semantic area.

In view of this, we carefully specify the prior of the loadings matrix Λ. We use the cumulative shrinkage process (CUSP) proposed by Legramanti *et al.* (2020), which exploits an over-parameterized model with an infinite number of factors and increasing probability of loadings being shrunk as the column index increases. In particular, we specify a spike and slab construction over the columns of Λ with spike probability mass ϖ_h increasing over the column index according to a stick-breaking construction. To allow for a local behaviour, we follow Schiavon *et al.* (2022) including a Bernoulli local scale $\phi_{jh} \sim \text{Ber}(c_p)$ in the variance of each element λ_{jh}. The mean $c_p \in (0,1)$ is set equal to a small positive offset to guarantee sparsity when p is large. Formally, we assume

$$\lambda_{jh} \sim N(0, \theta_h), \quad \theta_h = \phi_{jh}\rho_h(\vartheta_h - \theta_\infty) + \theta_\infty \tag{3}$$

$$\rho_h \sim \text{Ber}(1 - \varpi_h), \quad \vartheta_h^{-1} \sim \text{Ga}(a_\theta, b_\theta), \tag{4}$$

with θ_∞ a positive constant close to zero. Posterior distribution is approximated via MCMC exploiting an adaptive Gibbs sampling strategy.

3 Latent topic extraction

To illustrate the validity of our approach, we initially apply the model to a set of $n = 213$ newspaper sport headlines published by two newspapers of GEDI[*] in Autumn 2021. After removing the stopwords, we frame the headlines in a document-term matrix. Considering only the unigram and bigram which recur at least twice in the corpus, we obtain a binary matrix y registering the presence or absence of $p = 522$ distinct terms.

We follow Chandra *et al.* (2020) and Schiavon *et al.* (2022) to set the hyper-parameters. The offset c_p is set equal to the average word frequency in y. After running the MCMC algorithm, we recover meaningful posterior summary of the low-rank matrices Λ and η, we compute the posterior means only after having aligned the posterior samples. The algorithm estimates a nine latent factors model with 19 topics. Each factor can be interpreted as a concept characterized by high loadings in correspondence of terms belonging to a specific semantic area. Figure 1 reports a graph representation of the partial correlation matrix between the terms. Every term j, for $j = 1, \ldots, p$ is coloured according to its characterizing concept—i.e. $\text{argmax}_{h \in \{1,\ldots,9\}} \lambda_{jh}$—while the legend reports the term with the highest loading for every latent concept. As one may expect, different concepts refer to different sports or

[*] GEDI Gruppo Editoriale S.p.A. is an Italian media conglomerate based in Turin.

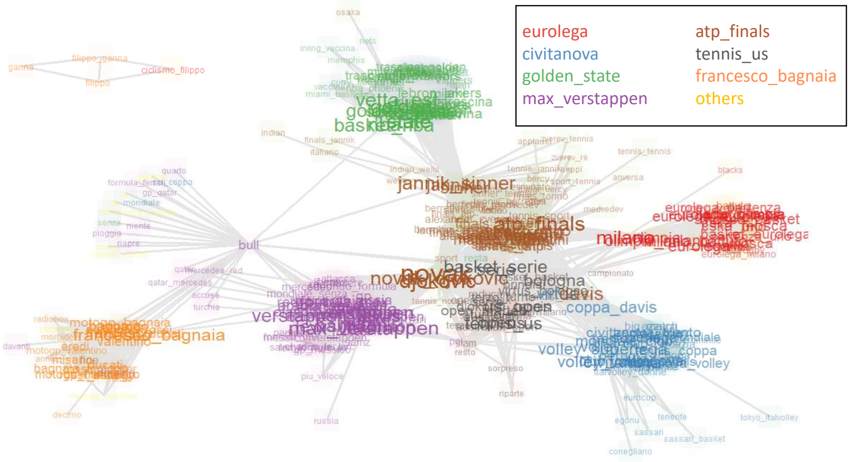

Figure 1. *Graphical representation based on posterior mean of partial correlation matrix. Terms are positioned using a FruchtermanReingold force-direct algorithm and coloured according to the highest loadings.*

competitions. Headlines are defined as weighted combination of such semantic concepts and grouped in topics with similar combination of concepts.

We aim to apply the same approach to documents and reports regarding health and ageing public plans published by Italian regional institutions to uncover the concepts and the themes on which Italian regions are focusing their efforts.

References

CHANDRA, NOIRRIT KIRAN, CANALE, ANTONIO, & DUNSON, DAVID B. 2020. Escaping the curse of dimensionality in Bayesian model based clustering. *arXiv preprint arXiv:2006.02700*.

LEGRAMANTI, SIRIO, DURANTE, DANIELE, & DUNSON, DAVID B. 2020. Bayesian cumulative shrinkage for infinite factorizations. *Biometrika*, **107**(3), 745–752.

MURPHY, KEEFE, VIROLI, CINZIA, & GORMLEY, ISOBEL CLAIRE. 2020. Infinite Mixtures of Infinite Factor Analysers. *Bayesian Analysis*, **15**(3), 937 – 963.

SCHIAVON, LORENZO, CANALE, ANTONIO, & DUNSON, DAVID B. 2022. Generalized infinite factorization models. *Biometrika*, **109**(3), 817–835.