A FLEXIBLE TOPIC MODEL

Ascari Roberto [1] and Giampino Alice [1]

[1] Department of Economics, Management, and Statistics, Piazza dell'Ateneo Nuovo, 1, Milan, University of Milano-Bicocca, Italy, (e-mail: roberto.ascari@unimib.it, a.giampino@campus.unimib.it)

ABSTRACT: In the last two decades, text modeling techniques have been used for various applications, including the analysis of topics in different text documents, where the aim is to provide a document representation in terms of topic distribution. This work aims to show some results on a generalization of the popular latent Dirichlet allocation model, with a particular focus on the clustering of text documents.

KEYWORDS: Dirichlet, latent variable, MCMC, mixture model, textual data.

1 Introduction

Let us consider a collection $\mathcal{C}$ of D text documents, commonly referred to as a "corpus". The d-th document can be thought of as a sequence $(w_{d,1},\ldots,w_{d,N_d})^\intercal$ of N_d words (i.e., $w_{d,n}$ represents the n-th word in the d-th document, $d = 1,\ldots,D$ and $n = 1,\ldots,N_d$). The set $\mathcal{V}$ of the V unique words appearing in the corpus represents a "vocabulary".

Topic modeling techniques assume that each word in a document is generated according to one among T possible topics. As a consequence, the d-th document can be represented through a vector $\boldsymbol{\theta}_d = (\theta_{d,1},\ldots,\theta_{d,T})^\intercal$, where $\theta_{d,t}$ represents the proportion of words in document d generated from topic t. Clearly, $\boldsymbol{\theta}_d$ belongs to the T-part simplex $\mathcal{S}^T = \{\boldsymbol{\theta} : \theta_t > 0, \sum_{t=1}^{T} \theta_t = 1\}$. Similarly, each topic is represented as a discrete probability distribution $\boldsymbol{\phi}_t$ over the vocabulary $\mathcal{V}$, $t = 1,\ldots,T$, thus $\boldsymbol{\phi}_t \in \mathcal{S}^V$. The most popular topic model is the latent Dirichlet allocation (LDA), introduced by Blei *et al.*, 2003, which supposes both the vectors $\boldsymbol{\theta}_d$ and ϕ_t following a Dirichlet distribution on $\mathcal{S}^T$ and $\mathcal{S}^V$, respectively. Thus,

$$\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha}),\ \ \boldsymbol{\alpha} \in \mathbb{R}_+^T \quad \text{and} \qquad \boldsymbol{\phi}_t \sim \text{Dir}(\boldsymbol{\beta}),\ \ \boldsymbol{\beta} \in \mathbb{R}_+^V.$$

Despite its popularity, the LDA suffers from the poor parameterization that the Dirichlet deserves for its covariance matrix. Then, the development of a more flexible technique seems to be a relevant issue.

2 The flexible LDA

In this section, we introduce a generalization of the LDA, namely the flexible LDA (FLDA). This model arises by assuming a flexible Dirichlet distribution (FD, Migliorati *et al.*, 2017) for each $\boldsymbol{\theta}_d$. The FD is a (structured) finite mixture model with Dirichlet components:

$$\mathrm{FD}(\boldsymbol{\theta};\boldsymbol{\alpha},\tau,\mathbf{p}) = \sum_{t=1}^{T} p_t \mathrm{Dir}(\boldsymbol{\theta};\boldsymbol{\alpha}+\tau\cdot\mathbf{e}_t),$$

where $\mathbf{p}\in\mathcal{S}^T$, $\tau>0$, and $\mathbf{e}_t$ is the null vector with the t-th element equal to 1. The additional parameters introduced by the mixture structure of the FD allow for a more flexible modelization of the covariance matrix, thus overcoming some limitations of the Dirichlet. It is noteworthy to mention that the FD includes the Dirichlet distribution as a special case if $\tau = 1$ and $p_t = \alpha_t/\alpha^+$ for $t = 1,\dots,T$, hence the FLDA model includes the LDA. The FD possesses several statistical properties, among which is the conjugacy to the multinomial scheme. Thus, if $\boldsymbol{\theta}_d \sim \mathrm{FD}(\boldsymbol{\alpha},\tau,\mathbf{p})$, then $\boldsymbol{\theta}_d$ given the corpus (i.e., the observed data) follows an FD distribution with updated parameters $\boldsymbol{\alpha}^*,\tau^*$, and $\mathbf{p}^*$.

To obtain estimates for the FLDA parameters $\{\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_D\}$ and $\{\boldsymbol{\phi}_1,\dots,\boldsymbol{\phi}_T\}$, we implement a collapsed Gibbs sampling (CGS), extending the approach proposed by Griffiths & Steyvers, 2004. The main difference with respect to a standard Gibbs sampling is that full conditionals are computed by marginalizing some parameters out. The estimates of the dropped parameters are computed by means of the conjugacy properties. To implement a CGS, we introduce a set of latent (i.e., unobservable) random variables $Z_{d,n}$ representing the topic label of the n-th word in the d-th document, $n = 1,\dots,N_d$, $d = 1,\dots,D$.

It is possible to show that the full conditionals, namely the probability that $\{Z_{d,n} = t\}$ (i.e., the word is assigned to topic t) given all the other topic assignments $\mathbf{z}_{-(d,n)}$, take the following form

$$p(Z_{d,n} = t|\mathbf{z}_{-(d,n)},\mathcal{C},\boldsymbol{\alpha},\tau,\mathbf{p},\boldsymbol{\beta}) \propto$$
$$\propto \frac{\left(\alpha_t + c^-_{t,d,\cdot}\right)\left(\beta_{v_{d,n}} + c^-_{t,\cdot,w_{d,n}}\right)}{\left(\beta^+ + c^-_{t,\cdot,\cdot}\right)} \cdot \left\{\sum_{h=1}^{T} p^*_{d,h} + p^*_{d,t}\left(\frac{\tau_t}{\alpha_t + c^-_{t,d,\cdot}}\right)\right\},$$

$t = 1,\dots,T$, where $p^*_{d,t} = p_t \dfrac{(\alpha_t+\tau)^{[c_{t,d,\cdot}]}}{(\alpha_t)^{[c_{t,d,\cdot}]}}$, $x^{[n]} = x(x+1)\cdot\dots\cdot(x+n-1)$ denotes the rising factorial function, and $w_{d,n}\in\mathcal{V}$ indicates which term of the vocabulary is associated with the n-th word in document d. Additionally, we define the quantities $c_{t,d,\cdot}$, $c_{t,\cdot,w}$, and $c_{t,d,\cdot}$ as summation over the proper index of the counts $c_{t,d,v} = \sum_{n=1}^{N_d} \mathbb{I}(z_{d,n} = t,\ w_{d,n} = v)$, the latter representing the number of times that word v is assigned to topic t in document d. Having the full conditionals, the CGS algorithm can be summarized by the following steps:

1. Initialize the vector $\mathbf{z}$ (randomly) and compute the counts $c_{t,d,v}^{(0)}$;
2. For $b = 1, \dots, B$:
   - For each word in the corpus:
     - sample a new topic $z_{d,n}^{(b)}$ for $w_{d,n}$ from $p(z)$;
     - update the counts $c_{t,d,v}^{(b)}$.
   - Use $\mathbf{z}^{(b)}$ to compute the estimates $\hat{\boldsymbol{\theta}}_d^{(b)}$ and $\hat{\boldsymbol{\phi}}_t^{(b)}$.

By having a sample of size B for the topic labels, namely $\mathbf{z}^{(b)}$, $b = 1, \dots, B$, and relying on the conjugacy properties, we can estimate $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_t$ as the mean of an FD and Dirichlet distributions with updated parameters, that is

$$\hat{\boldsymbol{\theta}}_d^{(b)} = \frac{\boldsymbol{\alpha} + \mathbf{c}_d^{(b)} + \tau \mathbf{p}_d^{*(b)} / p_+^{(b)}}{\alpha^+ + \tau + N_d} \quad \text{and} \quad \hat{\boldsymbol{\phi}}_t^{(b)} = \frac{\boldsymbol{\beta} + \mathbf{c}_t^{(b)}}{\beta^+ + c_{t,\cdot,\cdot}^{(b)}},$$

where $\mathbf{c}_d^{(b)} = (c_{1,d,\cdot}^{(b)}, \dots, c_{T,d,\cdot}^{(b)})^\intercal$ and $\mathbf{c}_t^{(b)} = (c_{t,\cdot,1}^{(b)}, \dots, c_{t,\cdot,V}^{(b)})^\intercal$.

3 Application: The Great Library Heist

During the night, a vandal broke into their professor's study and tore three books into single chapters. The single chapters are not labeled, so the professor is not able to cluster them so to restore the original books. In the following, we consider the $D = 166$ chapters as documents forming the corpus. We will consider $T = 3$ latent topics, each of them hopefully representing one of the destroyed books. Words in the corpus $\mathcal{C}$ compose a vocabulary $\mathcal{V}$ of $V = 16531$ unique terms. We run both the LDA and the FLDA models for $B = 5000$ iterations. Figure 1 displays the topic proportions $\boldsymbol{\theta}_d$ for all the documents, by conditioning on the true topic (i.e., the original book). We can note that both the LDA and FLDA models represent chapters from "Great Expectations" as mainly composed of terms arising from topic 1. The FLDA, thanks to the flexible covariance matrix of the FD, improves the LDA performance by providing more concentrated $\boldsymbol{\theta}_d$'s towards 0 or 1 than the LDA. Similar conclusions hold

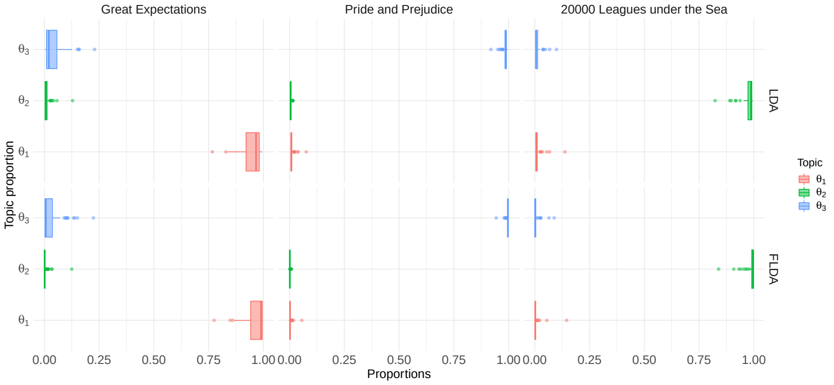

Figure 1. *Boxplots of the elements of* $\boldsymbol{\theta}_d$ *estimates by the LDA (upper panels) and the FLDA (bottom panels) conditioning on the true topic (i.e., the original book).*

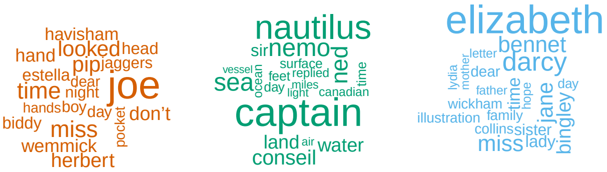

Figure 2. *Word clouds representing the 20 most probable words for each topic detected by the FLDA.*

true for chapters from "20000 Leagues Under the Sea" and "Pride and Prejudice", being characterized by high proportions of words from topics 2 and 3, respectively. Topics generated by the FLDA are represented by illustrating the 20 most probable words (Figure 2).

References

Blei, D.M., Ng, A.Y., & Jordan, M.I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

Griffiths, Thomas L., & Steyvers, Mark. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(SUPPL. 1), 5228 – 5235.

Migliorati, S., Ongaro, A., & Monti, G. S. 2017. A structured Dirichlet mixture model for compositional data: inferential and applicative issues. *Statistics and Computing*, **27**(4), 963–983.