

# Modeling topic dependencies in semantically coherent text spans with copulas

**Georgios Balikas** \*  
Université Grenoble-Alpes  
Computer Science Laboratory

**Hesam Amoualian**  
Université Grenoble-Alpes  
Computer Science Laboratory

**Marianne Clausel**  
Université Grenoble-Alpes  
Department of Statistics

**Eric Gaussier**  
Université Grenoble-Alpes  
Computer Science Laboratory

**Massih-Reza Amini**  
Université Grenoble-Alpes  
Computer Science Laboratory

## Abstract

The exchangeability assumption in topic models like Latent Dirichlet Allocation (LDA) often results in inferring inconsistent topics for the words of text spans like noun-phrases, which are usually expected to be topically coherent. We propose copulaLDA, that extends LDA by integrating part of the text structure to the model and relaxes the conditional independence assumption between the word-specific latent topics given the per-document topic distributions. To this end, we assume that the words of text spans like noun-phrases are topically bound and we model this dependence with copulas. We demonstrate empirically the effectiveness of copulaLDA on both intrinsic and extrinsic evaluation tasks on several publicly available corpora.

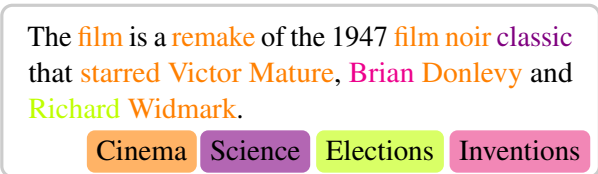
## 1 Introduction

Probabilistic topic models, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), are generative models that describe the content of documents by discovering the latent topics underlying them.

A limitation inherent from the bag-of-words representation in such state-of-the-art models concerns the independence assumption: given their topics, words are assumed to occur independently. While this exchangeability assumption greatly impacts the involved computations and, in particular, the calculations of the conditional probabilities, it is rather naive and unrealistic (Heinrich, 2005). As another limitation caused by the exchangeability assumption, the grouping of words in topically coherent spans, that is contiguous text spans like sentences, is lost.

On the other hand, text structure generally contains useful information that could be leveraged in inference process. Sentences or phrases, for instance, are by definition text spans complete in themselves that convey a concise statement. To better illustrate how text structure could help in topic identification, consider the example of Figure 1. It illustrates the topics inferred by LDA for the words (excluding stop-words) of a sentence drawn from a Wikipedia page. At the sentence level, one could argue that the sentence is generated by the “Cinema” topic since it discusses a film and its authors. LDA, however, fails and assigns several topics to the words of the sentence. Importantly, several of those topics like “Elections” and “Inventions” are unrelated. In finer text granularity, LDA also fails to assign consistent topics in noun-phrases like “film noir classic” and entities like “Brian Donlevy”. A binding mechanism among the topics of the words of a sentence, or a phrase, could have prevented those limitations and taking simple text structure into account would be beneficial.

Motivated by the previous example, we propose to incorporate text structure in the form of sentence or phrase boundaries as an intermediate structure in LDA. We plan to model this binding mechanism with copulas. Copulas have been found to be a flexible tool to model dependencies in the fields of



The film is a remake of the 1947 film noir classic that starred Victor Mature, Brian Donlevy and Richard Widmark.

Cinema Science Elections Inventions

Figure 1: Applying LDA on Wikipedia documents.

\* The author is also affiliated with Coffreo, Clermont Ferrand. The authors of the paper can be contacted at [firstname.lastname@imag.fr](mailto:firstname.lastname@imag.fr)

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

risk management and finance (Embrechts et al., 2002). They are a family of distribution functions that offer a flexible way to model the joint probability of random variables using only their marginals. This results in decoupling the marginal distributions by the underlying dependency. These properties make them appealing and some preliminary studies have started investigating their integration into different learning tasks (Wilson and Ghahramani, 2010; Tran et al., 2015; Amoualian et al., 2016).

The remainder of the paper is organized as follows: Section 2 presents the related work. The main contribution of this article is presented in Section 3, in which we propose to bind the latent topics that generate the words of a segment using copulas. We show that sampling word topics from copulas offers an elegant way to impose different levels and types of correlation between them. Section 4 then illustrates the behavior of *copulaLDA*, the copula-based version of LDA introduced in Section 3, while Section 5 concludes the paper.

## 2 Related Work

Despite the success that vector-space models (Salton et al., 1975) have enjoyed, they come with a number of limitations. We mention, for instance, their inability to model synonymy and polysemy and the sparse, high-dimensional induced representations. Many research studies have researched these problems, and Probabilistic Latent Semantic Analysis (Hofmann, 1999) was among the first attempts to model textual corpora using latent topics. In our work, we build on LDA (Blei et al., 2003), which is often used as a building block for topic models. In its context, the corpus is associated with a set of latent topics, and each document is associated with a random mixture of those topics. The words are assumed exchangeable, that is their joint probability is invariant to their permutation. Previous work proposed a variety of extensions to LDA in order to incorporate additional information such as class labels (Blei and McAuliffe, 2008) and temporal dependencies between stream documents (Wang et al., 2012). Here, our goal is to extend LDA by incorporating simple text structure in its generative and inference processes using copulas.

One may identify two lines of research to address the limitations due to the exchangeability assumption in LDA: extensions to account for the boundaries of text spans like sentences and extensions to account for the word order. With respect to the first line, (Wang et al., 2009) combine a unigram language model with topic models over sentences so that the latent topics are represented by sentences instead of terms. In (Griffiths et al., 2004), the authors investigate a combination of a topic model with a Hidden Markov Model (HMM). They assume that the HMM generates the words that handle the long-range dependencies (semantic dependencies) and the topic model the words that handle the short range dependencies (syntactic dependencies). Also, (Boyd-Graber and Blei, 2009) proposed the Syntactic Topic Model whose goal is to integrate the text semantics and the syntax in a non-parametric topic model. In another effort, (Zhu et al., 2006) propose *TagLDA*, where they replace the unigram word distributions by a factored representation that is conditioned on the topic and the part-of-speech tag of a term. Recently, (Balikas et al., 2016) introduced *senLDA*, that assumes that the terms occurring within a sentence are generated by the same topic. In our work here, we integrate part of the text structure in LDA by relying only on the boundaries of contiguous text spans like sentences, which can be obtained without deep linguistic analysis like the one required in the Syntactic Topic Model. Also, differently from *senLDA*, we do not restrict the words of the spans to be generated by the same topic. Instead, using copulas we pose correlations between those topics, which is more flexible.

The second line of research investigates how topic models can be extended to incorporate word order. In (Shafiei and Milios, 2006), the authors propose a four-level hierarchical structure where the latent topics of paragraphs are decided after performing a nested word-based LDA operation. In a similar context, (Wang et al., 2007) study how the word order in the form of n-grams can be leveraged to better capture a document's topical content. Their topical n-gram model extends LDA by determining unigram words and phrases based on context and assigning mixture of topics to both individual words and n-gram phrases.

Another interesting line of research studied the task of discovering and partitioning text in topically coherent spans. In (Du et al., 2010; Du et al., 2013) the authors rely on hierarchical Bayesian models to accomplish it. In this work, contrary to identifying such spans, we assume them to be topically coherent *a priori*, and we investigate how to leverage and incorporate this information to LDA.

Lately, there is an increasing interest over the integration of copulas in machine learning applications (Elidan, 2013) such as classification (Elidan, 2012) or structure learning (Liu et al., 2009). Interestingly, (Wilson and Ghahramani, 2010) have shown how to incorporate copulas in Gaussian processes in order to model the dependency between random variables with arbitrary marginals with a practical application on predicting the standard deviation of variables in the financial sector (volatility estimation). In another generic framework, (Tran et al., 2015) have shown the benefits of using copulas to model complex dependencies between latent variables in the general variational inference setting. The idea of using copulas with topic models was recently investigated in (Amoualian et al., 2016). In the context of document streams they proposed a topic model where the dependencies between the topic distributions of two consecutive documents are captured by copulas.

### 3 Integrating text structure to LDA using copulas

In this section we develop *copulaLDA* (hereafter *copLDA*), that extends LDA by integrating simple text structure in the model using copulas. We assume that the topics that generate the terms of coherent text spans are bound. A strong binding signifies high probability for the terms to have been generated by the same topic. Therefore, as we show, the conditional independence of topics given the per-document topic distributions does not hold. Before presenting the generative and inference processes of *copLDA*, we shortly discuss the idea of *coherent text spans*.

Each sentence is a coherent, meaningful segment of text and we consider them as coherent text spans in this study. However, each sentence can be further decomposed into smaller segments through syntactic analysis. Figure 2 illustrates the output of a shallow parsing step of the example sentence of Figure 1, generated using the Stanford Parser.<sup>1</sup> Among these different segments, noun phrases play a particular role as they are, for instance, at the basis of terminology extraction that aims at capturing concepts from a document. Noun phrases usually constitute a semantic unit, pertaining to a given concept related to few, related topics. For this reason, we also consider noun phrases as coherent text spans in this study. Another advantage of the two types of coherent text spans we consider (whole sentences and noun phrases) is that they can be easily extracted using shallow parsing techniques, and one needs not resort to complex syntactic analysis in practice.

*The film is a remake of the 1947 film noir classic that starred Victor Mature, Brian Donlevy and Richard Widmark.*

Figure 2: Shallow parsing using the Stanford Parser. Contiguous words in italics denote a noun-phrase.

#### 3.1 Copulas and random variables

Copulas are interesting because they separate the dependency structure of random variables from their marginals. Formally (Nelsen, 2007; Trivedi and Zimmer, 2007), a  $p$ -dimensional copula  $C$  is a  $p$ -variate distribution function with  $C : \mathbb{I}^p = [0, 1]^p \rightarrow [0, 1]$  whose univariate marginals are uniformly distributed on  $\mathbb{I}$  and  $C(u_1, \dots, u_p) = P(U_1 \leq u_1, \dots, U_p \leq u_p)$ . Copulas allow one to explicitly relate joint and marginal distributions, through Sklar's theorem (Sklar, 1959):

**Theorem 3.1** *Let  $F$  be a  $p$ -dimensional distribution function with univariate margins  $F_1, \dots, F_p$ . Let  $A_j$  denote the range of  $F_j$ . Then there exists a copula  $C$  such that for all  $(x_1, \dots, x_p) \in \mathbb{R}^p$*

$$F(x_1, \dots, x_p) = C(F_1(x_1), \dots, F_p(x_p)) \quad (1)$$

Furthermore, when  $F_1, \dots, F_p$  are all continuous, then  $C$  is unique.

As a result any multivariate distribution  $F$  can be decomposed into its marginals  $F_i, i \in \{1, \dots, p\}$  and a copula, allowing to study the multivariate distribution independently of the marginals. Sklar's theorem also provides a way of sampling multivariate distributions with a large number of random variables using copulas:  $F(x_1, \dots, x_p) = F(F_1^{-1}(u_1), \dots, F_p^{-1}(u_p)) = P[U_1 \leq u_1, \dots, U_p \leq u_p] = C(u_1, \dots, u_p)$ . Hence, to sample  $F$  it suffices to sample the dependence structure modeled by copulas and then transform

<sup>1</sup><http://nlp.stanford.edu/software/lex-parser.shtml>

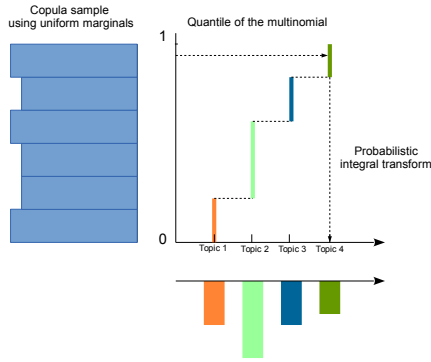


Figure 3: The transformation of a random variate to multinomial (or arbitrary) marginals. The arrows illustrate the generalized inverse; the histograms in  $y$  (resp.  $x$ ) axis depict the distributions of the initial (resp. transformed) samples.

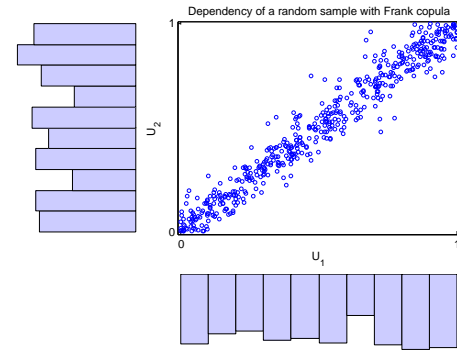


Figure 4: The positive correlation imposed to two random variates when sampling from a Frank copula with  $\lambda = 25$ . The histograms in  $x$  (resp.  $y$ ) axis show the distributions of each of the variates that generate the scatterplot.

the obtained sample in the marginals of interest using the probabilistic integral transform. We illustrate this transformation for one variable in Figure 3. Sampling the copula returns, for each variate, a sample as the one indicated in the histogram of the  $y$  axis. One can then transform the sample using the quantile ( $F^{-1}$ ) of an arbitrary marginal.

Before proceeding further, we visit some extreme conditions of dependence illustrating the respective copulas that model them: (1) *Independence*, which is a frequently assumed simplification in topic models and is obtained with  $\prod_{i=1}^p u_i$ , and (2) *Co-monotonicity*, which is the complete, positive correlation between the random variables  $u_p$ , obtained with  $\min(u_1, \dots, u_p)$ .

In the rest of our development we will be using a particular family of copulas, the Archimedean copulas. Archimedean copulas are widely used copulas and are defined with respect to a generator function  $\psi$ . They take the form:  $C(u_1, \dots, u_d) = \psi^{-1}(\psi(u_1) + \dots + \psi(u_d))$ . A special case of Archimedean copulas corresponds to Frank copulas, which are obtained by setting:  $\psi_\lambda(u) = \frac{-1}{\lambda} \log(1 - (1 - e^{-\lambda})e^{-u})$ . When  $\lambda \rightarrow 0$ , the Frank copula approaches the independency copula; when  $\lambda \rightarrow \infty$  it approaches the co-monotonicity copula. Hence, the Frank copula allows one to model all dependencies between complete independence to perfect dependence while varying  $\lambda$  from 0 to  $\infty$ . Therefore,  $\lambda$  can be seen as an additional hyper-parameter to be tuned or learned from the data. Figure 4 illustrates the positive dependence between two random variables sampled from a Frank copula with  $\lambda = 25$ . To sample from the Archimedean copulas, we rely on the algorithm proposed by (Marshall and Olkin, 1988), which was further improved in (McNeil, 2008; Hofert, 2011) and implemented in the R language (Hofert et al., 2011).

### 3.2 Extending LDA with copulas

As mentioned above, copulas provide a nice way to bind random variables. We are making use of them here to bind word-specific topics (the  $z$  variables in LDA) within coherent text spans, the rationale being that coherent text spans can not be generated by many different, uncorrelated topics. This leads us to the following generative model:

- For each topic  $k \in [1, K]$ , choose a per-word distribution:  $\phi_k \sim Dir(\beta)$ , with  $\phi_k, \beta \in \mathbb{R}^{|V|}$
- For each document  $d_i, i \in \{1, \dots, D\}$ :
  - Choose a per-document topic distribution:  $\theta_i \sim Dir(\alpha)$ , with  $\theta_i, \alpha \in \mathbb{R}^{|K|}$
  - Sample number of segments in  $d_i$ :  $S_i \sim Poisson(\xi)$ ;
  - For each segment  $s_{i,j}, j \in \{1, \dots, S_i\}$ :

- \* Sample number of words:  $N_{i,j} \sim \text{Poisson}(\xi_d)$ ;
- \* Sample topics  $\mathcal{Z}_{i,j} = (z_{i,j,1}, \dots, z_{i,j,N_{i,j}})$  from a distribution admitting  $\text{Mult}(1, \theta_i)$  as margins and  $C$  as copula;
- \* Sample words  $W_{i,j} = (w_{i,j,1}, \dots, w_{i,j,N_{i,j}})$ :  $w_{i,j,n} \sim \text{Mult}(1, \phi_{z_{i,j,n}})$ ,  $1 \leq n \leq N_{i,j}$ .

There are two main differences between *copLDA* and LDA. Firstly, the former assumes a hierarchical structure in the documents: the topics that generate the words in the coherent segments exhibit topical correlation, hence the conditional independence assumption between the terms of a segment given the document per-topic distribution ( $\theta_i$ ) no longer holds. Secondly, this topical correlation is modeled using copulas. Figure 5 provides the graphical model for *copLDA*. For clarity, we draw each word in a coherent segment  $S$  ( $w_1, \dots, w_N$ ) to make the dependencies explicit. Notice how the topics of those words depend on both the copula parameter  $\lambda$  and the per-document topic distribution  $\theta$ .

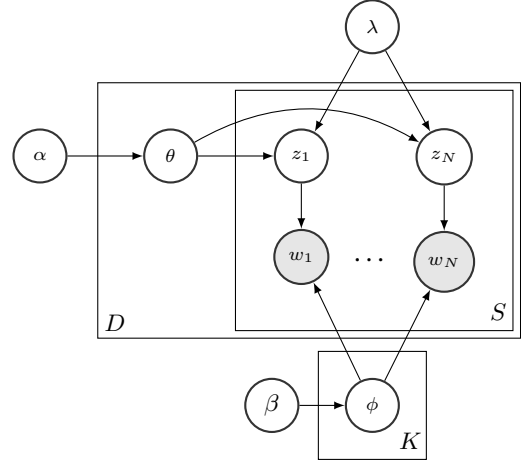


Figure 5: The *copLDA* generative model. We model the dependency between the topics underlying a segment with copulas.

The hyper-parameters  $\alpha$  and  $\beta$  correspond to priors of the model. Following (Blei et al., 2003), we assume them here to be symmetric and we fix them to  $\frac{1}{K}$ , with  $K$  the number of topics retained. The hyper-parameter  $\lambda$  is chosen after exploration of a grid of possible values, and is the same for the whole corpus. We choose the value that minimizes perplexity.

### 3.3 Inference with Gibbs sampling

The parameters of the above model, that are  $\phi, \theta$  and the topics of each segment  $\mathcal{Z}_{i,j} = (z_{i,j,1}, \dots, z_{i,j,N_{i,j}})$ , can be directly estimated through Gibbs sampling. Denoting  $\Omega$  and  $\Psi$  the count matrices such that  $\Omega = (\Omega_{i,k})$  (resp.  $\Psi = (\Psi_{k,v})$ ) represents the count of word belonging to topic  $k$  assigned to document  $d_i$  (resp. the count of word  $v$  being assigned to topic  $k$ ), the Gibbs updates for  $\theta$  and  $\phi$  are the same as the ones for the standard LDA model (Blei et al., 2003):

$$\theta_i \sim \text{Dir}(\alpha + \Omega_i) \quad \text{and} \quad \phi_k \sim \text{Dir}(\beta + \Psi_k) \quad (2)$$

The update for the variables  $z$  is obtained as follows:

$$\begin{aligned} p(\mathcal{Z}_{i,j} | \mathcal{Z}_{-i,j}, W, \Theta, \Phi, \alpha, \beta, \lambda) &= \frac{p(\mathcal{Z}_{i,j}, \mathcal{Z}_{-i,j}, W | \Theta, \Phi, \alpha, \beta, \lambda)}{p(\mathcal{Z}_{-i,j}, W | \Theta, \Phi, \alpha, \beta, \lambda)} = \\ \frac{p(\mathcal{Z}_{i,j}, W_{i,j} | \Theta, \Phi, \lambda) p(\mathcal{Z}_{-i,j}, W_{-i,j} | \Theta, \Phi, \lambda)}{p(W_{i,j} | \Theta, \Phi) p(\mathcal{Z}_{-i,j}, W_{-i,j} | \Theta, \Phi, \lambda)} &= \frac{p(\mathcal{Z}_{i,j}, W_{i,j} | \Theta, \Phi, \lambda)}{\sum_{\mathcal{Z}_{i,j}} p(\mathcal{Z}_{i,j}, W_{i,j} | \Theta, \Phi, \lambda)} = \\ \frac{p(W_{i,j} | \mathcal{Z}_{i,j}, \Phi) p(\mathcal{Z}_{i,j} | \Theta, \lambda)}{\sum_{\mathcal{Z}_{i,j}} p(W_{i,j} | \mathcal{Z}_{i,j}, \Phi) p(\mathcal{Z}_{i,j} | \Theta, \lambda)} &\sim p(W_{i,j} | \mathcal{Z}_{i,j}, \Phi) p(\mathcal{Z}_{i,j} | \Theta, \lambda) = p(\mathcal{Z}_{i,j} | \Theta, \lambda) \prod_{n=1}^{N_{i,j}} \phi_{w_{i,j,n}, z_{i,j,n}} \end{aligned} \quad (3)$$

where  $W, \Theta$  and  $\Phi$  stand for the whole parameter set of  $w, \theta$  and  $\phi$  and the probability outside the product in the last step admits a copula  $C_\lambda$  and  $\text{Mult}(1, \theta_i)$  as margins. As is standard in topic models, the notation  $-i, j$  means excluding the information for  $i, j$ . Note that in case where  $\lambda \rightarrow 0$ , the words of a segment become conditionally independent given the per-document distribution and one recovers the non collapsed Gibbs sampling updates of LDA.

From the expression of Eq. (3), a simple acceptance/rejection algorithm can be formulated: (1) Sample a random variable of pdf  $p(\mathcal{Z}_{i,j} | \Theta, \lambda)$  using copula, and, (2) Accept the sample with probability  $p(W_{i,j} | \mathcal{Z}_{i,j}, \Phi) = \prod_{n=1}^{N_{i,j}} \phi_{w_{i,j,n}, z_{i,j,n}}$ . Algorithm 1 summarizes the inference process.

### 3.4 Computational Considerations

As the values of  $\phi_{w_{i,j,1},z_{i,j,1}} \times \dots \times \phi_{w_{i,j,n},z_{i,j,n}}$  tend to be very low, the acceptance/rejection sampling step described above is very slow in practice (see below). We propose here to speed it up by considering, for each word  $w_{i,j,n}$  in a given segment, not the exact probability of  $z_{i,j,n}$ , but its mean (noted  $M$ ) over all the other words in the segment:

$$M(z_{i,j,n} | \mathcal{Z}_{-i,j}, W, \Theta, \Phi, \alpha, \beta, \lambda) = \sum_{w_{i,j,l}, l \neq n} \sum_{z_{i,j,l}, l \neq n} P(z_{i,j} | \mathcal{Z}_{-i,j}, W, \Theta, \Phi, \alpha, \beta, \lambda) \propto \phi_{w_{i,j,n}} \theta_{d,z_{i,j,n}}$$

as  $\sum_{w_{i,j,l}} \phi_{w_{i,j,l}} = 1$ . Note that the above form is a marginalization of  $P(z_{i,j} | \mathcal{Z}_{-i,j}, W, \Theta, \Phi, \alpha, \beta, \lambda)$  and thus defines a valid probability and a valid Gibbs sampler, even though on a joint distribution that slightly differs from the original one.

Figure 6 compares the perplexity scores achieved in 200 documents from the Wikipedia dataset ‘‘Wiki46’’ of Table 1 by the copLDA model, when considering noun-phrases as coherent spans, with and without rejection sampling. We repeat the experiment 10 times and also plot the standard deviation. We first note that approximating Algorithm 1 by ignoring the rejection sampling step results in slightly worse performance. On the other hand, without the rejection sampling, copLDA converges faster in terms of iterations. Furthermore, the cost in terms of running time of a single iteration is significantly smaller: for instance, for 30 iterations with rejection sampling, the algorithm needs almost 6 hours, that is 100 times more than the 3.5 minutes needed without the rejection sampling. Hence, in the rest of the study, for scaling purposes, we adopt the above mean approximation.

#### Algorithm 1: A Gibbs Sampling iteration for *copLDA*

```

Input: documents' words grouped in segments,  $\alpha, \beta, K$ , Copula family and its parameter  $\lambda$ 
//Initialize counters  $\Psi, \Omega$ 
for document  $d_i, i \in [1, D]$  do
  for segment  $s_{i,j} : j \in \{1, \dots, S_i\}$  do
    Draw a random vector  $U = (U_1, \dots, U_{N_{i,j}})$  that admits a copula  $C_\lambda$ 
    do /* If the mean approximation is used, the loop is done once, ignoring the acceptance condition */
      for words  $w_{i,j,k}, k \in [1, W_{N_{i,j}}]$  in  $s_{i,j}$  do
        Decrease counter variables  $\Psi, \Omega$ 
        Get  $z_{i,j,k}$  by transforming  $U_k$  to Mult. marginals with the generalized inverse
        Assign topic  $z_{i,j,k}$  to  $w_{i,j,k}$ 
        Increase counters  $\Psi, \Omega$ 
      end
    while Accept the new segment topic assignments with probability  $\phi_{w_{i,j,1},z_{i,j,1}} \times \dots \times \phi_{w_{i,j,n},z_{i,j,n}}$ 
  end
end

```

## 4 Experimental study

**Models** In our experiments, we compare the following topic models: (1) *copLDA<sub>sen</sub>* that considers sentences as coherent segments, (2) *copLDA<sub>np</sub>* that considers noun-phrases as coherent segments, (3) LDA as proposed in (Blei et al., 2003) using the collapsed Gibbs sampling inference of (Griffiths and Steyvers, 2004), and (4) *senLDA* described in (Balikas et al., 2016) using its public implementation. For *copLDA<sub>x</sub>* models, we use the Frank copula which was reported to obtain the best performance in similar tasks (Amoualian et al., 2016) and was also found to achieve the best performance in our local validation settings compared to Gumbel and Clayton copulas. We have implemented the models using Python;<sup>2</sup> for sampling the Frank copulas we used the R *copula* package (Hofert et al., 2011) and rPY.<sup>3</sup> As mentioned in Section 3.2,  $\lambda$  is set to 2 for *copLDA<sub>sen</sub>* and to 5 for *copLDA<sub>np</sub>* (values which we found to perform well in every dataset we tried). Furthermore, the hyper-parameters  $\alpha$  and  $\beta$  where set to  $1/K$ , where  $K$  is the number of topics, which was selected from  $\{50, 100, 200, 300, 400\}$  for each dataset. For the shallow parsing step, required for *copLDA<sub>np</sub>*, we used the Stanford Parser (Klein and Manning, 2003). The text pre-processing steps performed are: lower-casing, stemming using the Snowball Stemmer and removal of numeric strings.

<sup>2</sup>The models used in this paper are available for research purposes at <https://github.com/balिकासg/topicModelling>.

<sup>3</sup><https://pypi.python.org/pypi/rpy2>

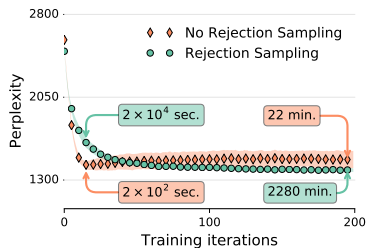


Figure 6: The effect of rejection sampling in efficiency and perplexity performance.

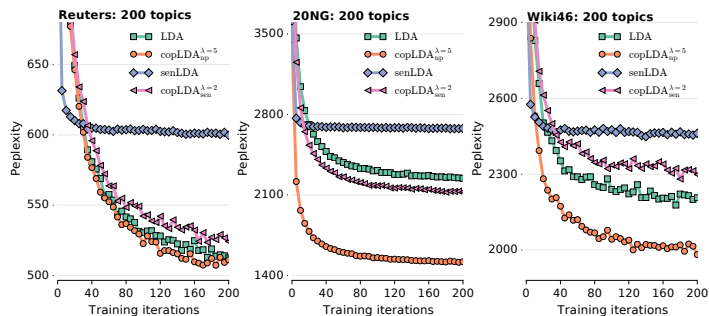


Figure 7: The perplexity curves of the investigated models for 200 Gibbs sampling iterations and different datasets.

	Basic Statistics				Perplexity Scores				Classification (MiF <sub>1</sub> ) scores			
	Docs.	N	V	Classes	senLDA	copLDA <sub>sen</sub>	LDA	copLDA <sub>np</sub>	senLDA	copLDA <sub>sen</sub>	LDA	copLDA <sub>np</sub>
20NG	19,056	1.7M	75.4K	20	2636	2083	2200	<b>1483</b>	0.5622	0.6328	0.6246	<b>0.6490</b>
TED	1,096	1.16M	30.4K	15	2099	1812	1805	<b>1775</b>	0.4612	0.4678	0.4633	<b>0.4764</b>
PubMed	5498	1.09M	28.7K	50	1601	1385	1384	<b>1085</b>	0.6666	<b>0.7525</b>	0.7406	0.7431
Reuters	10,788	875K	21.4K	90	579	512	501	<b>499</b>	0.7504	0.7692	<b>0.7893</b>	0.7851
Wiki15	1,198	162K	13.4K	15	2988	2766	2640	<b>2397</b>	0.6920	0.7230	0.74	<b>0.7403</b>
Wiki37	2,459	317K	19.7K	37	3103	2871	2711	<b>2395</b>	0.5717	0.6053	<b>0.6447</b>	0.6220
Wiki46	3,657	478K	23.4K	46	2220	2280	2135	<b>1978</b>	0.5326	0.6170	<b>0.6599</b>	0.6326
Austen	5,262	170K	6.3K	-	1110	898	<b>798</b>	805	-	-	-	-

Table 1: The basic statistics, the perplexity and the classification scores of the datasets used.

**Datasets** We have used the following publicly available data collections to test the performance of the topic models: (1) 20NG (20 news groups), which is a standard text dataset for such tasks as provided by (Bird et al., 2009), (2) Reuters (Reuters-21578, the “ModApte” version), also discussed in (Bird et al., 2009), (3) TED, that is transcriptions of TED talks released in the framework of the International Workshop on Spoken Language Translation 2013 evaluation campaign<sup>4</sup> (we have merged the train, development and test parts and we selected the transcriptions with at least one associated label among the 15 most common in the data<sup>5</sup>), (4) Wiki<sub>x</sub>, with  $x \in \{15, 37, 46\}$  and PubMed, both excerpts<sup>6</sup> from the Wikipedia dataset of (Partalas et al., 2015) and the PubMed dataset of (Tsatsaronis et al., 2015) used in (Balikas et al., 2016), and (5) “Austen”, where we concatenated three books<sup>7</sup> written by Jane Austen, available from the Gutenberg project (each paragraph is considered as a document). Table 1 presents some basic statistics for these datasets.

**Manual inspection of the topics** We begin by comparing LDA and  $copLDA_{np}$ . For presentation purposes, we train the two topic models using the Wiki<sub>47</sub> dataset with 10 topics and we illustrate the top-10 words learned for each topic by the two models in Table 2. As one can note, since the two models have been trained on the same data with the same training parameters, the identified topics are very similar. This said,  $copLDA_{np}$  manages to produce arguably better topics. This is for example the case for the topic “Birth”; although both models assign high probability to words like “born” and “american” due to the content of the dataset,  $copLDA_{np}$  manages to identify several words corresponding to months which makes the topic more thematically consistent and easier to interpret compared to its LDA counterpart. In the same line, Table 3 visualizes the inferred topics for parts of the Wiki<sub>47</sub> dataset. Notice here that given the topic interpretations of Table 2, both models manage to identify intuitive topics. Note however how in most of the cases the text structure information used by  $copLDA_{np}$  helps to obtain consistent topics to generate noun-phrases like “crime thriller film” and “raspy voice”, a consistency that LDA is lacking.

**Intrinsic evaluation: perplexity** We present in Table 1 the perplexity scores achieved by the 4 models in

<sup>4</sup><http://workshop2013.iwslt.org/59.php>

<sup>5</sup>Technology, Culture, Science, Global Issues, Design, Business, Entertainment, Arts, Politics, Education, Art, Creativity, Health, Biology and Music.

<sup>6</sup><https://github.com/balikasg/topicModelling/tree/master/data>

<sup>7</sup>We used the books: Emma, Persuasion, Sense. We considered each paragraph as a document.



Profession	Science	Books	Art	Cinema	Places	Music	Birth	Elections	Inventions
profession	univers	book	art	film	state	record	born	elect	california
world	research	new	new	televis	unit	music	american	canadian	plant
football	scienc	work	work	role	us	band	known	parti	use
wrestl	professor	american	paint	appear	township	album	best	member	invent
play	work	publish	york	also	school	song	actress	liber	flower
born	institut	time	american	actor	univers	also	decemb	minist	compani
american	award	author	artist	born	serv	produc	june	hous	north
championship	prize	also	museum	play	war	releas	april	canada	patent
team	born	year	painter	seri	nation	new	juli	serv	inventor
first	receiv	york	studi	star	build	singer	januari	conserv	found
known	univers	book	art	film	township	record	play	elect	work
wrestl	research	new	new	born	state	music	football	canadian	first
born	scienc	american	york	televis	counti	band	born	serv	year
world	professor	author	paint	role	us	album	american	parti	photograph
profession	work	publish	american	actor	california	song	tour	member	design
american	institut	novel	work	appear	michigan	also	golf	liber	state
name	born	time	artist	also	plant	singer	year	hous	new
wrestler	prize	also	painter	seri	civil	releas	profession	minist	use
best	studi	writer	museum	actress	popul	produc	first	state	also
championship	award	magazin	born	american	flower	american	season	born	build

Table 2: The top-10 words of copLDA (upper half) and LDA (lower half) in the Wiki46 dataset.

*Kiss of Death* is a 1995 *crime thriller film* starring *David Caruso Samuel L. Jackson* and *Nicolas Cage*. *The film* is a *very loosely based remake* of the *1947 film noir classic* of the same name that starred *Victor Mature, Brian Donlevy* and *Richard Widmark*.

*Bertram Stern* (born 3 October 1929) is *an American fashion and celebrity portrait photographer*.

*Dana Hill* (born *Dana Lynne Goetz* in *Los Angeles, California*; *May 6, 1964 - July 15, 1996*) was *an American actress and voice actor* with a *raspy voice* and *childlike appearance*, which *allowed* her to *play adolescent roles* well into her 20s.

*Kiss of Death* is a 1995 *crime thriller film* starring *David Caruso Samuel L. Jackson* and *Nicolas Cage*. *The film* is a *very loosely based remake* of the *1947 film noir classic* of the same name that starred *Victor Mature, Brian Donlevy* and *Richard Widmark*.

*Bertram Stern* (born 3 October 1929) is *an American fashion and celebrity portrait photographer*.

*Dana Hill* (born *Dana Lynne Goetz* in *Los Angeles, California*; *May 6, 1964 - July 15, 1996*) was *an American actress and voice actor* with a *raspy voice* and *childlike appearance*, which *allowed* her to *play adolescent roles* well into her 20s.

Table 3: The discovered topics underlying the words of example documents for LDA (left) and copLDA (right). The parts of the documents in italics indicate the noun-phrases obtained by the Stanford Parser. The text colours refer to the topics described in Table 2.

each of the datasets we examined. We split each dataset in two parts with 80%/20% of the documents: we use the former for learning the model and the second for calculating the perplexity scores. First note that  $copLDA_{np}$  achieves the lowest scores in most of the datasets. LDA is the second best performing model, whereas the third one is  $copLDA_{sen}$ . We believe that the difference between  $copLDA_{sen}$  and  $copLDA_{np}$  stems from the fact that perplexity is an evaluation measure that is calculated on the basis of words. Hence, considering sentences as coherent spans whose topics are bound results in less flexibility and this is reflected in higher perplexity scores. However, using copulas results in more flexibility than assigning the same topic in each term of the sentence which is illustrated in the performance difference between  $copLDA_{sen}$  and  $senLDA$ . The former being more flexible, due to the copulas, performs better. In the same line, Figure 7 illustrates the perplexity curves of the hold-out documents for the four models on three of the datasets of Table 1 for 200 Gibbs sampling iterations. Note that  $senLDA$  is the model with the fastest convergence rate with respect to the number of Gibbs iterations. On the other hand, LDA,  $copLDA_{sen}$  and  $copLDA_{np}$  require the same number of iterations, which depends on the dataset.  $copLDA_{np}$  manages to achieve the lowest perplexity scores: notice its steep curves in the first iterations.

**Extrinsic evaluation: text classification** To further highlight the merits of  $copLDA$ , we also present in Table 1 the classification results for the datasets used. The reported scores are the averages of 10-fold cross-validation. We use the per-document topic distributions as classification features fed to Support Vectors Machines (SVMs). We have used the implementation of (Pedregosa et al., 2011) with  $C = 1$  for the SVM regularization parameter. For the multi-label datasets (TED and PubMed) we employed one-versus-rest: the SVMs return every category with a positive distance from the separating hyper-planes. As one can note,  $copLDA_{np}$  and LDA achieve the highest MiF scores in most of the datasets, without a clear advantage to one vs the other. Binding the topics of sentence words with copulas improves over the



results of *senLDA*: *copLDA*<sub>sen</sub> performs only slightly worse than LDA and *copLDA*<sub>np</sub> on most datasets and outperforms them, only slightly again, on one dataset.

## 5 Conclusions

We proposed *copLDA* that extends LDA to incorporate the topical dependencies within sentences and noun-phrases using copulas. We have shown empirically the advantages of considering text structure and incorporating it in LDA with copulas. In our future work we plan to integrate procedures to learn the  $\lambda$  parameter of Frank copulas and to investigate ways to model not only dependencies within text segments like noun-phrases, but also dependencies between such segments with nested copulas.

## Acknowledgements

The authors would like to thank the anonymous reviewers for their useful comments. This work is partially supported by the CIFRE N 28/2015.

## References

- H. Amoualian, M. Clausel, E. Gaussier, and M.R. Amini. 2016. Streaming-LDA: A Copula-based Approach to Modeling Topic Dependencies in Document Streams. In *Proceedings of the 22th International ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM.
- G. Balikas, M.R. Amini, and M. Clausel. 2016. On a Topic Model for Sentences. *Journal of CoRR*, abs/1606.00253.
- S. Bird, E. Klein, and E. Loper. 2009. *Natural Language Processing with Python*. " O'Reilly Media, Inc."
- D.M. Blei and J.D. McAuliffe. 2008. Supervised Topic Models. In *Advances in Neural Information Processing Systems 20 NIPS*, pages 121–128. Curran Associates, Inc.
- D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning*, 3:993–1022, March.
- J. Boyd-Graber and D.M. Blei. 2009. Syntactic Topic Models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21 NIPS*, pages 185–192. Curran Associates, Inc.
- L. Du, W. Buntine, and H. Jin. 2010. A Segmented Topic Model Based on the Two-parameter Poisson-Dirichlet Process. *Journal of Machine learning*, 81(1):5–19.
- L. Du, W.L. Buntine, and M. Johnson. 2013. Topic Segmentation with a Structured Topic Model. In *Proceedings of HLT-NAACL*, pages 190–200.
- G. Elidan. 2012. Copula Network Classifiers (CNCs). In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pages 346–354.
- G. Elidan. 2013. Copulas in Machine Learning. In *Advances in Copulae in mathematical and quantitative finance*, pages 39–60. Springer.
- P. Embrechts, A. McNeil, and D. Straumann. 2002. Correlation and Dependence in Risk Management: Properties and Pitfalls. *Journal of Risk management: value at risk and beyond*, pages 176–223.
- T.L. Griffiths and M. Steyvers. 2004. Finding Scientific Topics. *Journal of the National Academy of Sciences*, 101(suppl 1):5228–5235.
- T.L. Griffiths, D.M. Steyvers, M. Blei, and J.B. Tenenbaum. 2004. Integrating Topics and Syntax. In *Proceedings of Neural Information Processing Systems 17 NIPS*, volume 4, pages 537–544.
- G. Heinrich. 2005. Parameter Estimation for Text Analysis. Technical report, Technical report.
- M. Hofert, M. Mächler, et al. 2011. Nested Archimedean Copulas Meet R: The nacopula Package. *Journal of Statistical Software*, 39(9):1–20.
- M. Hofert. 2011. Efficiently Sampling Nested Archimedean Copulas. *Journal of Computational Statistics & Data Analysis*, 55(1):57–70.

- T. Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99*, pages 50–57, New York, NY, USA. ACM.
- D. Klein and C.D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.
- H. Liu, J. Lafferty, and L. Wasserman. 2009. The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. *Journal of Machine Learning Research*, 10:2295–2328.
- A.W. Marshall and I. Olkin. 1988. Families of Multivariate Distributions. *Journal of the American Statistical Association*, 83(403):834–841.
- A.J. McNeil. 2008. Sampling Nested Archimedean Copulas. *Journal of Statistical Computation and Simulation*, 78(6):567–581.
- R.B. Nelsen. 2007. *An Introduction to Copulas*. Springer Science & Business Media.
- I. Partalas, A. Kosmopoulos, N. Baskiotis, T. Artieres, G. Paliouras, E. Gaussier, I. Androutopoulos, M.R. Amini, and P. Galinari. 2015. LSHTC: A Benchmark for Large-Scale Text Classification. *Journal of CoRR*, abs/1503.08581, march.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- G. Salton, A. Wong, and C. Yang. 1975. A Vector Space Model for Automatic Indexing. *Journal of Communications of the ACM*, 18(11):613–620.
- M. Shafiei and E. Milios. 2006. Latent Dirichlet Co-clustering. In *Proceedings of the Sixth International Conference on Data Mining, ICDM '06*, pages 542–551, Washington, DC, USA. IEEE Computer Society.
- M. Sklar. 1959. *Fonctions de Répartition à n Dimensions et Leurs Marges*. Université Paris 8.
- D. Tran, D.M. Blei, and E.M. Airoldi. 2015. Copula Variational Inference. In *Proceedings of Neural Information Processing Systems 28 NIPS*, pages 3564–3572.
- P.K. Trivedi and D.M. Zimmer. 2007. *Copula Modeling: An Introduction for Practitioners*. Now Publishers Inc.
- G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M.R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *Journal of BMC bioinformatics*, 16(1):1.
- Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval. In *Proceedings of ICDM 2007. Seventh IEEE International Conference on*, pages 697–702. IEEE.
- D. Wang, S. Zhu, T. Li, and Y. Gong. 2009. Multi-Document Summarization using Sentence-based Topic Models. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 297–300. Association for Computational Linguistics.
- Y. Wang, E. Agichtein, and M. Benzi. 2012. TM-LDA: Efficient Online Modeling of Latent Topic Transitions in Social Media. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12*, pages 123–131, New York, NY, USA. ACM.
- A.G. Wilson and Z. Ghahramani. 2010. Copula Processes. In *Advances in Neural Information Processing Systems 23 NIPS*, pages 2460–2468. Curran Associates, Inc.
- X. Zhu, D.M. Blei, and J. Lafferty. 2006. Taglda: Bringing Document Structure Knowledge into Topic Models. Technical report, Technical Report TR-1553, University of Wisconsin.