Abstract

Probabilistic topic models in low data resource scenarios are faced with less reliable estimates due to sparsity of discrete word co-occurrence counts, and do not have the luxury of retraining word or topic embeddings using neural methods. In this challenging resource constrained setting, we introduce an automatic trade-off between the discrete and continuous representations via an adaptive mixture coefficient, which places greater weight on the discrete representation when the corpus statistics are more reliable. The adaptive mixture coefficient takes into account global corpus statistics, and the uncertainty in each topic’s continuous distribution. Our approach outperforms the fully discrete, fully continuous, and static mixture model on topic coherence in low resource monolingual and multilingual settings.

1 Introduction

In topic modeling, the goal is to learn key themes in a corpus for exploratory document analysis (Boyd-Graber et al., 2017). Latent Dirichlet Allocation (LDA; Blei et al. (2003)) has been the bedrock for topic modeling and remains a hard to beat baseline for the general scenario which models with only words and documents.

We examine topic modeling in a low resource data setting (Hao et al., 2018), which has seen little attention but is commonly encountered in the digital humanities where document collections are potentially small (Jockers and Mimno, 2013; Schöch, 2017; Navarro-Colorado, 2018). In such scenarios, word co-occurrence statistics are less reliable due to sparsity of discrete counts.

With the rise of neural word embeddings (Mikolov et al., 2013), the defacto approach to improving over discrete models has been to utilise continuous representations (regardless of whether the setting is low resource). Early work by Liu et al. (2015) introduced topic dependent word embeddings, while others subsequently use embeddings to influence the discrete topic-word distribution (Zhao et al., 2017; Dieng et al., 2019). However, the low resource scenario constrains us to existing pre-trained embeddings, as the number of train documents is limited to several thousands and thus prohibitively small to train neural models (Srivastava and Sutton, 2017; Zhu et al., 2018; Liu et al., 2019; Hu et al., 2020; Zhu et al., 2020).

We therefore consider approaches that do not require further tuning of embeddings, and operate within the well established LDA probabilistic inference framework in the continuous space. There have been multiple attempts to replace discrete words with pre-trained embeddings (replacing the multinomial topic-word distribution with a continuous topic-word distribution), doing away with discrete words completely (Das et al., 2015; Bamghelich et al., 2016; Xun et al., 2017). Given the dominance of pre-trained word embeddings in modern NLP, would continuous representations outperform discrete representations even in low resource settings? Surprisingly, we find that discrete LDA outperforms its fully continuous counterpart on topic coherence measures which correlate with human judgement (Lau et al., 2014).

How then can we utilise pre-trained continuous representations for learning better topics? A natural direction is hybrid models based on statistical counts and pre-trained neural representations (Neubig and Dyer, 2016). Early work by Nguyen et al. (2015) used a mixture of discrete and continuous topic-word distributions with static mixture coefficients. However we find that this does not improve over discrete LDA, which motivates a more
nuanced treatment of the mixture coefficient.

In this work, we introduce an adaptive mixture coefficient specific to each word and each topic, which is updated at every step of Collapsed Gibbs Sampling (Griffiths and Steyvers, 2004). The intuition is as follows, topic anchor words (Lund et al., 2017) which have stronger signal from corpus statistics should rely more on the discrete distribution, while infrequent words should rely more on their embeddings (pre-trained on a large external corpus). Crucially, we do not assume any prior knowledge of the corpus used to train the word embeddings, and our parameterisation depends on the uncertainty of the continuous topic distributions at the current state of the Markov Chain during Gibbs Sampling. Our contributions are as follows:

1. By using adaptive mixed representations for the observed word with a data-dependent parameterisation, we provide an automatic trade-off between continuous and discrete representations during inference. Our method requires no additional tuning and relies purely on corpus statistics and statistics gathered from the current state of the Markov Chain.

2. We illustrate the extensibility of our approach to LDA variants with a combined topic model; Cross-lingual Adaptive LDA, and showed that adaptive mixing can balance between both discrete and continuous representations for better topic coherence on both monolingual and multilingual datasets.

2 Background

2.1 Unsupervised Learning with LDA

Discrete LDA (Blei et al., 2003) describes a generative probabilistic model of a corpus with latent topics. Formally we can define a corpus with $D$ documents and $K$ topics, where each document has a multinomial distribution over topics, $\Theta = \{\theta_1, \cdots, \theta_D\}$, and each topic has a multinomial distribution over words, $\Phi = \{\phi_1, \cdots, \phi_K\}$. $\Theta$ and $\Phi$ are the set of document-topic and topic-word distributions respectively. LDA relies on discrete counts and co-occurrence statistics, and therefore has poorer estimates in low resource scenarios due to data sparsity.

Gaussian LDA (Das et al., 2015) proposes a variant of LDA which operates on the continuous vector space rather than on discrete words. Each word is represented by an $M$-dimensional vector $v \in \mathbb{R}^M$ and is drawn from a multivariate Gaussian for that topic. That is, for $K$ topics, there are $K$ Gaussian distributions. While there have been extensions to more complex continuous distributions such as von Mises-Fisher (Batmanghelich et al., 2016; Li et al., 2016b)), we opted to work with a simpler distribution to demonstrate the approach, which can subsequently be extended in future work.

Polylingual LDA (Mimno et al., 2009) studies LDA across more than two languages using parallel corpora. The model assumes that the document-topic distribution $\theta_d$, is shared across languages, and that each language specific topic has a multinomial topic-word distribution, $\Phi^l_1, \Phi^l_2$ due to the discrete nature of words. Mimno et al. (2009); Ni et al. (2009) showed that Polylingual topic models can infer topic structure in multilingual corpora.

Latent Feature Topic Models A natural extension to discrete only or continuous only representations, is to model a word as being sampled with some probability from its discrete or continuous component. Nguyen et al. (2015) introduced an interpolation between the continuous and discrete representations, but convert the continuous representations back into discrete probability over word types by learning latent feature weights.

3 Discrete-Continuous Mixture LDA

We first establish an incremental extension to the Latent Feature Topic model using mixture of discrete categorical and continuous Gaussian distributions. We adopt a mixture model where each word has some probability of either coming from its categorical (discrete) or Gaussian (continuous) distribution. The generative process for this model with $K$ topics is as follows:

For topic $k = 1$ to $K$
1. Draw covariance $\Sigma_k \sim \mathcal{W}^{-1}(\Psi, \nu_0)$
2. Draw mean $\mu_k \sim \mathcal{N}(\mu_0, \frac{1}{\tau} \Sigma_k)$
3. Draw topic-word distribution $\phi_k \sim \text{Dir}(\lambda)$

For each document $d$ in corpus $C$
1. Draw a topic distribution $\theta_d \sim \text{Dir}(\eta)$
2. For each word $w_{d,i}$
   (a) Draw topic $z_{d,i} \sim \text{Multin}(\theta_d)$
   (b) Draw $\pi \sim \text{Beta}(\alpha, \beta)$
   (c) With $\pi$, draw $w_{d,i} \sim \text{Multin}(\phi_{z_{d,i}})$
   (d) With $(1-\pi)$, draw $v_{d,i} \sim \mathcal{N}(\mu_{z_{d,i}}, \Sigma_{z_{d,i}})$
where $W^{-1}$ is the Inverse Wishart distribution, $\Psi$ is the normalised Precision matrix, $\nu_0$ is degrees of freedom, $\mu_0$ is the prior mean for each Gaussian topic, and $\pi$ is a mixture coefficient.

3.1 Gibbs Sampling for Posterior Inference

Given a corpus, our goal is to infer the posterior distribution over $\Theta$ and $\Phi$ and latent topic assignments $z$, given the observations $x$. We perform inference with collapsed Gibbs sampling (Griffiths and Steyvers, 2004) which can be derived by analytically integrating out $\Theta$ and $\Phi$.

The key step in Gibbs sampling\(^2\) samples a new topic $z_{d,i}$ assignment for each word, $w_{d,i}$ at index $i$ in document $d$ based on the conditional distribution where the previous assignment is ignored (denoted with $\setminus$):

$$
p(z_{d,i} = k | z_{\setminus d,i}, x, \eta, \varphi) 
\propto p(x_{d,i} | z_{d,i} = k, \varphi, z_{\setminus d,i}, x) 
\times p(z_{d,i} = k | \eta, z_{\setminus d,i}, x) \quad (1)
$$

$\eta$ is the corresponding parameters of a Dirichlet prior for the document-topic distribution $\theta$, and $\varphi$ are parameters associated with the topic-word distribution. This is either $\lambda$ for the Dirichlet prior for multinomial $\phi$, or $\mu_0, \Sigma_0, \nu_0, \kappa$ for the Gaussian.

In our proposed model (section 5), $\varphi$ consists of both Dirichlet and Gaussian parameters.$^3$

The first term on the right in Equation 1 expresses the probability of the $i^{th}$ word in document $d$ under topic $k$, while the second term expresses the probability of topic $k$ in document $d$ (Griffiths and Steyvers, 2004). Gaussian LDA modifies the first term to use continuous representations instead of discrete, while cross-lingual models focus on the second term which reflects document level sharing.

We focus on the first term to incorporate adaptive mixed representations in section 5.

Mixture Models

Let $f_1$ be a discrete probability mass function with parameters $\varphi_1$ and $f_2$ be a continuous density function with parameters $\varphi_2$. The density can be expressed as a convex combination:

$$f(x|\varphi_1, \varphi_2) = \pi f_1(w|\varphi_1) + (1-\pi) f_2(v|\varphi_2) \quad (2)$$

Then, the second term in Equation 3 can be expressed as the density of $v_{d,i}$ under the multivariate $t$ distribution\(^4\) parameterised by mean $\mu_k$ and covariance $\frac{\kappa_k + 1}{\kappa_k} \Sigma_k$, with $\nu_k$ degrees of freedom. $\kappa$ is a prior confidence on $\mu_k$ and $\Sigma_k$ (Murphy, 2012).

$$\varphi = \{\lambda, \nu_0, \mu_0, \Sigma_0, \kappa\}, \text{ including parameters of both the Dirichlet and Gaussian priors, with the subscript 0 indicating parameters of the conjugate prior.}$$

4 Perspectives on Mixture Coefficient $\pi$

4.1 Perspectives on Static $\pi$

There are several ways to interpret the mixture coefficient $\pi$ which interpolates between the discrete and continuous representations. Both the discrete and Gaussian LDA can be viewed as special cases of a two component mixture model, where the mixture coefficient $\pi$ is either 1 or 0 respectively. $\pi$ can also be viewed as a tunable hyperparameter that emphasises either representation depending on the availability of discrete word units or quality of embeddings (Nguyen et al., 2015).

4.2 Perspectives on $\pi$ as a Static Random Variable Informed by Observations

From a Bayesian perspective, the mixture coefficient, $\pi \in [0, 1]$, can be modelled as a random variable following a Beta distribution. This provides a distribution over component proportions (discrete or continuous) with useful conjugate properties. By Bayes Rule, posterior inference of $\pi$ is proportional to prior times likelihood: $p(\pi|x) \propto p(\pi)p(o|\pi)$.

\(^2\)We refer readers to Resnik and Hardisty (2010) for a detailed explanation or refresher of this process.

\(^3\)Hyperparameters are not crucial to understanding our method, and we expand on the notation for them in Table 1.

\(^4\)The multivariate $t$ distribution arises in Bayesian Inference when the variance of a normally distributed random variable is unknown (Gelman et al., 2013).

\(^5\)When $w_{d,i}, w'$ are subscripts of $N$ or $\lambda$, they are integers that index a count vector or Dirichlet parameter vector $\lambda$, e.g. when used in the context of $N_{k,w_{d,i}}$, $w_{d,i}$ is the index of the word type, for the token of $i^{th}$ word of document $d$. 

Table 1 summarises current and previous work with respect to Equation 1.
We recommend a more pragmatic view for balancing between (noisy) learned word embeddings and discrete representations by modeling the mixture coefficient as a topic-specific random variable. At inference time, we sample \( \pi_{k,j} \sim Beta(\alpha_{k,j}, \beta_k) \) from a Beta distribution that is specific to each word type and each topic for the \( \alpha \) parameter, and topic specific for the \( \beta \) parameter to compute Equation 3.

\[
\pi_{k,j} \sim Beta(\alpha_{k,j}, \beta_k)
\]

The parameter \( \alpha_{k,j} \) represents the concentration on the discrete representation, while \( \beta_k \) represents the concentration on the continuous representation. As we do not assume any knowledge of the external corpus used to train the continuous representations, the \( \beta \) parameter is agnostic to the word type. On each Gibbs sampling update, \( \alpha_{k,j} \) is updated by discrete counts for the categorical distribution, and \( \beta_k \) is updated based on the uncertainty in the \( \ell \) distribution as measured by the trace of the covariance matrix \( \Sigma_k \).

### 5.1 Adaptive \( \alpha_{k,j} \) Based on Counts

We specify corpus specific \( \alpha' \) priors, \( \alpha^0_j \) for each word type indexed by \( j \) in the vocabulary as the number of word counts \( N_{k,j} \) normalised by \( K \), the number of topics, and the relative proportion of number of documents \( D \) to number of unique vocabulary words \( |V| \).

\[
\alpha_j^0 = \frac{D}{|V|} + \frac{\sum_k N_{k,j}}{K}
\]

Intuitively, we expect that if a word has a higher frequency in the corpus, its statistics based on

| LDA types | Topic-Word | \( p(x_{d,i}^j | z_{d,i}^j = k, \psi, z_{(d,i)}; x) \propto p(z_{d,i}^j = k | \eta, x) = \) |
|-----------|------------|--------------------------------------------------|
| Discrete  | Discrete   | \( \mathbb{D} = \frac{N_{k,w_{d,i}} + \alpha_{w_{d,i}}}{\sum_{w'} N_{k,w'} + \lambda_w} \) |
| Gaussian  | Continuous | \( \mathbb{C} = t \nu \left( \mu_k, \frac{\kappa_k + 1}{\kappa_k} \Sigma_k \right) \) |
| Static \( \pi \) | Mixture | \( (1 - \pi) \cdot (\mathbb{C}) + \pi \cdot (\mathbb{D}) \) |
| Adaptive \( \pi \) | Mixture | \( (1 - \pi_{k,j}) \cdot (\mathbb{C}) + \pi_{k,j} \cdot (\mathbb{D}) \) |
| Polylngual | Discrete | \( \mathbb{L} = \frac{N_{k,w_{d,i}} + \alpha_{w_{d,i}}}{\sum_{w'} N_{k,w'} + \lambda_w} \) |
| \( \ell \)-Adapt | Mixture | \( (1 - \pi_{k,j}^\ell) \cdot (\mathbb{L}_{k,w_{d,i}}) + \pi_{k,j}^\ell \cdot (\mathbb{L}) \) |

Table 1: Comparison of various LDA models. For topic \( k \), language \( \ell, N_{k,d,i}^{\ell,d} \) are counts of topic \( k \) in document \( d, N_{k,w_{d,i}}^{\ell} \) are counts of the word type for the \( i^{th} \) word in document \( d, w_{d,i} \) in topic \( k, j \) indexes the word type for the token \( w_{d,i} \) or \( v_{d,i} \), and \( t \nu \) is the probability density function of the multivariate \( \nu \) distribution parameterised by \( \nu \) degrees of freedom, mean \( \mu_k \) and covariance \( \Sigma_k \). \( \kappa_k = \kappa + \sum_{w'} N_{k,w'} \), where \( \kappa \) represents the belief on the prior of the multivariate Gaussian. For the cross-lingual model, \( \nu^\ell \) and \( \kappa^\ell \) are parameters of the Dirichlet prior distribution.
discrete counts are more reliable. However, if $|V| >> D$, count statistics become less reliable. The $\alpha_{k,j}$ parameter at each step where $N_{k,j}$ is the number of times word type at vocabulary index $j$ was assigned to topic $k$ is

$$\alpha_{k,j} = \alpha^0_j + N_{k,j} \quad (7)$$

which takes a similar form to the regular closed-form conjugate posterior update in Equation 4 for discrete counts.

5.2 Adaptive $\beta_k$ Based on Topic Uncertainty

Recall that while counts are appropriate for the discrete case, continuous representations are learned from an external corpus and should not constitute full observation counts. Hence there is no closed form update for the membership of the continuous representations (Koller and Friedman, 2009).

Instead we let $\beta_k$ be a random variable which reflects our current confidence in the multivariate $t$ distribution indexed by topic $k$.

We approximate the uncertainty of the $k$ topic distribution, as measured by the sum of eigenvalues of the square root of the topic covariance matrix $\Sigma_k$, equivalently written as $tr(\sqrt{\Sigma_k})$. We formulate $\beta_k$ as depending on the constant terms $\frac{M}{K}; M$ is the number of dimensions of the multivariate Gaussian, and (non-constant) $\Sigma_k$ which is updated at every step of Gibbs Sampling:

$$\beta_k = \frac{M}{K} (tr(\sqrt{\Sigma_k}))^{-1} \quad (8)$$

The intuition for the inverse relationship between $tr(\sqrt{\Sigma_k})$ and $\beta_k$ is as follows. If the topic has high variance, then $\beta_k$ should be smaller as we have less confidence in its density function. The square root is a computational convenience for working with the Cholesky decomposition $L_k^T L_k = \Sigma_k$, where the last step assumes most of the variance is contained along the diagonals\(^8\) of $L_k$. In the following equation, we simplify the notation of $L_k$ to $L$.

$$(tr(\sqrt{\Sigma_k}))^{-1} = (tr(\sqrt{L^T L}))^{-1} \approx (||\sqrt{L}||_F^{-2})^{-1} = (\sum_{i,j} L_{ij})^{-1} \approx (\sum_i L_{ii})^{-1} \quad (9)$$

We elaborate on the interpretation of $B_k$ in Appendix C.

\(^8\) We verified this assumption by inspecting $L_k$, and found that the off-diagonals tended to be smaller by a factor of 3.

### Figure 1: Cross-Lingual Adaptive LDA, with shared continuous parameters $\mu_k$, $\Sigma_k$ across languages and adaptive $\pi_{k,j}$ for every word type $j$ and topic $k$. The word type $j$ corresponds to the $i^{th}$ token of document $d$. $w_{d,i}$ indicates a token when it is not being used as a subscript.

6 Computational Complexity

We consider the computational cost for every Gibbs Sampling step. The main source of computational complexity comes from inverting $\Sigma_k$ which takes $O(M^3)$ when computing the probability density of $v_{d,i}$ in row 2 of Table 1.

Since the covariance matrix $\Sigma_k$ is symmetric and positive semi-definite, we can utilise the Cholesky decomposition where $\Sigma_k$ can be decomposed as a product of upper and lower triangular matrices, $\Sigma_k = L_k^T L_k$. Although this takes $O(M^3)$, we pay this cost only once during initialisation. $L_k$ is maintained by performing rank-1 updates and downdates (Seeger, 2004) at every step of Collapsed Gibbs Sampling.

As shown in Das et al. (2015), calculating the probability density takes $O(M^2)$ instead of $O(M^3)$. Our proposed prior for $\beta_k$ sum the diagonals of $L_k$ which takes $O(M)$ with little to no constant time overhead.

Therefore each Gibbs Sampling step takes $O(KM^2)$ where $K$ is the number of topics whose $p(v_{d,i} | z_{d,i} = k, \varphi, x)$ we need to compute. This is parallelisable to $O(M^2)$ as each term can be computed independently.

7 Cross-lingual Adaptive LDA

The following section describes the extension of our work from the monolingual to the cross-lingual setting. To test the robustness of our proposed model and extensibility to other models, we study the topic coherence in multilingual settings where the quality of word embeddings is thought to be
worse than monolingual embeddings. We introduce a new topic model for continuous multilingual representations building on our adaptive sampling scheme, Cross-lingual Adaptive LDA in (Figure 1).

**Modeling Assumptions** Following Mimno et al. (2009), we assume that the document-topic distribution $\theta_d$ is shared across paired language documents, and follow a bag-of-words assumption, i.e., they need not be sentence or word-aligned. We additionally assume that multilingual word embeddings $v_\ell^1, v_\ell^2$ have been mapped to the same embedding space, by adopting shared Gaussian mean $\mu_k$ and covariance $\Sigma_k$ across languages. This reduces the number of parameters and importantly ensures a continuous mapping across languages. Although this does not necessarily affect topic-coherence when measured within in each language, this would results in very poor cross-lingual document-topic representations. We checked this assumption by inspecting the learned topics without parameter sharing and found that topic indexes were mismatched across languages. Topic 5 in English may be about sports but Topic 5 in French may be about medicine.

7.1 Adaptive Mixing for Cross-lingual LDA

For the cross-lingual setting, our parameterisation of Equation 5 takes into account language $\ell \in \mathcal{L}$ for word type $j$ and topic $k$:

$$
\alpha_{\ell,j} = \frac{D}{|V_{\ell}|} + \frac{\sum_{k'}^{K'} N_{k',j}^{\ell}}{K} + N_{k,j}^{\ell}
$$

$$
\beta_k^\ell = \left(\frac{M}{K \cdot |L|}\right)(tr(\sqrt{\Sigma_k}))^{-1}
$$

Similar to the low resource monolingual setting, our approach relies on existing pre-trained multilingual word embeddings. Note that each language may have different vocabulary size.

8 Experiments

8.1 Experimental Setup

We conduct experiments on a standard monolingual dataset and multilingual wikipedia dataset, reflecting a resource constrained setting by reducing the number of train documents. Our experiments investigate the following:

- Does an adaptive mixture coefficient perform better than the fully continuous, fully discrete, and static mixture coefficient?
- How do the various mixture models perform across different number of training documents?

We compare the following models in Table 2, $\ell$- indicates the cross-lingual case in Table 3 and SMIX is as described in subsection 4.2:

- DISC: Discrete LDA ($\pi = 1$)
- GAUS: Gaussian LDA ($\pi = 0$)
- SMIX: Static Mix ($\pi = 0.5$)
- ADAP: Adaptive Mix (adaptive $\pi$)

8.2 Datasets

We use the 20 newsgroup dataset (20NG) which is a common text analysis dataset containing around 18000 documents and 20 categories. We perform stratified shuffled sampling, using 7000 docs as holdout test and varying the number of training documents from 1000 to 8000. For each model and each training size, we present the results averaged across 5 random splits of the dataset. Since the goal is to model the present corpora, our main results are evaluated on a held-out test set based on the same corpora. We additionally evaluate on a held-out test set following (Röder et al., 2015). GAUSS performs better in this setting, and we discuss possible reasons in Appendix F.

**Wikipedia paired document corpus.** For the multilingual scenario, we utilised a Wikipedia dataset (Sasaki et al., 2018) that was automatically constructed by inter-language link to the most relevant foreign language document. For the multilingual setting 1000 test pairs were standardized across all languages, and training data consisted of 8000 randomly selected document pairs for each language. We performed shuffled sampling on the training data for 5 random splits of 1000 and 7000 training document pairs.

Note that low resource topic modeling is not equivalent to low resource languages. A language can be considered high resource but the collection of documents that we are modeling could be small.\(^{11}\)

\(^{10}\)The dataset can be obtained at http://qwone.com/˜jason/20Newsgroups/\(^{11}\)An example of this is the modeling of Golden Age Spanish Sonnets (Navarro-Castro, 2018) which has a corpus size of around 5000 documents.
2457

8.3 Topic Coherence Measure

Topic models are often evaluated based on the likelihood of held-out documents. However, the likelihood of words from the discrete probability mass function and continuous probability density function is not directly comparable. Instead, we compute the coherence score $S_k$ of topic $k$ using the normalised point-wise mutual information (NPMI; Bouma 2009)) which has been found to correlate with human judgement of topic quality (Lau et al., 2014). We also evaluate on the ‘$C_v$’ metric, which is closely related (see Appendix F) from Röder et al. (2015).

NPMI ranges from $[-1, 1]$, where $-1$ indicates no co-occurrences and 1 indicates 100% co-occurrences. $^*$The score of each topic $S_k$ is computed from word pair combinations of the top $T$ words returned by that topic.

$$S_k = \sum_{i=1}^{T} \sum_{j=1}^{i-1} NPMI(w_i, w_j) \quad (11)$$

$$NPMI(w_i, w_j) = \frac{\log p(w_i, w_j) + \epsilon}{-\log p(w_i) p(w_j) + \epsilon} \quad (12)$$

We extract word co-occurrence statistics of the held-out documents to compute $p(w_i)$ and $p(w_i, w_j)$, and set $\epsilon$ to $1e^{-12}$ to avoid logarithm of 0. NPMI averaged across all topics are reported as $\frac{1}{K} \sum_k S_k$ in Table 2 and 3.

Note that the standard metric in Equation 12 will encounter division by 0 for the case where $p(w_i) \cdot p(w_j) = 0$, which is a case which frequently occurs in our low resource setting. We elaborate on this in Appendix D.

8.4 Results and Discussion

Finding 1: Adaptive Mixing performs best in resource constrained settings. We see that in Table 2, the adaptive mixture coefficient performs better under more resource scarce settings, and

Table 2: Performance of various models with variable number of train documents on NPMI (higher is better). Each NPMI score reported was averaged across 5 random train-test splits. $^p < 0.05$ significant difference for paired t-test against the strongest baseline (DISC).

| No. docs | DISC | GAUS | SMIX | ADAP |
|----------|------|------|------|-------|
| 1000     | -0.067 | -0.036 | -0.003 | 0.022* |
| 2000     | 0.006 | -0.055 | -0.046 | 0.043* |
| 3000     | 0.030 | -0.089 | -0.103 | 0.048* |
| 4000     | 0.044 | -0.111 | -0.141 | 0.052 |
| 5000     | 0.044 | -0.139 | -0.220 | 0.059 |
| 6000     | 0.059 | -0.283 | -0.251 | 0.076* |
| 7000     | 0.072 | -0.213 | -0.235 | 0.092* |
| 8000     | 0.093 | -0.192 | -0.261 | 0.094 |

Table 3: Performance of various models on languages ro: Romanian, fr: French, pl: Polish, es: Spanish on 1000 and 7000 documents. Each NPMI score reported was averaged across 5 random train-test splits. $^p < 0.05$ significant difference for paired t-test against the strongest baseline (DISC).

| $\ell$ | No. docs | DISC | GAUS | SMIX | ADAP |
|--------|----------|------|------|------|-------|
| ro 7000 | 0.045 | -0.307 | -0.105 | 0.081* |
| fr 7000 | 0.049 | -0.258 | -0.101 | 0.052 |
| pl 7000 | 0.032 | -0.273 | -0.174 | 0.024 |
| es 7000 | 0.039 | -0.283 | -0.112 | 0.053* |

Preprocessing  Standard text preprocessing steps were applied. Stopwords, digits, punctuations, words that appeared less than 5 times and the top 10 most frequent words were removed for efficiency. Wikipedia articles were restricted to the first 200 words and document titles were removed.

Model Settings  All experiments (both 20NG and the multilingual experiments) were conducted with pre-trained multilingual word embeddings from the MUSE library (Conneau et al., 2017). We trained for up to 100 iterations and checked for convergence by inspecting mixing of the posterior topic-word distributions.

Hyperparameters  We initialised the prior mean $\mu_0$ and covariance $\Sigma_0$ to the empirical mean and sample covariance respectively based on random assignment of words to topics. Following Das et al. (2015), we initialise $\kappa$ to 1, $\nu_0$ to the embedding size $M$ of 300. Parameters of the Dirichlet prior $\eta$ and $\lambda$ are set to 1 and 0.01 respectively, and $K = 20$. The same hyperparameter settings were used in the multilingual setting.

All parameters of our proposed approach are based on corpus statistics, and existing parameters such as number of topics, and embedding size.
Figure 2: Topic proportions for various mixture coefficients ($\pi$) in the 20NG dataset using $K = 20$. Slices do not correspond to ground truth, and only illustrate relative proportions. Although DISC can recover a similar proportion to the ground truth, the quality of topics are not as good as ADAP.

After a certain point, is nearly equivalent to the Discrete LDA. These results are in the direction that we expect, the discrete model performs increasingly well with larger corpus sizes.

Gaussian LDA (GAUS) performs poorly with increasing number of training documents. The authors report better performance using Pointwise Mutual Information (PMI) which assigns high scores to rare words such as human names such as “scott, graham, walker.” which are not representative of themes. In this work, we evaluate using normalized PMI (Bouma, 2009) which corrects for this. This is somewhat surprising given the dominance of neural methods in modern NLP, and motivates our analysis (see Observation 1 and Observation 2) in the next section.

Interestingly, even with a less optimal continuous distribution, the adaptive method is able to balance between both representations with low number of training documents, and has a ‘jump-start’ using embeddings. We note that ADAP performs slightly less convincingly in the multilingual setting in terms of achieving statistical significance (not poorer in absolute terms), which could be due to poorer quality of multilingual embeddings.

Finding 2: Static mixture coefficient of $\pi = 0.5$ performs poorly, and while this could potentially be tuned for better performance, our adaptive method requires no tuning at all. This is discussed further in subsection 8.6.

| ADAP train docs = 1000 | ADAP train docs = 7000 |
|------------------------|------------------------|
| government | law | ADAP | GAUS |
| public | color | government | word |
| laws | remember | christ | james |
| crime | days | bible | mary |
| court | idea | sin | smith |
| legal | told | christians | andrew |
| trust | post | death | gordon |
| police | list | paul | norton |
| fbi | process | church | thomas |
| | | mary | george |

Table 4: Top topic words on 20NG, bolded words are common across both topics. ADAP (Adaptive $\pi$) is able to construct topics with little training data (1000 docs), and correctly assigns human names to their ground truth topic.

One might expect that SMIX should not be worse than DISC or GAUSS, since it has access to both discrete and continuous distributions. However, the results suggest that equally weighting both the continuous and discrete topic representations, causes the model to not be able to learn effectively if they are in conflict, for e.g. continuous topic prefers topic 15 and the discrete topic prefers topic 3, and if weighted in equal proportions, this hinders the updates in Gibbs Sampling.

8.5 Analysis

Observation 1: GAUS produces narrow topics which are oddly narrow based on names (Table 4), American Cities, directions (North, South, East, West) etc. This phenomena is present in both the monolingual and multilingual models. While these groups of words may be semantically close, they are not representative ‘themes’ in a corpus.

This may be attributed to pre-training via skip-gram loss to predict neighboring words (Mikolov et al., 2013). Words which are used in similar contexts have similar embeddings, and the more unique the context is, the narrower the word clusters. To verify this, we compared word clusters from the Gaussian Mixture Model (with same $K$) (Bishop, 2006), which uses no corpus information. We observe a high word overlap with the topics from GAUS (see Appendix E), indicating that the continuous representations dominate the corpus co-occurrence statistics.

Observation 2: GAUS has a rich-get-richer phenomena. Figure 2 shows the size of topics produced by different models on the 20NG. With the exception of very narrow clusters of words,
most words collapse onto a single topic for GAUS (Figure 2b). If many words have been assigned to one topic, that topic covariance \( \Sigma_k \) becomes much larger than the others, leading to subsequent \( \nu_k \) then having a higher relative density under that topic during Gibbs Sampling.

Our proposed adaptive \( \pi \) (Figure 2d) counteracts this effect better than the static \( \pi \) (Figure 2c). If \( \Sigma_k \) is large, to balance the effect of words having a higher density under topic \( k \), the algorithm samples a larger \( \pi \), thereby placing less weight on the continuous representation.

**Observation 3:** ADAP is flexible and produces reasonable topics. Discrete LDA does not perform well with low training data due to sparsity of word co-occurrences. Table 4 shows that ADAP does not suffer from this and can make up for the lack of training data to produce a topic about ‘government’ and ‘law’. Next, we observe that while GAUS clusters all human names together based on their embedding space, ADAP is not overly reliant on embeddings and can correctly assign ‘Paul’ and ‘Mary’ to its ground truth topic of christianity. Additional topics and NPMI coherence scores are available in Appendix G.

### 8.6 Stability of Mixture Coefficient

As our experiments were conducted with a fixed number of topics, we study the expectation of \( \alpha, \beta, \pi \) under a varying number of topics \( (K \text{ from } 20 \text{ to } 200) \). We approximate the expectation by the arithmetic average:

\[
E[\alpha] = \frac{1}{K \sum_k} \sum_k \sum_j \alpha_{k,j} \quad \text{and} \quad E[\beta] = \frac{1}{\sum_j \sum_k} \sum_k \sum_j \beta_{k,j}
\]

for a fixed \( K \), where \( E[\pi] \) and \( E[\beta] \) are calculated in the same manner. We verified that as \( K \) increases, the variance of \( \pi \) increases as expected, as reflected by the smoothly decreasing \( E[\alpha] \) and \( E[\beta] \).

Note that \( \alpha \) and \( \beta \) take on different values for each word and topic during Gibbs Sampling. We observe that while \( E[\pi] \) is close to 0.5 for \( K = 20 \) for ADAP, it significantly outperforms SMIX (\( \pi = 0.5 \)) in Table 2. This lends confidence to the interpretation that the adaptive mixture coefficient \( \pi_{k,j} \) contributes to the better performance, as opposed to simply having a better static \( \pi \).

### 9 Conclusion

Low resource scenarios present an interesting challenge to topic modeling due to sparsity of counts and a lack of data to train neural models. Our work proposes an automatic trade-off between externally trained continuous representations and traditional co-occurrence count-based statistics that is specific to each word and topic. The method accounts for variations in number of topics and embedding dimensions, and requires no additional tuning beyond existing methods.

Importantly, it requires no additional retraining of word embeddings or learning of topic embeddings, allowing us to rely solely on pre-trained representations and existing corpus statistics. We showed the efficacy and extensibility of our approach on a monolingual and a multilingual dataset, while introducing a new Cross-lingual Adaptive LDA topic model in the process. In future work, we aim to study the different scenarios of low resource (e.g., when there are a lot of infrequent words such as named entities) and their interaction with different embedding methods.

### Acknowledgements

We thank the anonymous reviewers and JHU colleagues Anton Belyy, Sabrina Mielke, Matt Post, Tom Lippincott, Mei Hongyuan for their helpful comments on earlier drafts, Xu Yanxun, Desh Raj for technical discussion, and Tim Vieira for help with Cython.

### References

Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Sam Gershman. 2016. Nonparametric spherical topic modeling with word embeddings. In *Proceedings of the conference. Associ*
Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2020. Short text topic modeling techniques, applications, and performance: a survey. IEEE Transactions on Knowledge and Data Engineering.

Philip Resnik and Eric Hardisty. 2010. Gibbs sampling for the uninitiated. Technical report, Maryland Univ College Park Inst for Advanced Computer Studies.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining, pages 399–408.

Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 458–463.

Christof Schöch. 2017. Topic modeling genre: An exploration of french classical and enlightenment drama. DHQ: Digital Humanities Quarterly, 11(2).

Matthias Seeger. 2004. Low rank updates for the cholesky decomposition. Technical report.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. arXiv preprint arXiv:1703.01488.

Guangxu Xun, Yaliang Li, Wayne Xin Zhao, Jing Gao, and Aidong Zhang. 2017. A correlated topic model using word embeddings. In IJCAI, pages 4207–4213.

He Zhao, Lan Du, and Wray Buntine. 2017. A word embeddings informed focused topic model. In Asian Conference on Machine Learning, pages 423–438.

Lixing Zhu, Yulan He, and Deyu Zhou. 2020. A neural generative model for joint learning topics and topic-specific word embeddings. Transactions of the Association for Computational Linguistics, 8:471–485.

Qile Zhu, Zheng Feng, and Xiaolin Li. 2018. Graphbttm: Graph enhanced autoencoded variational inference for biterm topic model. In Proceedings of the 2018 conference on empirical methods in natural language processing, pages 4663–4672.
A Pseudocode for Crosslingual Adaptive LDA Inference

Algorithm 1: Adaptive Mixing LDA Inference

**Data:** Documents
\[ D = \{doc_1, \cdots, doc_D\}, \]
Vocab \( V = \{w_1, \cdots, w_{|V|}\}, \)
Embeddings \( \{v_1, \cdots, v_{|V|}\}, v \in \mathbb{R}^M \)

**Result:** \( \Phi = \{\phi_1, \cdots, \phi_K\}, \Theta = \{\theta_1, \cdots, \theta_D\} \)

1. **Initialization:**
   1. \( \mu_0 \leftarrow \frac{1}{|V|} \sum_{j=1}^{|V|} v_j \) (prior mean)
   2. \( \Sigma_0 \leftarrow \frac{1}{|V| - 1} \left( \sum_{j=1}^{|V|} (v_j - \mu_0)(v_j - \mu_0)^T \right) \) (prior cov)

2. **for each doc \( d \in \{1, \cdots, D\}, \) word \( i \) do**
   1. \( z_{d,i} \leftarrow k \) uniform sample from \( \{1, \cdots, K\} \)

3. **for topic \( k \in \{1, \cdots, K\} \) do**
   1. \( N_k \leftarrow |\{z_{d,i} = k|\forall i \in [d], \forall doc_d \in D\}| \)
   2. \( \mu_k \leftarrow \frac{\alpha_0 + N_k \varphi_k}{\nu_0 + \psi_k} \)
   3. \( \Sigma_k \leftarrow \frac{1}{\nu_0 + N_k - M + 1} \) for each \( \Sigma_k \)
   4. \( L_d^T L_d \leftarrow \Sigma_k \) (cholesky decomposition)
   5. \( N_{k,j} \leftarrow |\{z_{d,i} = k|\forall doc_d \in D\}| \) for each \( w_{d,i} \)
   6. \( N_{k} \leftarrow |\{z_{d,i} = k|\forall w_{d,i} \in doc_d \}| \) for each \( doc_d \in D \)

4. **while iter<maxiter or not converged do**
   1. **for each doc \( d \in \{1, \cdots, D\}, \) word \( i \) in \( doc \) do**
      1. \( z_{old} \leftarrow z_{d,i} \)
      2. \( N_{z_{old},j} \leftarrow N_{z_{old},j} + N_{z_{old},j}^d \)
      3. \( \alpha_{z_{old},j} \beta_{z_{old}} \) (Eq : 7, 9)
      4. \( \mu_{z_{old},L_{z_{old}}} \) (Eq : 7, 9)
      5. **for topic \( k \in \{1, \cdots, K\} \) do**
         1. Sample \( \pi_{k,j} \sim Beta(\alpha_{k,j}, \beta_{z_{old}}) \)
         2. Compute \( p(z_{d,i} = k|\varphi, x) \) (Eq : 1, 3)
      3. Sample \( z_{new} \sim p(z_{d,i} = k|\varphi, x) \)
      4. \( N_{z_{new},j} \leftarrow N_{z_{new},j}, N_{z_{new},j}^d \)
      5. \( \alpha_{z_{new},j} \beta_{z_{new}} \) (Eq : 6, 8)
      6. \( \mu_{z_{new},L_{z_{new}}} \) (Eq : 7, 9)

The full inference algorithm is given in Algorithm 1. For details on the parameterisation of the multivariate \( t \), update of \( \mu_k \) and computation of \( \Psi_k \), we refer readers to Murphy (2012). For update and downweights of \( L_{k} \), we refer readers to Seeger (2004) and Das et al. (2015).

B Accounting for Uncertainty in the Multivariate \( t \) Distribution

We present a small modification when calculating the density of the word vector \( \mathbf{v}_{d,i} \) for each topic (row 2 of Table 1). At each step of Gibbs Sampling, the model samples a topic based on the relative likelihood of a \( \mathbf{v}_{d,i} \) drawn from a \( t \)-distribution of topic \( k. \)

We observe that in Equation (1), the second term is dominated by the first term, where \( x_{d,i} \) is a word vector representation, \( \mathbf{v}_{d,i} \).

In high dimensions, \( p(\mathbf{v}_{d,i} | v_{d,i}^0) = k, \varphi, \mathbf{x}_{d,i} \) becomes highly skewed towards a certain topic, such that the influence of the document structure becomes negligible. This motivates a correction in the first term, as the embeddings are pre-trained rather than a true signal. We correct the degrees of freedom \( \nu_k \) to better account for uncertainty in the embedding representations.

B.1 Rescaling the Degrees of Freedom \( \nu_k \)

As given by Murphy (2012), \( \nu_k = \nu_0 + N_k - M + 1 \), where \( N_k \) is the number of words assigned to topic \( k \) and \( M \) is the embedding dimensions. Upon initialisation, under random assignment of words to topics, \( \mathbb{E}[N_k] = \frac{|V|}{|T|} \), where \( |V| \) are all the (non-unique) words in the corpus. Since for a typical corpus \( |V| \) is very large and \( \frac{|V|}{|T|} >> M \), the degrees of freedom \( \nu_k \) are very large resulting in an approximate normal distribution which is over-confident in its posterior predictions. This effectively dominates the priors for \( \Sigma_0, \nu_0 \) or \( \mu_0 \). Hence, we rescale \( \nu_k \) to \( \nu_k \) from 1 to 30\(^{16}\) to account for inherent uncertainty over \( \nu_i \) belonging to any particular topic.

The effect of rescaling \( \nu_k \) results in a heavier-tail distribution which results in higher density for \( \mathbf{v} \) which are further from \( \mu_k \). This encourages better mixing during Gibbs Sampling.

Comparison with the fully Bayesian treatment.

We found this heuristic to be numerically and em-
empirically more stable than a fully Bayesian treatment which encodes a higher variance in the distribution by having a larger prior on the covariance matrix \( \Sigma_0 \).

First, re-estimating the covariance matrix at every step of Gibbs Sampling is numerically unstable with a large \( \Sigma_0 \). Next, rescaling \( \nu_k \) guarantees that we maintain a heavy-tailed distribution at every iteration of Gibbs Sampling resulting in better mixing of the Markov Chain. By adopting the rescaling heuristic, we can directly set the prior covariance \( \Sigma_0 \) to its sample covariance, removing one adhoc parameter choice. Since both setting a large prior \( \Sigma_0 \) and scaling \( \nu_k \) are modeling decisions, we adopt the approach that is numerically and empirically more stable.

### C Interpretation of \( \beta_k \)

Note that \( \beta_k \) can be interpreted as a random variable drawn from a Gamma distribution, with shape parameter \( \frac{1}{K} \), and rate parameter \( \frac{tr(\sqrt{\Sigma_k})}{M} \).

\[
\beta_k \sim \text{Gamma}\left( \frac{1}{K}, \frac{tr(\sqrt{\Sigma_k})}{M} \right)
\] (13)

Then, Equation 8 is the point estimate of \( \beta_k \) obtained from the expectation of the Gamma distribution, where \( \beta_k \in (0, \infty) \) can be interpreted as real-valued ‘counts’ for observing the continuous representation. The rate parameter is scaled by \( \frac{1}{M} \) to make the numerator robust to dimension size. Since \( \Sigma_k \) is positive semi-definite, and square root is a monotonically increasing function, as \( M \) increases, the trace of \( \Sigma_k \) increases (\( \sum_{i=1}^{M} \sigma_i, \sigma_i \geq 0, \forall i \)) and \( \beta_k \) decreases.

### D NPMI when \( p(w_i) \cdot p(w_j) = 0 \)

In our implementation of NPMI, we do not consider the pair if either \( p(w_i) \) is \( p(w_j) \) is 0, as this simply indicates a “mismatch” between training and test corpus. However if they are non-zero, and \( p(w_i) \cdot p(w_j) = 0 \), then the model has predicted a poor word pair that never co-occurs despite them individually appearing in the test corpus, and the score for \( NPMI(w_i, w_j) = -1 \).

This differs from many online implementations of NPMI which will simply set \( NPMI(w_i, w_j) = 0 \) if \( p(w_i) \cdot p(w_j) = 0 \), and ‘does not penalise very poor word pairs of this nature.

| Low GMM Overlap | High GMM Overlap |
|-----------------|-----------------|
| Century named   | January France  |
| modern live     | February French |
| centuries written | December Paris |
| white including | March Vendée    |
| events wrote    | July Allier     |
| built based     | September Ghirone |
| renaissance countries | June Spain |
| growth history  | October Picardie |
| list published  | April Belgium   |

Table 5: ‘Genuine’ Topic model clusters learnt from the documents vs clusters with \( \geq 0.8 \) GMM overlap.

### E Overlap with GMM

### F Evaluation on held-out test set using \( C_v \) Topic coherence measure.

\( C_v \) combines the indirect cosine measure with the NPMI and the boolean sliding window and was introduced in Röder et al. (2015). The implementation of the metric and held-out wikipedia dataset is provided by [https://github.com/dice-group/Palmetto](https://github.com/dice-group/Palmetto).

| No. docs | DISC | GAUSS | SMIX | ALDA |
|----------|------|-------|------|------|
| 1000     | 0.3839 | 0.3886 | 0.4014 | 0.3964 |
| 2000     | 0.3895 | 0.4180 | 0.4319 | 0.3993 |
| 3000     | 0.4014 | 0.4374 | 0.4269 | 0.4162 |
| 4000     | 0.3985 | 0.4300 | 0.4289 | 0.4111 |
| 5000     | 0.4045 | 0.4278 | 0.4185 | 0.4066 |
| 6000     | 0.4079 | 0.4300 | 0.4161 | 0.4061 |
| 7000     | 0.4039 | 0.4262 | 0.3981 | 0.4092 |
| 8000     | 0.4090 | 0.4298 | 0.3936 | 0.4072 |

Table 6: \( C_v \) score on held-out wikipedia dataset.

We believe the main reason for \( GAUSS \) to score highly on this measure is most likely due to the scoring of word pairs as described in Appendix D. This is supported by the observation that with some very rare words, the effect of \( \epsilon \) in the NPMI score in \( C_v \) is large, resulting in higher scores than expected. This is described in [https://github.com/dice-group/Palmetto/issues/12](https://github.com/dice-group/Palmetto/issues/12).

### G Topics for 20NG
### Adaptive LDA (ADAP)

| Topic:0 | Topic:1 | Topic:2 | Topic:3 | Topic:4 | Topic:5 | Topic:6 |
|---------|---------|---------|---------|---------|---------|---------|
| jesus   | price   | religion| color   | israel  | win     | car     |
| bible   | bike    | true    | data    | war     | april   | power   |
| church  | money   | faith   | video   | jews    | white   | heard   |
| christian| sell | real    | power   | jewish  | night   | local   |
| christ  | cost    | agree   | mode    | american| close   | speed   |
| sin     | list    | argument| set     | israeli | gun     | cars    |
| life    | worth   | evidence| bit     | country | period  | miles   |
| gods    | pay     | truth   | software| armenians| red    | model   |
| word    | insurance| exist   | systems | university| steve   | friend  |
| earth   | shipping| religious| apple   | arab    | record  | deleted |
| christians | mark | reading | control| usa     | start  | engine  |
| lord    | market  | science | serial  | turkish | boston | ford    |
| heaven  | paid    | person  | speed   | armenian| arms   | told    |
| live    | business| belief  | standard| greek   | pens   | service |
| john    | ride    | theory  | output  | muslims | guns   | bought  |
| paul    | deal    | moral   | science | countries| cut    | weeks   |
| christianity | prices | statement| light   | national| straight| driving |
| spirit  | quality | values  | current | canada  | pts    | stuff   |
| sense   | buying  | claim   | simple  | press   | congress| dealer  |
| sense   | extra   | christians| fine    | germany | pittsburgh| couple  |

| $S_k$:0.292 | $S_k$:0.03 | $S_k$:0.139 | $S_k$:0.085 | $S_k$:0.016 | $S_k$:0.108 | $S_k$:0.136 |
|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| wc:2789     | wc:2720     | wc:2886     | wc:34212    | wc:14319    | wc:2178     | wc:2750     |

| Topic:7 | Topic:8 | Topic:9 | Topic:10 | Topic:11 | Topic:12 |
|---------|---------|---------|---------|---------|---------|
| lost    | original| book    | support | key     | games   |
| called  | idea    | sale    | image   | chip    | team    |
| left    | set     | offer   | info    | space   | players |
| hit     | love    | condition| sun     | clipper | season  |
| bad     | answer  | books   | graphics| encryption| hockey |
| start   | bad     | excellent| appreciated| government| play    |
| happen  | hear    | software| based   | phone   | play    |
| base    | sort    | mouse   | university| message | mike    |
| started | sound   | manual  | technology| algorithm| baseball |
| pitcher | solution| graphics| programs| security | league  |
| single  | light   | includes| convert | data    | series  |
| watch   | thinking| send    | job     | source  | teams   |
| field   | hate    | tape    | wondering| nsa     | runs    |
| cubs    | head    | complete| images  | david   | fan     |
| expect  | position| event   | product | press   | average |
| braves  | ideas   | items   | conference| des     | fans    |
| major   | times   | title   | design  | secure  | nhl     |
| minutes | true    | events  | tiff    | chips   | pick    |
| performance | reason | brand  | june    | launch  | goal    |
| james   | stuff   | manuals | gif     | agencies| guy     |

| $S_k$:0.001 | $S_k$:0.055 | $S_k$:0.017 | $S_k$:0.068 | $S_k$:0.189 | $S_k$:0.238 |
|-------------|-------------|-------------|-------------|-------------|-------------|
| wc:2344     | wc:2373     | wc:2478     | wc:2312     | wc:2984     | wc:3725     |

2464
### Adaptive LDA (ADAP)

| Topic:13 | Topic:14 | Topic:15 | Topic:16 | Topic:17 | Topic:18 | Topic:19 |
|----------|----------|----------|----------|----------|----------|----------|
| file     | list     | day      | card     | post     | person   | law      |
| window   | version  | bit      | scsi     | lot      | reason   | government |
| dos      | send     | remember | monitor  | note     | simply   | public   |
| running  | address  | current  | mac      | wrong    | true     | private  |
| code     | mail     | lot      | ram      | change   | children | federal  |
| machine  | post     | correct  | controller | guess   | wrong    | police   |
| screen   | subject  | road     | memory   | real     | feel     | house    |
| software | faq      | days     | ide      | understand | hand   | legal    |
| server   | posted   | company  | data     | sounds   | life     | laws     |
| error    | posting  | stuff    | bus      | yeah     | times    | court    |
| disk     | drivers  | difference | drives | add     | called   | weapons  |
| display  | ftp      | fast     | vga      | figure   | remember | clinton  |
| format   | driver   | office   | port     | basically | death   | class    |
| set      | internet | talking  | modern   | thread   | day      | warrant  |
| keyboard | reply    | notice   | cards    | hard     | questions | authority |
| disks    | lines    | mentioned | disk    | main     | bad      | citizens |
| size     | called   | start    | meg      | reason   | issue    | national |
| box      | printer  | dave     | standard | discussion | fbi    | president |
| manager  | dod      | type     | dos      | agree    | news     | tax      |

\[
S_k:0.2 \quad S_k:0.086 \quad S_k:0.042 \quad S_k:0.251 \quad S_k:0.011 \quad S_k:0.083 \quad S_k:0.136
\]

\[
wc:4613 \quad wc:3344 \quad wc:2532 \quad wc:4209 \quad wc:2800 \quad wc:49408 \quad wc:2531
\]

### Discrete LDA (DISC)

| Topic:0 | Topic:1 | Topic:2 | Topic:3 | Topic:4 | Topic:5 | Topic:6 |
|---------|---------|---------|---------|---------|---------|---------|
| car     | power   | war     | post    | book    | religion | key     |
| cars    | battery | armenians | posting | lost    | rights   | government |
| engine  | light   | turkish | list    | study   | gun      | chip    |
| miles   | design  | armenian | article | msg     | government | public |
| speed   | idea    | muslims | send    | pain    | news     | clinton |
| driving | bit     | population | source | york    | support  | encryption |
| ford    | quality | jewish | mail    | school  | control  | phone   |
| oil     | single  | answer  | questions | kids   | article  | security |
| heavy   | type    | history | address | disease | post     | police  |
| clean   | model   | killed  | hope    | drug    | freedom  | private |
| rear    | noise   | source  | posted  | books   | guns     | algorithm |
| box     | heater  | turkey  | note    | usa     | court    | message |
| manager | dod     | type    | dos     | agree   | news     | tax     |

\[
S_k:0.062 \quad S_k:0.0108 \quad S_k:0.106 \quad S_k:0.107 \quad S_k:0.202 \quad S_k:0.037 \quad S_k:0.246
\]

\[
wct:6189 \quad wct:6973 \quad wct:6362 \quad wct:6503 \quad wct:5782 \quad wct:7046 \quad wct:7628
\]

2465
### Discrete LDA (DISC)

| Topic:7 | Topic:8 | Topic:9 | Topic:10 | Topic:11 | Topic:12 | Topic:13 |
|---------|---------|---------|----------|----------|----------|----------|
| bike    | jesus   | local   | space    | games    | window   | file     |
| stuff   | church  | friend  | university| company | players  | code     |
| real    | faith   | book    | april    | play     | screen   | software |
| lot     | christian | weeks | technology | win     | application | graphics |
| level   | bible   | dealer  | press    | season   | color    | version  |
| times   | ride    | love    | check    | conference| hockey   | display  |
| guess   | thinking | human  | hand     | science  | league   | server   |
| sort    | deleted | goods   | talk     | radio    | runs     | images   |
| short   |field    | true    | picture  | station  | pick     | user     |
| left    | hit     | person  | bought   | office   | boston   | manager  |
| canada  | word    | yeah    | force    | mike     | event    | applications |

| $S_k$ | wc | $S_k$ | wc | $S_k$ | wc | $S_k$ | wc |
|-------|----|-------|----|-------|----|-------|----|
| 0.06  | 7097 | 0.251 | 5690 | 0.119 | 6855 | 0.038 | 9185 |
| 0.262 | 10345| 0.187 | 7249 | 0.202 | 8386 |

---

### Discrete LDA (DISC)

| Topic:14 | Topic:15 | Topic:16 | Topic:17 | Topic:18 | Topic:19 |
|----------|----------|----------|----------|----------|----------|
| card     | price    | day      | true     | israel   | remember |
| monitor  | sale     | feel     | death    | claim    | pretty   |
| sci      | offer    | law      | matter   | fbi      | bad      |
| mac      | sell     | remember | theory   | israeli  | hear     |
| video    | condition | water    | argument | evidence | days     |
| apple    | list     | talking  | correct  | happened | chance   |
| machine  | money    | food     | dead     | arab     | heard    |
| drivers  | shipping | word     | position | jews     | day      |
| controller | cost   | days     | reason   | started  | guys     |
| ram      | box      | called   | completely | claims  | feel     |
| mode     | sold     | common   | sex      | gas      | deal     |
| board    | excellent | written | homosexual | agree   | understand |
| drives   | pay      | language | evidence | leave    | worth    |
| bus      | power    | article  | issue    | countries | lot      |
| data     | includes | sense    | change   | comment  | minutes  |
| driver   | blue     | die      | note     | peace    | clinton  |
| ide      | stuff    | term     | wrong    | statement | reason   |
| speed    | original | supposed | nature   | children | watching |
| memory   | selling  | week     | sexual   | response | night    |
| modern   | including | english | statements | policy  | wait     |

| $S_k$ | wc | $S_k$ | wc | $S_k$ | wc | $S_k$ | wc |
|-------|----|-------|----|-------|----|-------|----|
| 0.221 | 10829| 0.163 | 7104| 0.02  | 6217| 0.056 | 7824|
| 0.116 | 7575 | 0.032 | 6668|        |     |        |    |
Static mix $\pi$, (SMIX)

| Topic:0 | Topic:1 | Topic:2 | Topic:3 | Topic:4 | Topic:5 | Topic:6 |
|---------|---------|---------|---------|---------|---------|---------|
| period  | lost    | andor   | power   | government | legal   | bit     |
| paul    | form    | close   | car     | public   | court   | die     |
| software| map     | command | card    | local    | judge   | advance |
| handbook| pens    | class   | bit     | university| trial   | address |
| book    | fall    | shift   | software| children | justice | count   |
| rules   | rob     | bds     | window  | israel   | criminal| kent    |
| held    | force   | virtual | key     | science  | federal | hardware|
| offers  | flag    | event   | price   | church   | supreme | clinton |
| final   | list    | managed | space   | press    | amendment| cut     |
| modern  | support | black   | monitor | country  | police  | string  |
| study   | named   | win     | speed   | american | courts  | bds     |
| shadow  | register| myers   | disk    | house    | lawyers | est     |
| riding  | cat     | string  | machine | war      | convicted | van   |
| bowman  | press   | center  | advance | jewish   | gun     | programmer |
| swift   | phone   | ticket  | color   | jews     | weaver  | und     |
| graham  | bear    | friendly | sale    | religious | crimes | les     |
| happen  | code    | morning | code    | private  | offensive| ground |
| writing | student | taurus  | screen  | history  | lawyer  | nlt     |
| tia     | straight| lot     | phone   | national | jury    | des     |
| manual  | table   | weight  | systems | israeli  | closed  | internet|

$S_k$:0.38  $S_k$:0.523  $S_k$:0.483  $S_k$:0.061  $S_k$:0.024  $S_k$:0.187  $S_k$:0.484

wc:423  wc:474  wc:590  wc:36446  wc:15595  wc:922  wc:486

Static mix $\pi$, (SMIX)

| Topic:7 | Topic:8 | Topic:9 | Topic:10 | Topic:11 | Topic:12 |
|---------|---------|---------|----------|----------|----------|
| john    | jesus   | start   | pittsburgh | times    | left     |
| david   | bible   | started | boston   | manager  | harry    |
| michael | christians | starting | san     | local    | det      |
| james   | christ  | active  | york     | picture  | eric     |
| andrew  | gods    | session | texas    | hold     | pre      |
| robert  | sin     | gordon  | chicago  | power    | wife     |
| peter   | heaven  | sad     | detroit  | capitol  | died     |
| joseph  | christianity | banks | toronto  | sites    | att      |
| daniel  | holy    | surrender | angeles | tim      | jason    |
| matthew | scripture | cursor | los     | names    | ted      |
| patrick | lord    | weeks   | buffalo  | master   | forged   |
| stephen | sabbath | root    | francisco | managed | bing     |
| francis | church  | helped  | montreal | string   | spot     |
| charles | resurrection | closed | philadelphia | slave | van     |
| martin  | son     | stopped | baltimore | finland | viola    |
| graeme  | scriptures | stopping | louis  | jumper   | managed  |
| lewis   | atheists | responsible | jose  | pay      | roommate |
| alan    | biblical | defending | minnesota | location | courier |
| richard | spiritual | aura    | red      | beach    | har     |
| craig   | mary    | traders | vancouver | creation | maria    |

$S_k$:0.529  $S_k$:0.288  $S_k$:0.503  $S_k$:0.258  $S_k$:0.437  $S_k$:0.751

wc:1214  wc:1945  wc:963  wc:1424  wc:545  wc:627
| Static mix \( \pi \), (SMIX) | Topic:13 | Topic:14 | Topic:15 | Topic:16 | Topic:17 | Topic:18 | Topic:19 |
|-----------------------------|---------|---------|---------|---------|---------|---------|---------|
| chip file major day mike games include | dos set pink days steve team including | graphics data track april dave times includes | sci true send months jim play included | keyboard post fbi night bob players features | hardware lot thrush week chris lost listed | unix hard minor weeks tom series edition | floppy list article month ron runs refer | ibm bad insurance hours larry win typical |
| chip file major day mike games include | dos set pink days steve team including | graphics data track april dave times includes | sci true send months jim play included | keyboard post fbi night bob players features | hardware lot thrush week chris lost listed | unix hard minor weeks tom series edition | floppy list article month ron runs refer | ibm bad insurance hours larry win typical |
| chip file major day mike games include | dos set pink days steve team including | graphics data track april dave times includes | sci true send months jim play included | keyboard post fbi night bob players features | hardware lot thrush week chris lost listed | unix hard minor weeks tom series edition | floppy list article month ron runs refer | ibm bad insurance hours larry win typical |

Gaussian LDA, (GAUS) | Topic:0 | Topic:1 | Topic:2 | Topic:3 | Topic:4 | Topic:5 | Topic:6 |
|-----------------|---------|---------|---------|---------|---------|---------|---------|
| file wrong american software government university games | image real stuff country files list science | include remember canada dos federal school baseball | images included person countries dos graphics national congress | image real stuff country files list science | include remember canada dos federal school baseball | images included person countries dos graphics national congress | image real stuff country files list science | include remember canada dos federal school baseball |
| image real stuff country files list science | include remember canada dos federal school baseball | images included person countries dos graphics national congress | image real stuff country files list science | include remember canada dos federal school baseball | images included person countries dos graphics national congress | image real stuff country files list science | include remember canada dos federal school baseball | images included person countries dos graphics national congress |
| image real stuff country files list science | include remember canada dos federal school baseball | images included person countries dos graphics national congress | image real stuff country files list science | include remember canada dos federal school baseball | images included person countries dos graphics national congress | image real stuff country files list science | include remember canada dos federal school baseball | images included person countries dos graphics national congress |

wc:3145 wc:64156 wc:432 wc:2285 wc:1562 wc:13073 wc:1200
| Gaussian LDA, (GAUS) | \( S_k:-0.005 \) | \( S_k:-0.008 \) | \( S_k:-0.013 \) | \( S_k:-0.259 \) | \( S_k:-0.085 \) | \( S_k:-0.158 \) |
|----------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| \( w_c:2507 \)       | \( w_c:37857 \) | \( w_c:1808 \)  | \( w_c:3049 \)  | \( w_c:2601 \)  | \( w_c:2230 \)  |                  |

| Gaussian LDA, (GAUS) | \( S_k:0.05 \)  | \( S_k:0.048 \) | \( S_k:0.013 \) | \( S_k:0.078 \)  | \( S_k:-0.151 \) | \( S_k:-0.536 \) | \( S_k:-0.239 \) |
|----------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| \( w_c:1793 \)       | \( w_c:68542 \) | \( w_c:1992 \)  | \( w_c:2166 \)  | \( w_c:1848 \)  | \( w_c:1702 \)  | \( w_c:1584 \)  |                  |