Adaptive Mixed Component LDA for Low Resource Topic Modeling

Suzanna Sia **Kevin Duh**
Johns Hopkins University
Baltimore, MD, USA
ssia1@jhu.edu kevinduh@cs.jhu.edu

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). [1] In such scenarios, word co-occurence 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; Batmanghelich 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

[1] This differs from the short text setting which has a large number of train documents, that has been addressed by multiple work (Li et al., 2016a; Qiang et al., 2020).

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2451–2469
April 19 - 23, 2021.

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, $\Phi = \{\phi_1, \cdots, \phi_K\}$. Θ and Φ 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 $\mathbf{v} \in \mathbb{R}^M$ 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 θ_d, is shared across languages, and that each language specific topic has a multinomial topic-word distribution, $\Phi^{\ell_1}, \Phi^{\ell_2}$ 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 $\mu_k \sim \mathcal{N}(\mu_0, \frac{1}{\kappa}\Sigma_k)$
3. Draw topic-word distribution $\phi_k \sim Dir(\lambda)$

For each document d in corpus C
1. Draw a topic distribution $\theta_d \sim Dir(\eta)$
2. For each word $w_{d,i}$
   (a) Draw topic $z_{d,i} \sim Multin(\theta_d)$
   (b) Draw $\pi \sim Beta(\alpha, \beta)$
   (c) With π, draw $w_{d,i} \sim Multin(\phi_{z_{d,i}})$
   (d) With $(1-\pi)$, draw $\mathbf{v}_{d,i} \sim \mathcal{N}(\mu_{z_{d,i}}, \Sigma_{z_{d,i}})$

where $\mathcal{W}^{-1}$ is the Inverse Wishart distribution, Ψ is the normalised Precision matrix, ν_0 is degrees of freedom, μ_0 is the prior mean for each Gaussian topic, and π is a mixture coefficient.

3.1 Gibbs Sampling for Posterior Inference

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

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 $\backslash$):

$$\begin{aligned} &p(z_{d,i}=k|\mathbf{z}_{\backslash d,i}, \mathbf{x}, \eta, \varphi) \\ &\propto p(x_{d,i}|z_{d,i}=k, \varphi, \mathbf{z}_{\backslash d,i}, \mathbf{x}) \\ &\times p(z_{d,i}=k|\eta, \mathbf{z}_{\backslash d,i}, \mathbf{x}) \end{aligned} \quad (1)$$

η is the corresponding parameters of a Dirichlet prior for the document-topic distribution θ, and φ are parameters associated with the topic-word distribution. This is either λ for the Dirichlet prior for multinomial ϕ, or $\mu_0, \Sigma_0, \nu_0, \kappa$ for the Gaussian. In our proposed model (section 5), φ 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 φ_1 and f_2 be a continuous density function with parameters φ_2. The density can be expressed as a convex combination:

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

Then, the second term in Equation 3 can be expressed as the density of $\mathbf{v}_{d,i}$ under the multivariate t distribution[4] parameterised by mean μ_k and covariance $\frac{\kappa_k+1}{\kappa_k}\Sigma_k$, with ν_k degrees of freedom. κ is a prior confidence on μ_k and Σ_k (Murphy, 2012). $\varphi = \{\lambda, \nu_0, \mu_0, \Sigma_0, \kappa\}$, including parameters of both the Dirichlet and Gaussian priors, with the subscript $_0$ indicating parameters of the conjugate prior. N indicates counts; for the first term in the RHS of Equation 3, $N_{k,w_{d,i}}$ are the counts of that particular word type (for the token $w_{d,i}$) assigned to topic k, and $N_{k,w'}$ is the number of counts of word type w' assigned to topic k, with V being the vocabulary.[5]

$$\begin{aligned} &p(x_{d,i}|z_{d,i}=k, \varphi, \mathbf{z}_{\backslash d,i}, \mathbf{x}) \\ &\propto \pi \frac{N_{k,w_{d,i}} + \lambda_{w_{d,i}}}{\sum_{w'}^{|V|} N_{k,w'} + \lambda_{w'}} \\ &+ (1-\pi) t_{\nu_k}(\mathbf{v}_{d,i}|\mu_k, \frac{\kappa_k+1}{\kappa_k}\Sigma_k) \end{aligned} \quad (3)$$

Table 1 summarises current and previous work with respect to Equation 1.

4 Perspectives on Mixture Coefficient π

4.1 Perspectives on Static π

There are several ways to interpret the mixture coefficient π 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 π is either 1 or 0 respectively. π 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 π 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 π is proportional to prior times likelihood: $p(\pi|o) \propto p(\pi)p(o|\pi)$.

[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 λ, they are integers that index a count vector or Dirichlet parameter vector λ. 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.

LDA types	Topic-Word	$p(x^{\ell_1}_{d,i} \mid z^{\ell_1}_{d,i} = k, \varphi, \mathbf{z}_{\setminus d,i}, \mathbf{x}) \propto$	$p(z^{\ell_1}_{d,i} = k \mid \eta, \mathbf{x}) =$
Discrete	Discrete	$\mathbb{D} = \frac{N_{k,w_{d,i}} + \lambda_{w_{d,i}}}{\sum_{w'}^{\lvert V \rvert} N_{k,w'} + \lambda_{w'}}$	$\frac{N^{d}_{k\setminus d,i} + \eta_k}{\sum_{k'} N^{d}_{k'\setminus d,i} + \eta - 1}$
Gaussian	Continuous	$\mathbb{C} = t_{\nu_k}(\mathbf{v}_{d,i} \mid \mu_k, \frac{\kappa_k + 1}{\kappa_k}\Sigma_k)$	
Static π	Mixture	$(1-\pi)\cdot(\mathbb{C}) + \pi \cdot (\mathbb{D})$	
Adaptive π	Mixture	$(1-\pi_{k,j})\cdot(\mathbb{C}) + \pi_{k,j} \cdot (\mathbb{D})$	
Polylingual	Discrete	$\mathbb{L} = \frac{N^{\ell_1}_{k,w_{d,i}} + \lambda^{\ell_1}_{w_{d,i}}}{\sum_{w'}^{\lvert V^{\ell_1} \rvert} N^{\ell_1}_{k,w'} + \lambda^{\ell_1}_{w'}}$	$\frac{N^{\ell_1,d}_{k\setminus d,i} + N^{\ell_2,d}_{k} + \eta_k}{\sum_{k'} N^{\ell_1,d}_{k'\setminus d,i} + \sum_{k'} N^{\ell_2,d}_{k'} + \eta - 1}$
ℓ-Adapt	Mixture	$(1-\pi^{\ell_1}_{k,j}) t_{\nu^{\ell}_k}(\mathbf{v}^{\ell_1}_{d,i} \mid \mu_k, \frac{\kappa_k+1}{\kappa_k}\Sigma_k) + \pi^{\ell_1}_{k,j} \cdot (\mathbb{L})$	

Table 1: **Comparison of various LDA models.** For topic k, language ℓ_1, $N^{\ell_1,d}_{k}$ are counts of topic k in document d, $N^{\ell_1}_{k,w_{d,i}}$ 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_{ν_k} is the probability density function of the multivariate t distribution parameterised by ν_k degrees of freedom, mean μ_k and covariance Σ_k. $\kappa_k = \kappa + \sum_{w'}^{\lvert V \rvert} N_{k,w'}$, where κ represents the belief on the prior of the multivariate Gaussian. For the cross-lingual model, ν^{ℓ} and κ^{ℓ}_k sum counts in ℓ_1 and ℓ_2. "\" denotes counts when excluding that variable. $\lambda \in \mathbb{R}^{\lvert V \rvert}$ and $\eta \in \mathbb{R}^{K}$ are hyperparameters of the Dirichlet prior distribution on topic-word and document-topic distribution respectively.

Here the observations o correspond to the discrete and continuous representations.

It can be shown due to conjugacy of the beta-binomial distribution that when the prior $p(\pi)$ is $Beta(\alpha_0, \beta_0)$, the posterior $p(\pi|o)$ is also a Beta distribution, where α' and β' are counts of words that have a discrete and continuous representation available, and α_0 and β_0 are set to 1 in the absence of any information.

$$\pi \sim Beta(\alpha_0 + \alpha', \beta_0 + \beta') \tag{4}$$

Note that with modern word embeddings such as FastText,[6] and Byte Pair Encoding methods, both discrete and continuous representations are mostly always observed together and $|V| = \alpha' \approx \beta'$, when $|V|$ is large, $\mathbb{E}[\pi] = 0.5$ with $Var[\pi] \approx 0$. Unfortunately, this view is overly "naive" as the continuous representations are not true observations, but learned representations which should not constitute full observation counts. We refer to this setting as "Static Mixing (SMIX)" in section 8, where we directly adopt $\pi = 0.5$. [7]

5 Adaptive Mixture Coefficient $\pi_{k,j}$

We recommend a more pragmatic view for balancing between (noisy) learned word embeddings and discrete counts by modeling the mixture coefficient as a topic k and word type indexed by j, $\pi_{k,j}$ specific random variable. At inference time, we sample $\pi_{k,j} \in [0, 1]$ from a $Beta$ distribution that is specific to each word type and each topic for the α parameter, and topic specific for the β parameter to compute Equation 3.

$$\pi_{k,j} \sim Beta(\alpha_{k,j}, \beta_k) \tag{5}$$

The parameter $\alpha_{k,j}$ represents the concentration on the discrete representation, while β_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 β 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 β_k is updated based on the uncertainty in the t distribution as measured by the trace of the covariance matrix Σ_k.

5.1 Adaptive $\alpha_{k,j}$ Based on Counts

We specify corpus specific 'α' priors, α^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 $|V|$.

$$\alpha^0_j = \frac{D}{|V|} + \frac{\sum_k N_{k,j}}{K} \tag{6}$$

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

[6] FastText can generate a representation for previously unseen vocabulary words based on their character Ngrams.

[7] For a vocabulary size of just 1000, $Var[\pi] = 0.00026$.

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

$$\alpha_{k,j} = \alpha_j^0 + N_{k,j} \tag{7}$$

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

5.2 Adaptive β_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 β_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 Σ_k, equivalently written as $tr(\sqrt{\Sigma_k})$. We formulate β_k as depending on the constant terms $\frac{M}{K}$, M is the number of dimensions of the multivariate Gaussian, and (non-constant) Σ_k which is updated at every step of Gibbs Sampling:

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

The intuition for the inverse relationship between $tr(\sqrt{\Sigma_k})$ and β_k is as follows. If the topic has high variance, then β_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.

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

We elaborate on the 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.

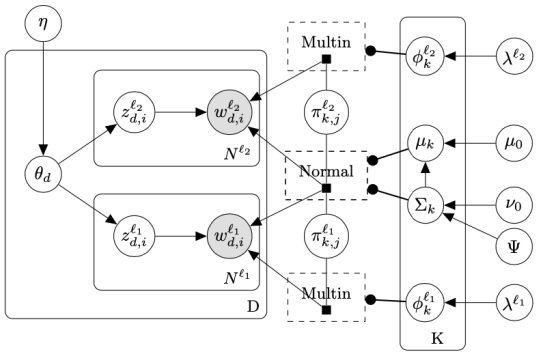


Figure 1: Cross-Lingual Adaptive LDA, with shared continuous parameters μ_k, Σ_k across languages and adaptive $\pi_{k,j}$ for every word type j and topic k. The word type j corresponds to the i^{th} token of document d. $w_{d,i}$ indicates a token when it is not being used as a subscript.

6 Computational Complexity

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

Since the covariance matrix Σ_k is symmetric and positive semi-definite, we can utilise the Cholesky decomposition where Σ_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 β_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 $p(\mathbf{v}_{d,i}|z_{d,i} = k, \varphi, \mathbf{x})$ 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 θ_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 $\mathbf{v}^{\ell_1}, \mathbf{v}^{\ell_2}$ have been mapped to the same embedding space, by adopting shared Gaussian mean μ_k and covariance Σ_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:

$$\alpha_{k,j}^{\ell} = \frac{D}{|V^{\ell}|} + \frac{\sum_{k'}^{K} N_{k',j}^{\ell}}{K} + N_{k,j}^{\ell}$$
$$\beta_{k}^{\ell} = \left(\frac{M}{K \cdot |\mathcal{L}|}\right)\left(tr\left(\sqrt{\Sigma_k}\right)\right)^{-1} \quad (10)$$

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[9] 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, ℓ- 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 π)

8.2 Datasets

We use the **20 newsgroup dataset (20NG)** which is a common text analysis dataset containing around 18000 documents and 20 categories.[10] 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 samping 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]

[9] Code made available at https://github.com/suzyahyah/adaptive_mixture_topic_model

[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-Colorado, 2018) which has a corpus size of around 5000 documents.

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 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).

ℓ	No. docs	DISC	GAUS	SMIX	ADAP
ro	1000	-0.014	-0.134	-0.139	**0.012**
fr	1000	-0.010	-0.172	-0.151	**0.005**
pl	1000	-0.030	-0.138	-0.300	**-0.011**
es	1000	**0.010**	-0.280	-0.119	0.008
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***

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).

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 μ_0 and covariance Σ_0 to the empirical mean and sample covariance respectively based on random assignment of words to topics. Following Das et al. (2015), we initialise κ to 1, ν_0 to the embedding size M of 300. Parameters of the Dirichlet prior η and λ 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.

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-occurences and 1 indicates 100% co-occurences. [12] The score of each topic S_k is computed from word pair combinations of the top T words returned by that topic.

$$S_k = \sum_{i=2}^{T} \sum_{j=1}^{i-1} NPMI(w_i, w_j) \quad (11)$$

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

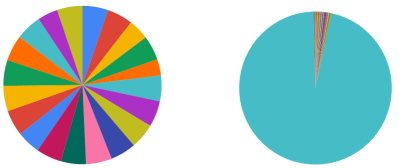
We extract word co-occurrence statistics of the held-out documents to compute $p(w_i)$ and $p(w_i, w_j)$, and set ϵ to $1e^{-12}$ to avoid logarithm of 0. NPMI averaged across all topics are reported as $\frac{1}{K} \sum_k S_k$ in Table 2 and 3.

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.

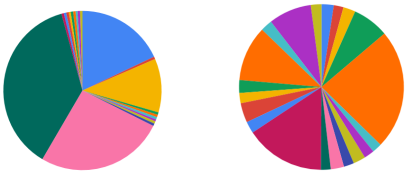
8.4 Results and Discussion

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

[12] Hao et al. (2018) introduced a multilingual NPMI for low resource languages and proposed the bible as a held-out test set, but note that it is "archaic" - good at evaluating topics such as family and religion but poor at evaluating modern topics such as biology. We use regular NPMI for consistency with the monolingual setup.



(a) DISC, $\pi = 1$ (b) GAUS, $\pi = 0$



(c) SMIX, Static π (d) ADAP dynamic π

Figure 2: Topic proportions for various mixture coefficients (π) 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.."[13] which are not representative of themes. In this work, we evaluate using normalized PMI (Bouma, 2009) which corrects for this. This is somewhat suprising 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.

[13] See Table 1 of Das et al. (2015)

No. train docs = 1000		No. train docs = 7000	
ADAP	DISC	ADAP	GAUS
government	**law**	jesus	john
law	**government**	word	**paul**
public	color	christ	james
laws	remember	bible	**mary**
crime	days	sin	smith
court	idea	christians	andrew
legal	told	death	gordon
trust	post	**paul**	norton
police	list	church	thomas
fbi	process	**mary**	george

Table 4: Top topic words on 20NG, bolded words are common across both topics. ADAP (Adaptive π) 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 neighbouring 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-occurence 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).[14] If many words have been assigned to one topic, that topic covariance Σ_k becomes much larger than the others, leading to subsequent $\mathbf{v}_i$ then having a higher relative density under that topic during Gibbs Sampling.

Our proposed adaptive π (Figure 2d) counteracts this effect better than the static π (Figure 2c). If Σ_k is large, to balance the effect of words having a higher density under topic k, the algorithm samples a larger π, 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-occurences. 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 α, β, π under a varying number of topics (K from 20 to 200).

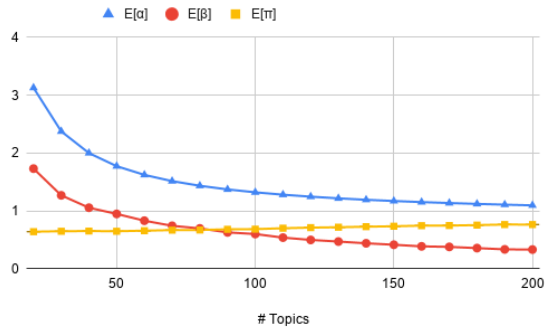


Figure 3: Stability of adaptive mixture coefficient $\pi_{i,k}$ with increasing number of topics in 20NG using 7000 documents. All α, β, π are expected values across all vocabulary words and all topics. We observe that the expected values vary smoothly with increasing K.

We approximate the expectation by the arithmetic average: $\mathbb{E}[\alpha] = \frac{1}{K}\frac{1}{|V|}\sum_k^K \sum_j^{|V|} \alpha_{k,j}$ for a fixed K, where $\mathbb{E}[\pi]$ and $\mathbb{E}[\beta]$ are calculated in the same manner. We verified that as K increases, the variance of π increases as expected, as reflected by the smoothly decreasing $\mathbb{E}[\alpha]$ and $\mathbb{E}[\beta]$.

Note that α and β take on different values for each word and topic during Gibbs Sampling. We observe that while $\mathbb{E}[\pi]$ 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 π.

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-*

[14] This is still true for $K = 50$.

ation for Computational Linguistics. Meeting, volume 2016, page 537. NIH Public Access.

Christopher M Bishop. 2006. *Pattern recognition and machine learning*. springer.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 795–804.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. Topic modeling in embedding spaces. *arXiv preprint arXiv:1907.04907*.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian data analysis*. Chapman and Hall/CRC.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Shudong Hao, Jordan Boyd-Graber, and Michael J Paul. 2018. Lessons from the bible on modern topics: Low-resource multilingual topic model evaluation. *arXiv preprint arXiv:1804.10184*.

Xuemeng Hu, Rui Wang, Deyu Zhou, and Yuxuan Xiong. 2020. Neural topic modeling with cycle-consistent adversarial training. *arXiv preprint arXiv:2009.13971*.

Matthew L Jockers and David Mimno. 2013. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769.

Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016a. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174.

Ximing Li, Jinjin Chi, Changchun Li, Jihong Ouyang, and Bo Fu. 2016b. Integrating topic modeling with word embeddings by mixtures of vmfs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 151–160.

Luyang Liu, Heyan Huang, Yang Gao, Yongfeng Zhang, and Xiaochi Wei. 2019. Neural variational correlated topic modeling. In *The World Wide Web Conference*, pages 1142–1152.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical word embeddings. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Jeffrey Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. 2017. Tandem anchoring: A multiword anchor approach for interactive topic modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 896–905.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 880–889. Association for Computational Linguistics.

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*.

Borja Navarro-Colorado. 2018. On poetic topic modeling: extracting themes and motifs from a corpus of spanish poetry. *Frontiers in Digital Humanities*, 5:15.

Graham Neubig and Chris Dyer. 2016. Generalizing and hybridizing count-based and neural language models. *arXiv preprint arXiv:1606.00499*.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from wikipedia. In *Proceedings of the 18th international conference on World wide web*, pages 1155–1156. ACM.

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. Graphbtm: 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 $\mathscr{D} = \{doc_1, \cdots, doc_D\}$,
Vocab $V = \{w_1, \cdots, w_{|V|}\}$,
Embeddings $\{\mathbf{v}_1, \cdots, \mathbf{v}_{|V|}\}, \mathbf{v} \in \mathbb{R}^M$
Result: $\Phi = \{\phi_1, \cdots, \phi_K\}, \Theta = \{\theta_1, \cdots, \theta_D\}$

Initialization:
$\mu_0 \leftarrow \frac{1}{|V|}\sum_{j=1}^{|V|} \mathbf{v}_j$ (prior mean)
$\Sigma_0 \leftarrow \frac{1}{|V|-1}\sum_{j=1}^{|V|}(\mathbf{v_j} - \mu_0)(\mathbf{v_j} - \mu_0)^T$ (prior cov)
for *doc* $d \in \{1, \cdots, D\}$, *word* i **do**
  $z_{d,i} \leftarrow k$ uniform sample from $\{1, \cdots, K\}$
for *topic* $k \in \{1, \cdots, K\}$ **do**
  $N_k \leftarrow |\{z_{d,i} = k | \forall i \in |d|, \forall doc_d \in \mathscr{D}\}|$
  $\mu_k \leftarrow \frac{\kappa\mu_0 + N_k\bar{\mathbf{v}}_k}{\kappa_k}$
  $\Sigma_k \leftarrow \frac{\Psi_k}{\nu + N_k - M + 1}$
  $L_k^T L_k \leftarrow \Sigma_k$ (cholesky decomposition)
  $N_{k,j} \leftarrow |\{z_{d,i} = k | \forall doc_d \in \mathscr{D}\}|$ for each $w_j \in V$, j is word type of token $w_{d,i}$
  $N_k^d \leftarrow |\{z_{d,i} = k | \forall\ w_i \text{ in } doc_d\}|$ for each $doc_d \in \mathscr{D}$
while *iter<maxiter or not converged* **do**
  for *doc* $d \in \{1, \cdots, D\}$, *word* i *in doc* **do**
    $z_{old} \leftarrow z_{d,i}$
    Decrement by 1, $N_{z_{old}}, N_{z_{old},j}, N_{z_{old}}^d$
    Update $\alpha_{z_{old},j}, \beta_{z_{old}} (Eq: 7, 9)$
    Update $\mu_{z_{old}}, L_{z_{old}}$
    for *topic* $k \in \{1, \cdots, K\}$ **do**
      Sample $\pi_{k,j} \sim Beta(\alpha_{k,j}, \beta_k)$
      Compute $\_p(z_{d,i} = k | \varphi, \mathbf{x}) (Eq: 1, 3$
)
  Sample $z_{new} \sim p(z_{d,i} | \varphi, \mathbf{x})$
  Increment by 1, $N_{z_{new}}, N_{z_{new},j}, N_{z_{new}}^d$
  Update $\alpha_{z_{new},j}, \beta_{z_{new}} (Eq: 6, 8)$
  Update $\mu_{z_{new}}, L_{z_{new}}$

The full inference algorithm is given in Algorithm 1. For details on the parameterisation of the multivariate t, update of μ_k and computation of Ψ_k, we refer readers to Murphy (2012). For update and downdates 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.[15] 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, $p(\mathbf{v}_{d,i}^{\ell_1} | z_{d,i}^{\ell_1} = k, \varphi, \mathbf{z}_{\setminus d,i}, \mathbf{x})$ becomes highly skewed towards a certain topic, such that the influence of the document structure becomes negligent. 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 ν_k to better account for uncertainty in the embedding representations.

B.1 Rescaling the Degrees of Freedom ν_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, $\mathbb{E}[N_k] = \frac{|\tilde{V}|}{K}$, where $|\tilde{V}|$ are all the (non-unique) words in the corpus. Since for a typical corpus $|\tilde{V}|$ is very large and $\frac{|\tilde{V}|}{K} >> M$, the degrees of freedom ν_k are very large resulting in an approximate normal distribution which is over-confident in its posterior predictions. This effectively dominates the priors for Σ_0, ν_0 or μ_0. Hence, we rescale ν_k to $\hat{\nu_k}$ from 1 to 30[16] to account for inherent uncertainty over $\mathbf{v}_i$ belonging to any particular topic.

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

Comparison with the fully Bayesian treatment. We found this heuristic to be numerically and em-

[15] Readers are referred Murphy (2012) for an exposition on the form for posterior inference under a Gaussian Prior.

[16] As the degrees of freedom increase, the t distribution approaches the normal distribution. $\nu \geq 30$ is a rule of thumb for when the difference between the t distribution and normal distribution becomes negligible.

pirically more stable than a fully Bayesian treatment which encodes a higher variance in the t distribution by having a larger prior on the covariance matrix Σ_0.

First, re-estimating the covariance matrix at every step of Gibbs Sampling is numerically unstable with a large Σ_0. Next, rescaling ν_k guarantees that we maintain a heavy-tailed t 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 Σ_0 to its sample covariance, removing one adhoc parameter choice. Since both setting a large prior Σ_0 and scaling ν_k are modeling decisions, we adopt the approach that is numerically and empirically more stable.

C Interpretation of β_k

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

$$\beta_k \sim Gamma(\frac{1}{K}, \frac{tr(\sqrt{\Sigma_k})}{M}) \tag{13}$$

Then, Equation 8 is the point estimate of β_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 Σ_k is positive semi-definite, and square root is a monotonically increasing function, as M increases, the trace of Σ_k increases ($\sum_i^M \sigma_i, \sigma_i \geq 0, \forall i$) and β_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	Gironde
reinaissance	countries	June	Spain
growth	history	October	Picardie
list	published	April	Belgium

Table 5: 'Genuine' Topic model clusters learnt from the documents vs clusters with ≥ 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.

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 ϵ 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.

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	space	countries	cut	weeks
christianity	prices	statement	light	national	straight	driving
spirit	quality	values	current	canada	pts	stuff
mary	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

Adaptive LDA (ADAP)

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	player
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

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
files	article	pretty	hard	nice	human	public
dos	send	remember	monitor	note	simply	rights
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	modem	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	S_k:0.086	S_k:0.042	S_k:0.251	S_k:0.011	S_k:0.083	S_k:0.136
wc:4613	wc:3344	wc:2532	wc:4209	wc:2800	wc:49408	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	clipper
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
white	systems	muslim	lines	cancer	action	data
left	normal	children	faq	cheers	society	search
heard	ground	genocide	based	double	subject	des
fun	control	human	version	original	land	law
looked	boot	shuttle	product	april	americans	nsa
air	fit	real	subject	effects	weapons	secure
tires	signal	cut	write	studies	questions	david
road	heat	turkey	note	usa	court	message
weight	fine	western	response	patients	congress	warrant
S_k:0.062	S_k:-0.008	S_k:0.106	S_k:0.107	S_k:-0.202	S_k:0.037	S_k:0.246
wc:6189	wc:6973	wc:6362	wc:6503	wc:5782	wc:7046	wc:7628

Discrete LDA (DISC)

Topic:7	**Topic:8**	**Topic:9**	**Topic:10**	**Topic:11**	**Topic:12**	**Topic:13**
bike	jesus	local	space	games	window	file
bad	life	told	university	team	image	files
stuff	church	friend	company	players	code	dos
real	faith	book	april	play	screen	software
lot	christian	weeks	technology	win	application	graphics
level	bible	dealer	press	season	color	version
times	christians	hand	conference	hockey	display	format
ride	love	check	science	league	server	advance
guess	christ	references	launch	player	set	info
sort	sin	james	news	baseball	error	package
thinking	human	heard	earth	series	size	directory
deleted	gods	talk	june	teams	running	disk
couple	agree	cover	radio	runs	images	unix
short	wrong	trouble	internet	fans	include	ftp
dod	truth	remember	greek	fan	change	hard
field	true	picture	station	pick	user	type
left	moral	experience	center	nhl	create	programs
hit	belief	set	contact	goal	widget	convert
job	person	bought	office	boston	manager	bit
canada	word	yeah	force	mike	event	applications
S_k:0.06	S_k:0.251	S_k:-0.119	S_k:-0.038	S_k:0.262	S_k:0.187	S_k:0.202
wc:7097	wc:10345	wc:5690	wc:6855	wc:9185	wc:7249	wc: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
scsi	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
modem	including	english	statements	policy	wait
S_k:0.221	S_k:0.163	S_k:0.02	S_k:0.056	S_k:0.116	S_k:0.032
wc:10829	wc:7104	wc:6217	wc:7824	wc:7575	wc:6668

Static mix π, (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	mit
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 π, (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 π, (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
scsi	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
vga	called	guy	sunday	brian	major	variety
chips	real	sounds	june	joe	player	mouse
motherboard	wrong	history	march	doug	season	addition
bios	article	ulf	morning	bobby	hockey	feature
interface	reason	auto	saturday	scott	black	covers
cpu	files	guys	friday	andy	hit	sale
cdrom	version	total	thursday	frank	head	notes
computers	stuff	march	july	kevin	league	map
amiga	heard	arts	tuesday	keith	baseball	van
macintosh	support	condition	monday	jeff	pick	consists
processor	send	late	daily	ro	average	runs
S_k:0.093	S_k:0.043	S_k:-0.583	S_k:0.017	S_k:-0.277	S_k:0.155	S_k:-0.289
wc:3145	wc:64156	wc:432	wc:2285	wc:1562	wc:13073	wc:1200

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	country	files	list	science	hockey
include	stuff	usa	chip	local	study	manager
images	remember	canada	dos	federal	school	baseball
included	person	countries	graphics	national	society	nhl
picture	pretty	international	scsi	congress	department	pitcher
gif	nice	germany	hardware	party	scientific	office
count	guess	america	encryption	governments	student	cubs
poster	simply	united	server	population	engineering	stats
listed	guy	americans	tape	membership	studies	braves
refer	talking	europe	unix	foreign	master	rangers
jpeg	yeah	british	floppy	committee	degree	record
listing	suppose	english	chips	land	education	leafs
recognized	guys	middle	ibm	bds	medicine	sox
photo	stupid	modern	vga	authorities	college	flyers
pom	understand	national	interface	andor	institute	coach
apr	hey	european	motherboard	administration	teaching	bruins
van	msg	nation	bios	legislation	students	pitchers
map	forget	japanese	cpu	liberal	literature	batting
counts	imagine	australia	cache	senate	astronomy	caps
S_k:-0.412	S_k:0.043	S_k:0.117	S_k:-0.019	S_k:-0.213	S_k:-0.058	S_k:-0.204
wc:1704	wc:3985	wc:1906	wc:4382	wc:1867	wc:1955	wc:2029

Gaussian LDA, (GAUS)

Topic:7	**Topic:8**	**Topic:9**	**Topic:10**	**Topic:11**	**Topic:12**
day	car	war	jesus	israel	mike
days	card	military	religion	jews	david
april	window	citizens	church	jewish	steve
months	price	doctor	faith	israeli	mark
night	monitor	civil	christian	armenians	dave
weeks	speed	army	bible	arab	jim
week	disk	soldiers	christians	turkish	bob
month	sale	population	christ	armenian	chris
hours	machine	century	religious	greek	tom
sunday	mac	persons	gods	muslims	brian
march	code	surrender	christianity	muslim	tim
june	color	forces	atheists	islamic	adam
apr	systems	fighting	beliefs	genocide	rob
morning	screen	civilians	holy	turkey	larry
saturday	standard	women	scripture	arabs	joe
friday	display	troops	atheist	russian	ron
thursday	video	local	catholic	turks	frank
july	bike	victims	atheism	palestinian	doug
tuesday	sell	deaths	religions	greece	bobby
spring	box	crimes	islam	azerbaijan	jack
S_k:-0.005	S_k:0.068	S_k:-0.13	S_k:0.259	S_k:0.085	S_k:-0.158
wc:2507	wc:37857	wc:1808	wc:3049	wc:2601	wc:2230

Gaussian LDA, (GAUS)

Topic:13	**Topic:14**	**Topic:15**	**Topic:16**	**Topic:17**	**Topic:18**	**Topic:19**
pittsburgh	power	team	love	drug	john	road
boston	set	win	death	food	paul	city
san	data	season	children	pain	michael	local
york	bit	league	friend	drugs	james	western
washington	post	teams	house	disease	mary	west
texas	call	title	family	cancer	andrew	land
clinton	lot	final	friends	brain	smith	east
chicago	hard	cup	lord	health	george	south
detroit	called	playoffs	child	blood	morris	north
waco	bad	playoff	wife	treatment	peter	town
california	key	winner	son	patients	thomas	location
toronto	article	beat	died	risk	gordon	houses
angeles	reason	score	father	medical	norton	region
colorado	space	record	born	ice	stanley	roads
los	support	match	meg	heart	joseph	village
buffalo	send	scored	mother	diet	johnson	central
kent	version	winning	marriage	eat	grant	route
philadelphia	heard	pts	woman	alcohol	allen	border
ottawa	book	scoring	parents	patient	grace	eastern
francisco	life	division	named	chronic	adams	cities
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
wc:1793	wc:68542	wc:1992	wc:2166	wc:1848	wc:1702	wc:1584