Computational Statistics (2020) 35:175–201
https://doi.org/10.1007/s00180-019-00891-1

ORIGINAL PAPER

DOLDA: a regularized supervised topic model for high-dimensional multi-class regression

Måns Magnusson[1,2] · Leif Jonsson[3] · Mattias Villani[1,4]

Received: 13 June 2017 / Accepted: 10 April 2019 / Published online: 12 June 2019

Abstract
Generating user interpretable multi-class predictions in data-rich environments with many classes and explanatory covariates is a daunting task. We introduce Diagonal Orthant Latent Dirichlet Allocation (DOLDA), a supervised topic model for multi-class classification that can handle many classes as well as many covariates. To handle many classes we use the recently proposed Diagonal Orthant probit model (Johndrow et al., in: Proceedings of the sixteenth international conference on artificial intelligence and statistics, 2013) together with an efficient Horseshoe prior for variable selection/shrinkage (Carvalho et al. in Biometrika 97:465–480, 2010). We propose a computationally efficient parallel Gibbs sampler for the new model. An important advantage of DOLDA is that learned topics are directly connected to individual classes without the need for a reference class. We evaluate the model's predictive accuracy and scalability, and demonstrate DOLDA's advantage in interpreting the generated predictions.

Keywords Text classification · Latent Dirichlet Allocation · Horseshoe prior · Diagonal Orthant probit model · Interpretable models

✉ Måns Magnusson
mans.magnusson@liu.se

Leif Jonsson
leif.jonsson@ericsson.com

Mattias Villani
mattias.villani@liu.se

[1] Department of Computer and Information Science, Linköping University, 581 83 Linköping, Sweden

[2] Department of Computer Science, Aalto University, Konemiehentie 2, 02150 Espoo, Finland

[3] Department of Computer and Information Science, Ericsson AB, 164 80 Stockholm, Sweden

[4] Department of Statistics, Stockholm University, 114 19 Stockholm, Sweden

1 Introduction

During recent decades more and more textual data has become available, creating a growing need to statistically analyze large amounts of textual data. The popular Latent Dirichlet Allocation (LDA) model introduced by Blei et al. (2003) is a generative probabilistic model in which each document is summarized by a set of latent semantic themes, often called *topics*. Formally, a topic is a probability distribution over the vocabulary. An estimated LDA model is, therefore, a compressed latent representation of the documents where each document is a mixture of topics and where each word (token) in a document belongs to a single topic. Most probabilistic topic models, such as LDA, are unsupervised, i.e. the topics are learned solely from the words in the documents without access to other document meta-data.

In many situations, though, there is other information we would like to incorporate in modeling a corpus of documents. A common example is when we have labeled documents, such as ratings of movies together with a movie description, illness categories in medical journals, or the locations of identified bugs together with bug reports in software engineering applications. The simplest approach would be to use a standard topic model and then use the estimated topic distributions per document in another model, such as a logistic regression model. This two-step approach would result in topics that are not produced for the purpose of explaining the dependent variable of interest. Alternatively, one could use a supervised topic model to find the semantic topic structure in the documents that are related to the class of interest. The difference between a supervised topic model and a two-step approach is similar to the difference between principal component regression (PCR) and partial least squares (PLS). In PCR, the principal components are first computed and then a regression model is estimated based on the estimated components. In PLS the components are estimated together with the regression model with the purpose of estimating components that have good predictive performance for the dependent variable of interest (Geladi and Kowalski 1986).

Most of the proposed supervised topic models have been designed with the objective of by trying to find good text classification models, and the focus has naturally been on the predictive performance. However, the predictive performance of most supervised topic models is similar to that of using a Support Vector Machine (SVM) with covariates based on word frequencies (Jameel et al. 2015). While predictive performance is certainly important, the real attraction of supervised topic models comes from their ability to learn semantically relevant topics and to use those topics to produce accurate *interpretable* predictions of documents or other textual data. The interpretability of a model is an often-neglected feature, but it is crucial in real-world applications. As an example, Parnin and Orso (2011) show that bug fault localization systems are quickly disregarded when the users cannot understand how the system has reached its predictive conclusion. Compared to other text classification systems, topic models are very well suited for interpretable predictions since topics are abstract entities that humans can easily grasp. The problems of interpretability in multi-class supervised topic models can be divided into three main areas.

First, most supervised topic models use a logit or probit approach, where the model is identified by the use of a *reference category* to which the effect of any covariate is compared. This defeats one of the main purposes of supervised topic models since it complicates the interpretability of the models. Instead of interpreting the effect of a topic on a class, we need to interpret it as the effect on a class compared to the reference category.

Second, to handle multi-class categorization *a topic should be able to affect multiple classes*, and some topics *may not influence any class at all*. In most supervised topic modeling approaches (such as Jiang et al. 2012; Zhu et al. 2013; Jameel et al. 2015) the multi-class problem is solved using binary classifiers in a "one-vs-all" classification approach. This approach works well in the situation of evenly distributed classes, but may not work well for skewed class distributions (Rubin et al. 2012). A one-vs-all approach also makes it more difficult to interpret the model. Estimating one model per class makes it impossible to see which classes are affected by the same topic and which topics do not predict any label. In these situations, we would like to have *one* topic model to interpret. The approach of one-vs-all is also costly from an estimation point of view since we need to estimate one model per class (Zheng et al. 2015), something that can be difficult in a situation with hundreds of classes.

Third, there can be situations with hundreds of classes and hundreds of topics [see Jonsson et al. (2016) for an example]. Without *regularization* or variable selection we would end up with a model with too many parameters to interpret and uncertain parameter estimates. In a good predictive supervised topic model, one would like to find a small set of topics that are strong determinants of a single document class label. This is especially relevant when the numbers of observations in different classes are skewed, which is a common problem in real-world situations (Rubin et al. 2012). In the more rare classes, we would like to induce more shrinkage compared to more common classes.

Multi-class regression is a non-trivial problem in Bayesian modeling. Historically, the multinomial probit model has been preferred due to the data augmentation approach proposed by Albert and Chib (1993). Augmenting the sampler using latent variables lead to straightforward Gibbs sampling with conditionally conjugate updates of the regression coefficients. The Albert-Chib sampler often tend to mix slowly, and the same holds for improved samplers such as the parameter expansion approach in Imai and van Dyk (2005). Recently, Polson et al. (2013) have proposed a similar data augmentation approach using Polya-gamma variables for the Bayesian logistic regression model. This approach preserves conditional conjugacy in the case of a Normal prior for the regression coefficients and was the foundation for the supervised topic model in Zhu et al. (2013).

In addition to the issue of interpretability of models, scalability of topic models is of crucial importance. MCMC algorithms are generally considered to be computationally costly. In the case of probabilistic topic models, this is even more prominent, since we often sample at least one parameter per word for the whole corpus. Modern corpora can be very large, with millions of documents, making efficient and parallel sampling a crucial component.

In this paper we explore a new approach to supervised topic models that produces accurate multi-class predictions from semantically interpretable topics using a fully Bayesian approach, hence solving all three of the above-mentioned problems. The model combines LDA with the recently proposed Diagonal Orthant (DO) probit model (Johndrow et al. 2013) for multi-class classification, with an efficient Horseshoe prior that achieves sparsity and interpretation by aggressive shrinkage (Carvalho et al. 2010). The new Diagonal Orthant Latent Dirichlet Allocation (DOLDA)[1] model has been demonstrated to have competitive predictive performance, while still producing interpretable multi-class predictions from semantically relevant topics. In addition, we also derive an efficient and parallel MCMC sampler that can be used to scale up model inference to larger corpora.

The paper is organized as follows. In Sect. 2, we describe the new proposed model and in Sect. 3 we then describe the proposed scalable MCMC sampler for the proposed model. In Sect. 4, the experimental results are presented and in Sect. 5 we conclude the paper. In the "Appendix", a full derivation of the sampler is supplied.

2 Related work

Incorporating supervised information or meta-data in the estimation of topic models has been done in a large number of papers, such as Rosen-Zvi et al. (2004), Griffiths et al. (2005) and Chemudugunta et al. (2007) that incorporate author information, syntax and background structures, to give a few early examples. How topic models incorporate the labeled information, such as classes, can broadly be classified into two groups, *downstream* supervised models and *upstream* supervised models. In upstream topic models, loosely defined, the topics are conditioned on the labeled information, while in downstream topic models, the labeled information is conditioned on the topics. Examples of upstream topic models are topics conditioned on authorship (Rosen-Zvi et al. 2004), topical perspectives (Ahmed and Xing 2010) and more general supervised information (Mimno and McCallum 2012).

In downstream topic models, of which our proposed model is an example, the label information is instead conditioned on the topics, similar to conditioning on covariates in a linear regression model or a logistic regression model. One of the first approaches was proposed by McAuliffe and Blei (2008) were the authors propose a supervised topic model based on the generalized linear model framework, thereby making it possible to link binary, count, and continuous response variables to topics that are inferred jointly with the regression/classification effects. This idea has been elaborated further, especially in the case of classification, in a series of paper, all closely connected to this work. The three most related approaches are Jiang et al. (2012), that propose a downstream supervised topic model using a max-margin classification, Zhu et al. (2013) that propose a logistic supervised topic model using data augmentation with Polya-gamma variates and Perotte et al. (2011) that use a hierarchical binary probit approach to model a hierarchical label structure in the form of a binary tree. All of these models are downstream supervised topic models using MCMC for inference and different forms of data augmentation approaches to model classes. Zhu et al. (2013) and Perotte et al. (2011) in combine a data augmentation strategy together with a linear model, just our proposed model. Compared to these earlier work we differ in two major ways. First, we use the diagonal orthant data augmentation scheme

[1] DOLDA is Swedish for hidden or latent.

that is computationally attractive for many classes, essentially addressing the issue of scalability in the number of classes. In addition, unlike the work of Zhu et al. (2013), Perotte et al. (2011) and Jiang et al. (2012), we focus on interpretability by using the horseshoe prior (Carvalho et al. 2010), something not done previously to ours. We also, just like (Zheng et al. 2015), focus on scalability in supervised topic models, but unlike (Zheng et al. 2015), we do not use Metropolis–Hastings to improve computational complexity. Instead, our approach focuses on the use of exchangeability to enable parallelism and analytically updates to make computations more efficient.

3 Diagonal Orthant Latent Dirichlet Allocation

3.1 Handling the challenges of high-dimensional interpretable supervised topic models

To solve the first and second challenges identified in the Introduction, i.e., reference classes and multi-class models, we propose the use of the Diagonal Orthant (DO) probit model in Johndrow et al. (2013) as an alternative to the multinomial probit and logit models. Johndrow et al. (2013) propose a Gibbs sampler for the Diagonal Orthant model and show that it mixes well. One of the benefits of the DO model is that all classes can be independently modeled using binary probit models when conditioning on the latent variable, thereby removing the need for a reference class. The parameters of the model can be interpreted as the effect of the covariate on the marginal probability of a specific class, which makes this model especially attractive when it comes to interpreting the inferred topics. This model also includes multiple classes in an efficient way that makes it possible to estimate a multi-class linear model in parallel over the classes.

The third problem of modeling supervised topic models is that the semantic meanings of all topics do not necessarily have an effect on our label of interest; one topic may have an effect on one or more classes, and some topics may just be noise that we do not want to use in the supervision. In cases where there are many topics and many classes, we will also have a very large number of parameters to analyze. The Horseshoe prior in Carvalho et al. (2010) was specifically designed to filter out signals from massive noise. This prior uses a local-global shrinkage approach to shrink some (or most) coefficients to zero while allowing for sparse signals to be estimated without any shrinkage. This approach has shown good performance in linear regression-type situations (Castillo et al. 2015), with many predictors (Nalenz and Villani 2018), which makes it straightforward to incorporate other covariates into our model, something that is rarely done in the area of supervised topic models. Different global shrinkage parameters are used for the different classes to handle the problem with an unbalanced number of observations in different classes. This makes it possible to shrink more when there is less data for a given class and shrink less in classes with more observations.

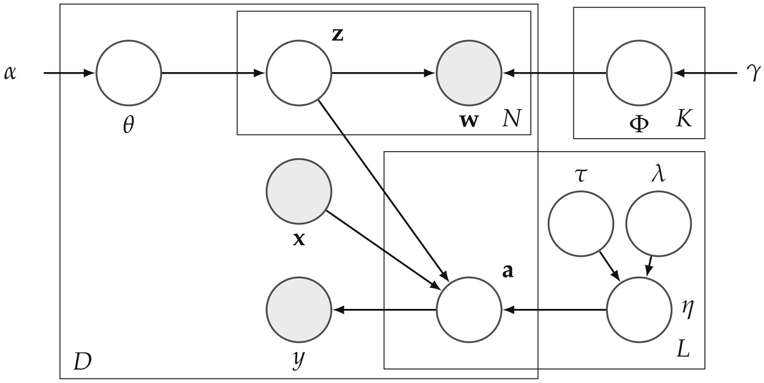

Fig. 1 The Diagonal Orthant probit supervised topic model (DOLDA)

3.2 Generative model

The DOLDA generative model is described below. See also a graphical description of the model in Fig. 1. A summary of the notation is given in Table 1.

(1) For each topic $k = 1, \ldots, K$

  (a) Draw a distribution over words $\phi_k \sim \mathrm{Dir}_V(\beta)$

(2) For each label $l \in L$

  (a) Draw a global shrinkage parameter $\tau_l \sim C^+(0, 1)$
  (b) For each covariate and topic $p = 1, \ldots, K + P$
    (i) Draw local shrinkage parameter $\lambda_{l,p} \sim C^+(0, 1)$
    (ii) Draw coefficients[2] $\eta_{l,p} \sim \mathcal{N}(0, \tau_l^2 \lambda_{l,p}^2)$

(3) For each observation/document $d = 1, \ldots, D$

  (a) Draw topic proportions $\theta_d \sim \mathrm{Dir}_K(\alpha)$
  (b) For each token $n = 1, \ldots, N_d$
    (i) Draw topic assignment $z_{n,d}|\theta_d \sim \mathrm{Categorical}(\theta_d)$
    (ii) Draw word $w_{n,d}|z_{n,d}, \phi_{z_{n,d}} \sim \mathrm{Categorical}(\phi_{z_{n,d}})$
  (c) $y_d \sim \mathrm{Categorical}(\mathbf{p}_d)$ where

$$\mathbf{p}_d = \left[\sum_{l}^{L} cdf_{\mathcal{N},l} \left((\bar{\mathbf{z}}, \mathbf{x})_d^\top \eta_{l\cdot} \right) \right]^{-1} \left(F_{\mathcal{N}}\left((\bar{\mathbf{z}}, \mathbf{x})_d^\top \eta_{1\cdot} \right), \ldots, F_{\mathcal{N}}\left((\bar{\mathbf{z}}, \mathbf{x})_d^\top \eta_{L\cdot} \right) \right)$$

and $F_{\mathcal{N}}(\cdot)$ is the univariate normal CDF (Johndrow et al. 2013).

[2] The intercept is assigned a normal prior.

Table 1 DOLDA model notation

Symbol	Description	Symbol	Description
$\mathcal{V}$	The set of word types/vocabulary	β	The prior for Φ: $K \times V$
V	The size of the vocabulary i.e $V = \|\mathcal{V}\|$	Θ	Document-topic proportions: $D \times K$
v	Word type	θ_d	Topic probability for document d
$\mathcal{K}$	The set of topics	α	The prior for Θ: $D \times K$
K	The number of topics i.e $K = \|\mathcal{K}\|$	$\mathbf{M}$	# of topic indicators in each document: $D \times K$
L	The number of labels/categories	$\mathbf{a}$	Matrix of latent gaussian variables: $D \times L$
$\mathcal{L}$	The set of labels/categories	η	Coefficient matrix for each label and covariate: $(K + P) \times L$
D	The number of observations/documents i.e. $D = \|\mathcal{D}\|$	η_0	Prior for η: $L \times (K + P)$
$\mathcal{D}$	The set of observations/documents	$z_{n,d}$	Topic indicator for token n in document d
P	The number of non-topic covariates/features	$\bar{\mathbf{Z}}$	Proportion of topic indicators by document: $D \times K$
$F_{\mathcal{N}}(\cdot)$	The univariate standard Normal cdf	$\bar{\mathbf{z}}_d$	Proportion of topic indicators for document d
N	The total number of tokens	$w_{n,d}$	Token n in document d
N_d	The number of tokens in document d	$\mathbf{w}_d$	Vector of tokens in document d: $1 \times N_d$
$\mathbf{N}$	# obs topic-word type indicators: $K \times V$	y_d	Label for document d
Φ	The matrix with word-topic probabilities : $K \times V$	$\mathbf{X}$	Covariate/feature matrix (including intercept): $D \times P$
ϕ_k	The word probabilities for topic k: $1 \times V$	$\mathbf{x}_d$	Covariate/features for document d

4 Inference

4.1 The MCMC algorithm

Markov Chain Monte Carlo (MCMC) is used to estimate the model parameters. We use different global shrinkage parameters τ_l for each class, based on the fact that the different classes can have a different number of observations. This gives the following sampler for inference in DOLDA, see "Appendix A" for details.

(1) Sample the latent variables $a_{d,l} \sim \mathcal{N}_{+}((\mathbf{x}\,\bar{\mathbf{z}})_d^T \eta_l, 1)$ for $l = y_d$ and $a_{d,l} \sim \mathcal{N}_{-}((\mathbf{x}\,\bar{\mathbf{z}})_d^T \eta_l, 1)$ for $l \neq y_d$, where $\mathcal{N}_{+}$ and $\mathcal{N}_{-}$ are the positive and negative truncated normal distribution, truncated at 0.

(2) Sample all of the regression coefficients as in an ordinary Bayesian linear regression per class label l where $\eta_l \sim \mathcal{MVN}\left(\mu_l, ((\mathbf{X}\,\bar{\mathbf{Z}})^T(\mathbf{X}\,\bar{\mathbf{Z}}) + \tau_l^2\Lambda_l)^{-1}\right)$ and Λ_l is a diagonal matrix with the local shrinkage parameters λ_i per parameter in η_l and $\mu_l = ((\mathbf{X}\,\bar{\mathbf{z}})^T(\mathbf{X}\,\bar{\mathbf{z}}) + \tau_l^2\Lambda_l)^{-1}(\mathbf{X}\,\bar{\mathbf{z}})^T\mathbf{a}_l$

(3) Sample the global shrinkage parameters τ_l at iteration j using the following two step slice sampling:

$$u \sim \mathcal{U}\left(0, \left[1 + \frac{1}{\tau_{l,(j-1)}}\right]^{-1}\right)$$

$$\frac{1}{\tau_{l,j}^2} \sim \mathcal{G}\left((p+1)/2, \frac{1}{2}\sum_{p=1}^{K+P}\left(\frac{\eta_{l,p}}{\lambda_{l,p}}\right)^2\right) I\left[\frac{1}{\tau_{l,(j-1)}^2} < (1-u)/u\right]$$

where I indicates the truncation region for the truncated gamma.

(4) Sample each local shrinkage parameter $\lambda_{i,l}$ at iteration j as

$$u \sim \mathcal{U}\left(0, \left[1 + \frac{1}{\lambda_{p,l,(j-1)}^2}\right]^{-1}\right)$$

$$\frac{1}{\lambda_{p,l,j}^2} \sim \mathrm{Exp}\left(\frac{1}{2}\left(\frac{\eta_{l,p}}{\tau_l}\right)^2\right) I\left[\frac{1}{\lambda_{p,l,(j-1)}^2} < (1-u)/u\right]$$

where I indicates the truncation region for the truncated exponential distribution.

(5) Sample the topic indicators $\mathbf{z}$

$$p(z_{i,d} = k | w_i, \mathbf{z}^{\neg i}, \eta, \mathbf{a}) \propto \phi_{v,k} \cdot \left(\mathbf{M}_{d,k}^{\neg i} + \alpha\right)$$
$$\times \exp\left(-\frac{1}{2}\sum_{l}^{L}\left[-2\frac{\eta_{l,k}}{N_d}\left(a_{d,l} - (\bar{\mathbf{z}}_d^{\neg i}\,\mathbf{x}_d)\eta_l^{\mathsf{T}}\right) + \left(\frac{\eta_{l,k}}{N_d}\right)^2\right]\right)$$

where $\mathbf{M}$ is a $D \times K$ count matrix containing the sufficient statistics for Θ and $\mathbf{M}^{\neg i}$ is this matrix with topic indicator $z_{i,d}$ removed from $\mathbf{M}$.

(6) Sample the topic-vocabulary distributions Φ

$$\phi_k \sim \mathrm{Dir}(\beta + \mathbf{N}_{k,\cdot})$$

where $\mathbf{N}$ is a $K \times V$ count matrix containing the sufficient statistics for Φ.

4.2 Efficient parallel sampling of z

To improve the speed of the sampler we cache the calculations done in the supervised part of the topic indicator sampler and parallelize the sampler. Very large text corpora are increasingly common, so efficient sampling of the $\mathbf{z}$ is absolutely crucial in practice. The basic sampler for $\mathbf{z}$ can be slow due to the serial nature of the collapsed sampler and the fact that the supervised part of $p(z_{i,d})$ involves a dot product. A naive implementation would result in a complexity of $O((K+P)\cdot L \cdot K)$ to sample just one topic indicator z_i.

The supervised part of document d can be expressed as $g_{d,k}^{\neg i}$ where

$$g_{d,k}^{\neg i} = \exp -\frac{1}{2}\sum_{l}^{L}\left[-2\frac{\eta_{l,k}}{N_d}\left(a_{d,l} - (\bar{\mathbf{z}}_d^{\neg i}\,\mathbf{x}_d)\eta_l^{\intercal}\right) + \left(\frac{\eta_{l,k}}{N_d}\right)^2\right].$$

By realizing that sampling a topic indicator $z_{i,d}$ will only change this part a little, we can derive the relationship

$$g_{d,k}^{\neg i} = g_{d,k}^{\neg(i-1)} + \frac{1}{N_d^2}\left[\sum_{l}^{L}\eta_{l,k}\eta_{l,z_{i,d}} - \sum_{l}^{L}\eta_{l,k}\eta_{l,z_{i-1,d}}\right],$$

where $g_{d,k}^{\neg(i-1)}$ is the supervised effect computed for the previous token and where the expression $\sum_l^L \eta_{l,k}\eta_{l,z_{i,d}}$ can be calculated once per iteration in η and can be stored in a two-dimensional array of size K^2. We can use the above relationship to update the supervision after sampling each topic indicator by calculating $g_{d,k}^{\neg i}$ "on the fly" based on the previous supervised contribution. This means that we only need to compute $g_{d,k}^{\neg i}$ once per document, and then we just need to update these values. This approach to computing $g_{d,k}^{\neg i}$ increases the speed by an order of magnitude for a model with 100 topics and reduces the computational complexity of sampling one $z_{i,d}$ from $O((K+P)\cdot L\cdot K)$ to $O(K)$ for all but the first token per document. For details, see "Appendix B".

To further improve the performance we parallelize the sampler and use the fact that documents are conditionally independent given Φ. By sampling Φ instead of marginalizing it out we will gain from parallelization with the additional cost of sampling Φ. This approach to parallelizing topic models give us a sampler that correctly samples the posterior using an ergodic Markov chain, unlike other parallel approaches such as AD-LDA (Magnusson et al. 2018; Newman et al. 2009).

In summary, we propose a sampler that samples the $z_{i,d}$ in parallel over the documents, the elements in Φ sampled in parallel over topics, and sampling η can be in parallel over classes. The code is publicly available at https://github.com/lejon/DiagonalOrthantLDA.

4.3 Computational complexity

The computational complexity of the sampler depends on the different parameters sampled. Below we analyze the different parts of the sampler for one iteration. Sampling the regression coefficients has three parts, (1) computing $\Lambda_{post} = ((\mathbf{X}\,\bar{\mathbf{z}})^T(\mathbf{X}\,\bar{\mathbf{z}}) + \tau_l^2\Lambda_l)$ is of complexity $O(L \cdot D \cdot (K + P)^2)$, (2) inverting Λ_{post} for all classes is of complexity $O(L \cdot (K + P)^3)$, and (3) sampling from the multivariate Gaussian distribution of each η_l is also of complexity $O(L \cdot (K + P)^3)$. Sampling the topic indicators has complexity $O((K + P) \cdot L \cdot K \cdot D)$ for the first topic indicator in each document, a complexity dominated by $O(L \cdot D \cdot (K + P)^2)$. If we use the method for increased efficiency proposed above, all other topic indicators can be sampled with complexity $O(K \cdot N)$. Sampling the latent variables $\mathbf{a}$ is of complexity $(K + P) \cdot D \cdot L$ and is hence dominated by the sampling of η. Similarly, sampling τ and λ is of complexity $(K + P) \cdot L$ and is also dominated by the sampling of η. Sampling Φ is of complexity $O(K \cdot V)$, something that is generally dominated by sampling the topic indicators $O(K \cdot N)$, since topically $N >> V$ (Magnusson et al. 2018).

Finally, the total complexity of the sampler, with regard to the number of classes (L), the number of topics (K), the number of documents (D), the mean document size $(\bar{N})$, and the number of covariates (P) is $O(L \cdot (K + P)^3 + L \cdot D \cdot (K + P)^2 + K \cdot D \cdot \bar{N})$ where $\bar{N} = N/D$. From this analysis, we can see that as the corpus grows $(D \to \infty)$ we see that sampling η and the topic indicators $\mathbf{z}$ will dominate the computations. But we would also expect the number of topics to grow as the number of documents grows. In this situation, the main cost of the algorithm would be sampling the ηs and the first topic indicator of each document.

Due to the similarity of the DOLDA sampler to that of the MedLDA sampler, it is straightforward to use the cyclical Metropolis–Hastings proposals in Zheng et al. (2015) for inference in DOLDA. But, as shown in Magnusson et al. (2018), it is not obvious that the reduction in sampling complexity will result in a faster sampler when MCMC efficiency is taken into account.

4.4 Evaluation of convergence and prediction

We evaluate the convergence of the MCMC algorithm by monitoring the unnormalized log-likelihood over the iterations:

$$\begin{aligned}&\log \mathcal{L}(\mathbf{w}, \mathbf{y}|\mathbf{z}, \eta, \mathbf{X}, \alpha, \beta)\\&\quad = \log p(\mathbf{y}|\mathbf{z}, \eta, \mathbf{X}) + \log p(\mathbf{w}|\mathbf{z}, \alpha, \beta)\end{aligned}$$

$$
\begin{aligned}
&\propto \sum_{d}^{D} \log \left[(1 - F_{\mathcal{N}}(-(\bar{\mathbf{z}}_d \, \mathbf{x}_d)\eta_j^{\mathsf{T}})) \prod_{l \neq j} F_{\mathcal{N}}(-(\bar{\mathbf{z}}_d \, \mathbf{x}_d)\eta_l^{\mathsf{T}}) \right] \\
&\quad - \sum_{d}^{D} \log \left[\sum_{j=1}^{J} (1 - F_{\mathcal{N}}(-(\bar{\mathbf{z}}_d \, \mathbf{x}_d)\eta_j^{\mathsf{T}})) \prod_{l \neq s} F_{\mathcal{N}}(-(\bar{\mathbf{z}}_d \, \mathbf{x}_d)\eta_l^{\mathsf{T}}) \right] \\
&\quad + K \log \Gamma \left(\sum^{V} \beta \right) - KV \log \Gamma (\beta) + \sum^{K} \sum^{V} \log \Gamma \left(\mathbf{N}_{k,v} + \beta \right) \\
&\quad - \sum^{K} \log \Gamma \left(\sum^{V} \mathbf{N}_{k,v} + \beta \right) \\
&\quad + D \log \Gamma \left(\sum^{K} \alpha \right) - DK \log \Gamma (\alpha) + \sum^{D} \sum^{K} \log \Gamma \left(\mathbf{M}_{d,k} + \alpha \right) \\
&\quad - \sum^{D} \log \Gamma \left(\sum^{K} \mathbf{M}_{d,k} + \alpha \right),
\end{aligned}
$$

where $F_{\mathcal{N}}$ is the univariate normal distribution and the last part is the same computations commonly used in evaluating the standard LDA model.

To make predictions for a new document $d^{\star}$ we first need to sample the topic indicators of the given document from

$$p(z_{i,d^{\star}} = k | \mathbf{w}^{\star}, \Phi) \propto \bar{\phi}_{k,v} \cdot \left(\mathbf{M}_{d,k}^{\neg i} + \alpha \right),$$

where $\bar{\phi}_{k,v}$ is the mean of the last part of the posterior draws of Φ. We use the posterior mean based on the last iterations instead of integrating out Φ to avoid potential problems with label switching. However, we have not seen any indications of label switching after convergence in our experiment, probably because the data sets used for document predictions are usually quite large. The topic indicators are sampled for the predicted document using the fast PC-LDA sampler in Magnusson et al. (2018). The mean of the sampled topic indicator vector for the predicted document, $\bar{\mathbf{z}}^{\star}$, is then used for class predictions:

$$y^{\star} = \arg\max \left((\bar{\mathbf{z}}^{\star}, \mathbf{x}^{\star})^{\top} \eta \right).$$

This is a maximum a posteriori estimate, but it is straightforward to calculate the whole predictive distribution for the label.

5 Experiments

We study model performance in four different ways: the classification accuracy, the interpretability of the model, the topic quality and supervision effects on topics, and the scalability of the sampler. The experiments are performed on 2 sockets with 8-core Intel Xeon E5-2660 Sandy Bridge processors at 2.2GHz at the National Supercomputer Center (NSC) at Linköping University.

5.1 Corpora and priors

To study the different aspects of the DOLDA model we use multiple corpora. We collected a corpus containing the 10,810 highest-rated movies at IMDb.com. We use both the textual description and information about producers and directors to classify a given movie into a genre. We also analyze the classical 20 Newsgroup corpus.

In addition, we also include two corpora based on the New York Times Annotated Corpus (Sandhaus 2008) for our experiments. To label the documents we use the classification in the "online section". Thus, we only use articles from 2001 and later, when the "online section" was added. From these documents, we extract labels that we call "Top level" and "Hierarchical". These labels are used as the class of the documents. An example of an "online section" is

Arts; Dining and Wine; Education; Books

A document with the above example online-section would get the top label "Arts" and the hierarchical (2 level) label "Arts; Dining and Wine". For the hierarchical classification, we extracted only the articles which had at least two levels ("Arts" being the first level and "Dining and Wine" the second level in the example above). After extracting the classes for the documents we create four subsets from the corpora described above. These subsets contain 90%, 60%, 20%, and 10% of documents sampled from the above corpus. None of the documents in the 10% subset exists in the other subsets. The purpose of the NYT corpora is to show how the sampler scales, both with regard to documents ($\sim$600 K) and with a large number of classes (240).

Our companion paper (Jonsson et al. 2016) applies the DOLDA model developed here to bug localization in a large-scale software engineering context using a corpus with 15,000 bug reports, each belonging to one of 118 classes.

We include the corpora for different purposes. IMDb is a smaller corpus, but contains additional covariates. The 20 Newsgroups corpus is included to enable accuracy performance with other comparable supervised topic models. The New York Times corpus is included to show the scalability of the MCMC sampler with regard to the number of classes as well as the number of documents.

The corpora are tokenized and a standard stop list of English words are removed, as well as the rarest word types that make up 1% of the total tokens, or in some experiments, the words that occur less than 10 times. In the IMDb corpus, we only include genres with at least 10 movies (Table 2).

In all experiments, we use a relatively vague prior setting $\alpha = \beta = 0.01$ for the LDA part of the model and $c = 100$ for the prior variance of the η coefficients in the normal model prior and for the intercept coefficient when using the Horseshoe prior. The accuracy experiment for IMDb uses 5-fold cross-validation and the 20 Newsgroups corpus uses the same training and test set as in Zhu et al. (2012) to enable

Table 2 Corpora used in experiment, by the number of classes (L), the number of documents (D), the vocabulary size (V), and the total number of tokens (N)

Corpus	L	D	V	N
IMDb	20	10,810	47,371	967,255
20 Newsgroups	20	18,846	187,321	4,913,292
New York Times (hiearchical)	240	183,751	1,540,464	129,151,602
New York Times (top level)	31	595,635	3,326,778	339,298,734

Statistics have been computed using the word tokenizer in the `tokenizers` R package with default settings (Mullen 2016)

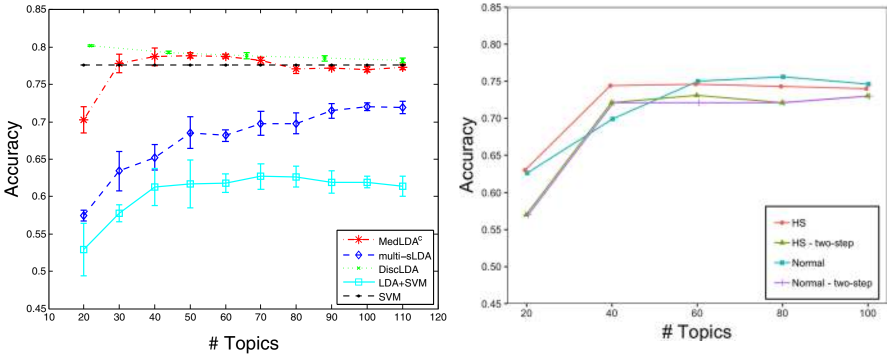

Fig. 2 Accuracy of MedLDA, taken from Zhu et al. (2012) (left) and accuracy of DOLDA for the 20 Newsgroup test set (right)

direct comparisons of accuracy. In the interpretability analysis of the IMDb corpus we use the whole corpus, without cross-validation.

5.2 Results

5.2.1 Classification accuracy

Figure 2 shows the accuracy on the hold-out test set for the 20 Newsgroups corpus for different numbers of topics. The accuracy of our model is slightly lower than MedLDA and SVM using only textual features, but higher than both the classical supervised multi-class LDA and the ordinary LDA together with an SVM approach.

We can also see from Fig. 2 that the accuracy of using the DOLDA model with the topics jointly estimated with the supervision part outperforms a two-step approach of first estimating LDA and then using the DO probit model with the pre-estimated mean topic indicators as covariates. This is true for both the Horseshoe prior and the normal prior, but the difference with regard to accuracy is just a few percentage points.

The advantage of DOLDA is that it produces interpretable predictions with semantically relevant topics. Therefore, it is reassuring that DOLDA can compete in accuracy with other less interpretable models such as the SVM, even when the model is dramatically simplified by aggressive Horseshoe shrinkage for interpretational purposes. Our

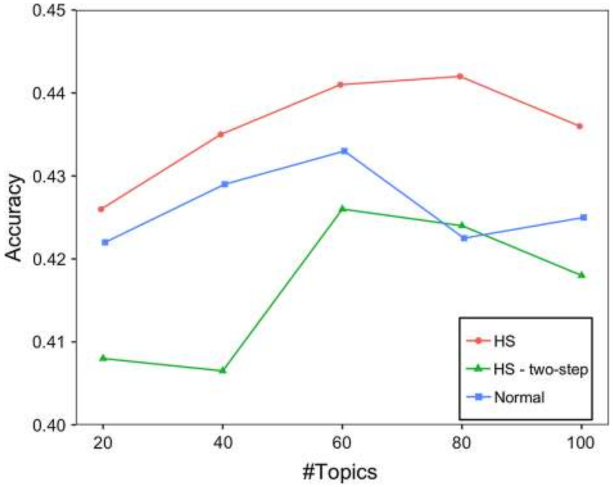

Fig. 3 Accuracy for DOLDA on the IMDb data with normal and Horseshoe prior and using a two step approach with the Horseshoe prior

next analysis illustrates the interpretational strength of DOLDA. See also our companion paper (Jonsson et al. 2016) in the software engineering literature for further demonstrations of DOLDA's ability to produce interpretable predictions in industrial applications.

Figure 3 displays the accuracy on the IMDb corpus as a function of the number of topics. The estimated DOLDA model also contains several other discrete covariates, such as the film's director and producer. The accuracy of the more aggressive Horseshoe prior is better than the normal prior for all topic sizes. A supervised approach with topics and supervision inferred jointly again outperforms a two-step approach.

5.2.2 Model interpretability

To illustrate the interpretability of DOLDA, we fit a new model to the IMDb corpus using only topics as covariates. Note first in Fig. 4 how the Horseshoe prior is able to distinguish between so-called signal topics and noise topics; the Horseshoe prior aggressively shrinks a large fraction of the regression coefficient toward zero, making it much easier to interpret how different latent aspects of the documents affect the class label. This is achieved without the need of setting any additional hyper-parameters in the model.

The Horseshoe shrinkage makes it easy to identify the topics that affect a given class. This is illustrated for the *Romance* genre in the IMDb corpus in Fig. 5. This genre consists of relatively few observations (only 39 movies), and the Horseshoe prior, therefore, shrinks most coefficients to zero, keeping only one large signal topic that happens to have a negative effect on the Romance genre. The normal prior, on the other, hand gives a much denser, and therefore a much less interpretable solution.

For a further analysis of what triggers a Romance genre label, Table 3 shows the 10 top words for Topic 39. From this table, it is clear that the signal topic identified

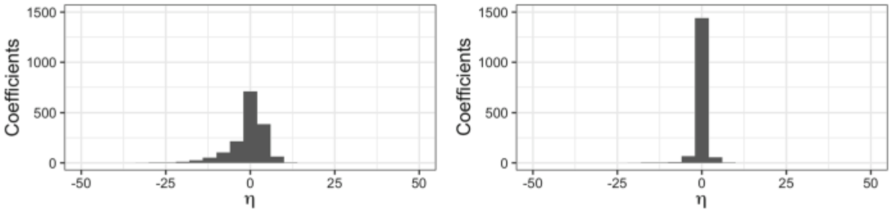

Fig. 4 Coefficients for the IMDb corpus with 80 topics using the normal prior (left) and the Horseshoe prior (right)

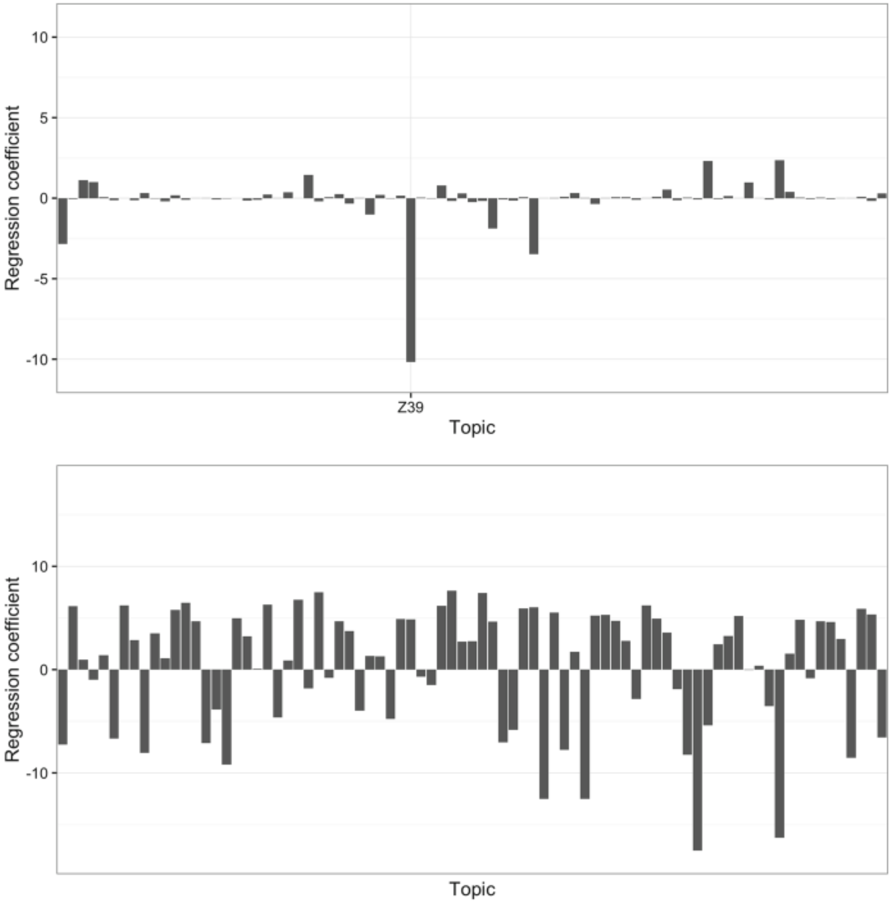

Fig. 5 Coefficients for the genre *Romance* in the IMDb corpus with 80 topics using the Horseshoe prior (upper) and a normal prior (below)

using the Horseshoe prior is some sort of "crime" topic that is negatively related to the Romance genre, which makes intuitive sense. The crime topic is clearly expected to be positively related to the *Crime* genre, and Fig. 6 shows that this is indeed the case.

Table 3 Top words in topics using the Horseshoe prior

Topic 33	Earth space planet alien human future years world time mission
Topic 39	Police murder detective killer case investigation crime crimes solve murdered

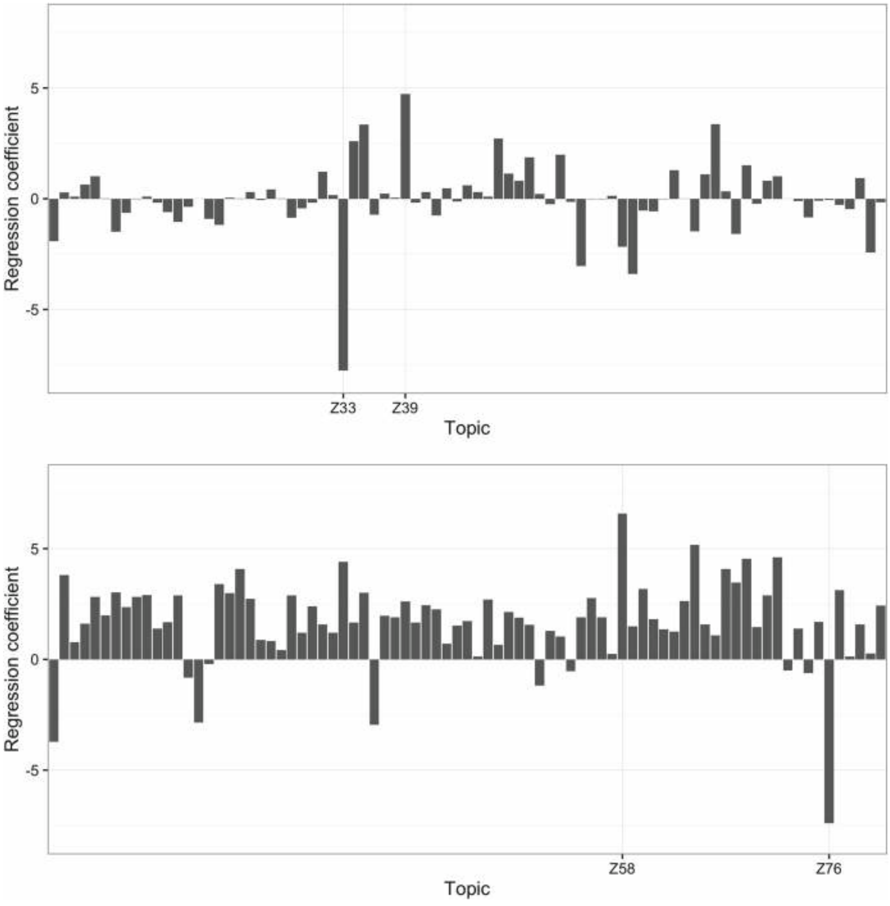

Fig. 6 Regression coefficients for the class *Crime* for the IMDb corpus with 80 topics using the Horseshoe prior (upper) and a normal prior (below)

We can also see from Fig. 6 that Topic 33 has a strong negative effect on the Crime genre. In Table 3 we can see that Topic 33 seems to be some sort of Sci-Fi topic. This topic has, in turn, the largest positive relationship with the Sci-Fi movie genre.

This illustrates an example how the aggressive shrinkage of the Horseshoe prior not only increases the prediction accuracy, but also simplifies interpretations since a much smaller number of topics is estimated to affect a given label - making it easier to focus on the topics that actually have an effect in the analysis. This is much more difficult in the Normal prior situation.

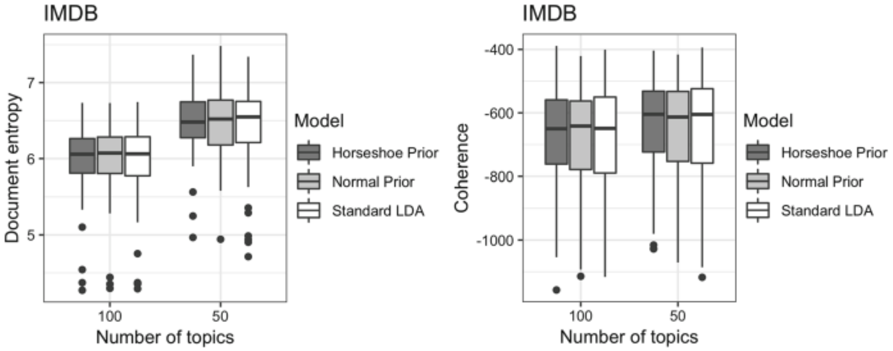

Fig. 7 Document entropy (left) and topic coherence (right) for the IMDb corpus

5.2.3 Topic quality and effect of supervision

Even though the accuracy improves using a supervised approach, this raises the question of the effect of the supervision on the quality of the individual topics. How are the topics affected by the supervision and by shrinkage priors?

To study the effect on the topics, we focus on two measurements of topic quality. First, we study the effect of *topic coherence* using the measure proposed by Mimno et al. (2011). This measure has been shown in experiments to be a good approximation of topical coherence, estimated using manual annotations, such as topic intrusion (Chang et al. 2009).

We also study the *document entropy* of the topics. This measure gives us an indication of how the topics are distributed over documents. Are topics evenly distributed over documents (high entropy) or more sparsely distributed over documents (low entropy). This is an indication of the effect that supervision has on the topics. Is the supervised information making the distribution over documents more or less sparse?

In Fig. 7 we can see that, in general, there is no large difference in coherence or document entropy between the different models and priors, which is also true for the other corpora (not shown). This indicates that the effect of the supervision on the inferred topics is small; document entropy and topic coherence remains more or less the same with and without supervision.

To study the effect of the supervision in more detail, we focus instead on those topics that are actually affected by the supervision in the model. Since we have L number of coefficients that affect each topic, we choose to study the supervised effect on topic k, called r_k, by looking at the sum of the absolute values of the η coefficients, i.e.

$$r_k = \sum_{l=1}^{L} |\eta_{l,k}|.$$

This is a rough estimate of the overall supervised effect on the individual topics. We also studied negative and positive regression coefficients separately, but the results

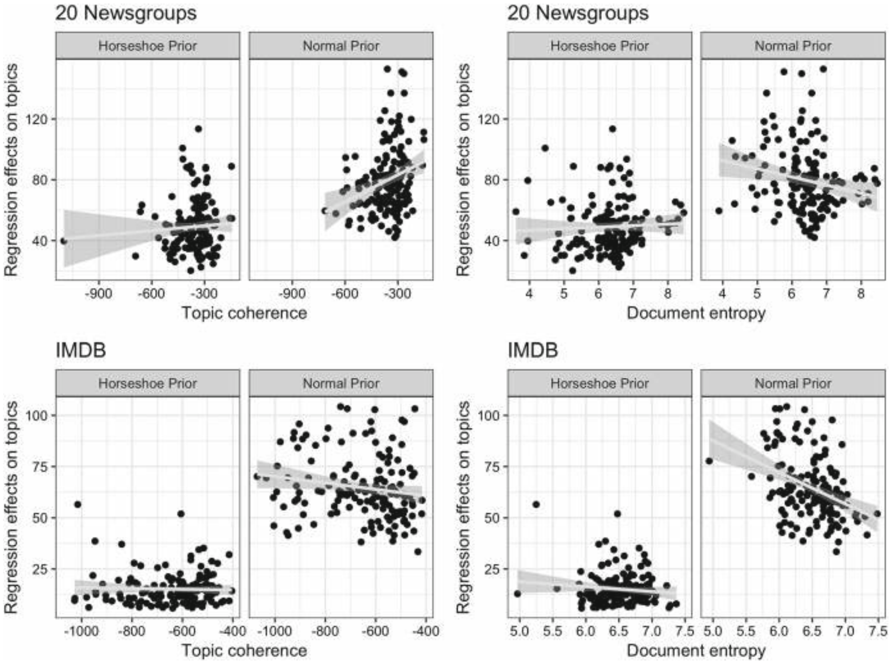

Fig. 8 Coherence and document entropy by supervised effect with 50 topics

are similar. In Fig. 8 the results are presented for different corpora, priors, and topic quality measures. In all cases, we use $K = 50$. The results are similar to those of $K = 100$.

Figure 8 show the effect of the supervision. We can see the effect of the horseshoe prior in that the regression supervised effects are lower in general, due to the shrinkage imposed by the prior. With regard to topic coherence and document entropy, the supervision has no clear effect. Instead, it seems like the effect that the supervision has on the topics is corpus-specific. In the 20 Newsgroups corpus, we can see a small positive relationship between coherence and supervision effect, something that is not shared by the other corpora.

These results indicate that the effect of the supervision on individual topics depends on the corpora (and the label). This makes sense in that different types of labels will relate to different aspects of the text and the underlying topics. For the 20 Newsgroups, we can see a slight positive correlation between coherence and the supervised effects, and this is also a corpus with higher prediction accuracy. But overall, the results seem to indicate that there is no clear effect of the supervision on topic quality, as measured by document entropy and coherence.

5.2.4 Scaling and parallelism performance

One of the contributions of this model is its ability to scale to larger corpora using a parallel and efficient MCMC sampler. To study the scalability of the model we ran

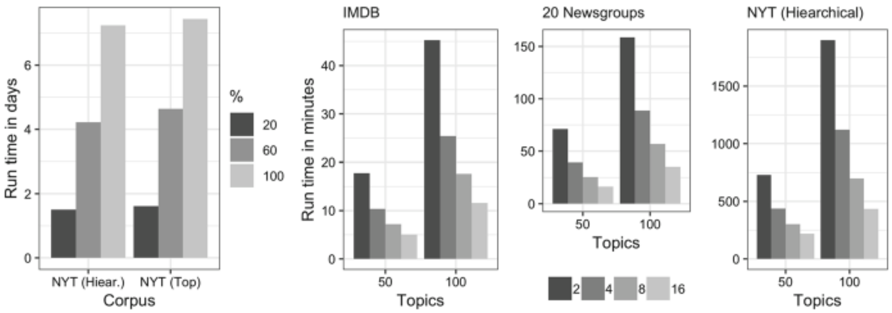

Fig. 9 Scaling performance (left) and parallel performance (right). The scaling experiments were run for 5000 iterations and the parallel performance experiments were run for 1000 iterations each. All were run with 3 different random seeds and the average runtime was computed. In the parallel experiment, the 20% NYT Hierarchical data was used and 2, 4, 8, and 16 cores

experiments on the runtime effects on the different aspects identified in the complexity analysis, the number of classes, the number of documents and the number of topics.

Figure 9 show the results of the scaling and parallel performance experiments. From the results, we can see that the scaling in size of the corpus is more or less linear, as we expect. More interestingly, the runtime of the two NYT corpora is very similar. The hierarchical NYT corpus has roughly ten times more classes (240 vs. 31) than the top-level NYT corpus while being roughly one third in size. Hence we can conclude that, empirically, the largest effect on runtime is the number of documents, or tokens, rather than the number of classes for a standard setting with 100 topics.

Figure 9 also shows the parallel performance of the sampler. Unlike most other large-scale MCMC samplers for topic models, this sampler is both parallel and samples using an ergodic Markov chain with the posterior as the target. This still gives good parallel performance, even on a smaller corpus, such as the IMDb corpus. It is also obvious that the concurrency with regard to the different classes is also of importance. For the hierarchical NYT corpus, the increased number of classes affects the overall sampling time, but the parallel performance is not affected. We can also see that the parallel performance is needed mainly when the number of topics is larger.

6 Conclusions

Several supervised topic models have been proposed with the purpose of identifying topics that can be used successfully to classify documents. We have proposed DOLDA, a supervised topic model with special emphasis on generating semantically interpretable predictions together with an efficient and scalable MCMC sampler for inference. An important component of the model to ease interpretation is the DO-probit model without a reference class. By coupling the DO-probit model with an aggressive Horseshoe prior with a shrinkage that is allowed to vary over the different classes, it is possible to create an interpretable classification model that automatically identifies the most interesting “signal” topics. At the same time, the DOLDA model comes with very few hyperparameters - only the standard LDA parameters α and β are needed, which has been extensively studied in Wallach et al. (2009). The fact that there are so few parameters is different from most other supervised topic models (Jiang et al. 2012; Zhu et al. 2012; Li et al. 2015).

Our experiments show that the gain in interpretation from using DOLDA comes with only a small reduction in prediction accuracy compared to the state-of-the-art supervised topic models; moreover, DOLDA outperforms other fully Bayesian models such as the original supervised LDA model. We have also shown that learning the topics jointly with the classification part of the model gives more accurate predictions than a two-step approach, where a topic model is first estimated and a classifier is then trained on the learned topics, showing a general benefit of supervised topic modeling.

The horseshoe prior has also shown benefits in supervised topic models, leading to a much more clear picture of the important topics for a given label with similar, or better, prediction accuracy. The computational cost of the horseshoe is small compared to the other parts of the sampler, making it an attractive prior for use in other supervised models as well.

The supervision effect on the topics is generally small and in line with previous results. The supervision, in general, does not seem to affect the topic interpretability much, but there seems to be an indication that this is corpus (and label) dependent. In the 20 Newsgroups corpus, where the accuracy is higher, the relationship between topic coherence and supervision effects are slightly positive.

Finally, we show that the DOLDA model scales well for large corpora and many classes. Still, further improvement in scalability can be achieved with regard to the number of topics K. The ideas of Zheng et al. (2015) can probably improve the scalability with respect to K, the number of topics, but this is something we leave for future work.

Acknowledgements Open access funding provided by Aalto University.

Appendix A: Derivation of the MCMC sampler

Here we will derive the sampler presented in Sect. 4. The full joint posterior distribution is

$$\prod_{d \in D} \left[p(y_d|\mathbf{a}_d) p(\mathbf{w}_d|\mathbf{z}_d, \Phi) \right] p(\mathbf{a}|\eta, \mathbf{z}, \mathbf{X}) p(\eta|\lambda, \tau) p(\lambda) p(\tau) p(\mathbf{z}|\Theta, \Phi) p(\Theta) p(\Phi),$$

with notation summarized in Table 1.

Sampling the latent variables a

This is achieved using the same approach as in Johndrow et al. (2013, p. 34), using truncated Normal distributions.

Sampling the regression coefficients $\boldsymbol{\eta}$

The prior for η_l is given by

$$\Sigma_{\eta,l} = \mathit{\Lambda}_{\eta,l}^{-1} = \tau_l^2 \begin{pmatrix} c/\tau_l^2 & 0 & \cdots & 0 \\ 0 & \lambda_{l,1}^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{l,p}^2 \end{pmatrix},$$

where η_1 is the intercept of the model and p is the total number of η for class l. Conditioned on $\mathbf{a}$, we sample updates of each η_l as in ordinary Bayesian linear regression.

$$\eta_l \sim N\left(\mu_{n,l}, \Lambda_{post,l}^{-1}\right),$$

where

$$\mu_n = ((\mathbf{X}\bar{\mathbf{Z}})^T(\mathbf{X}\bar{\mathbf{Z}}) + \Lambda_\eta)^{-1}(\mathbf{X}\bar{\mathbf{Z}})^T\mathbf{a},$$

and

$$\Lambda_{post,l} = (\mathbf{X}\bar{\mathbf{z}})^T(\mathbf{X}\mathbf{Z}) + \Lambda_{\eta,l}.$$

Sampling of category global shrinkage parameter τ_l

The derivations are done for a given class l so we suppress the index l in the derivations. We start by deriving the unnormalized posterior distribution of τ.

$$\begin{aligned} p(\tau|\lambda,\eta) &\propto p(\eta|\lambda,\tau)\cdot p(\tau) \\ &= \frac{1}{\sqrt{(2\pi)^p c\tau^{2p}\lambda_1^2\cdots\lambda_p^2}}\exp\left(-\frac{1}{2}\eta^{\mathrm{T}}\Lambda_\eta\eta\right)\cdot\frac{2}{\pi}\cdot\frac{1}{1+\tau^2} \\ &\propto \frac{1}{\tau^p}\exp\left(-\frac{1}{2\tau^2}\left(\sum_i\frac{\eta_i^2}{\lambda_i^2}\right)\right)\cdot\frac{1}{1+\tau^2} \end{aligned}$$

We use slice sampling as presented by Scott (2010, p. 6f.). We set $\gamma = \frac{1}{\tau^2}$ that implies $\tau = \gamma^{-\frac{1}{2}}$ and let $\hat{\mu} = \sum_p^P \left(\frac{\eta_p}{\lambda_p}\right)^2$. This gives the following unnormalized posterior for γ:

$$p(\gamma|\lambda, \eta) \propto \gamma^{\frac{p}{2}} \exp\left(-\frac{1}{2}\left(\sum_i \frac{\eta_i^2}{\lambda_i^2}\right)\gamma\right) \cdot \frac{1}{1+\gamma^{-1}} \left|\frac{d}{d\gamma}\gamma^{-\frac{1}{2}}\right|$$
$$\propto \exp\left(-\frac{1}{2}\hat{\mu}\gamma\right)\gamma^{\frac{p-1}{2}} \frac{1}{\gamma+1}.$$

To sample τ we use the slice sampling algorithm of Damlen et al. (1999, Section 3.2) by setting

$$l(\gamma) = \frac{1}{1+\gamma}$$
$$\pi(\gamma) = \exp\left(-\frac{1}{2}\hat{\mu}\gamma\right)\gamma^{\frac{p-1}{2}}.$$

We can see that $\pi(\gamma)$ is the density of a Gamma distribution with $\alpha = (p+1)/2$ and $\beta = \frac{1}{2}\hat{\mu}$. We can hence sample γ in two steps:

$$u \sim U(0, (1+\gamma)^{-1})$$
$$\gamma \sim \mathrm{G}\left((p+1)/2, \frac{1}{2}\hat{\mu}^2\right) I(\gamma < (1-u)/u)$$

where $I(\cdot)$ indicates the truncation region. After sampling we transform back to τ by setting $\tau = \gamma^{-\frac{1}{2}}$.

Sampling of local shrinkage parameter λ_l per category

As with the global shrinkage parameter, the category index l is suppressed. The computations follow that of τ in large parts. We have that for each $\lambda_{i,l}$

$$p(\eta|\lambda_i, \tau)p(\lambda_i) \propto \frac{1}{\lambda_i} \exp\left(-\frac{\eta_i^2}{2\tau^2\lambda_i^2}\right) \cdot \frac{1}{1+\lambda_i^2},$$

and then we set $\gamma_i = \frac{1}{\lambda_i^2}$ with $\lambda_i = \gamma_i^{-\frac{1}{2}}$ and hence

$$p(\gamma_i) \propto \gamma_i^{\frac{1}{2}} \exp\left(-\frac{\eta_i^2}{2\tau^2\gamma_i^{-1}}\right) \cdot \frac{1}{1+\gamma_i^{-1}} \left|\frac{d}{d\gamma}\gamma_i^{-\frac{1}{2}}\right|$$
$$\propto \exp\left(-\frac{\eta_i^2}{2\tau^2}\gamma_i\right) \cdot \frac{1}{\gamma_i+1}.$$

To sample λ_i we use a similar algorithm to that of the slice sampling approach used for τ:

$$l(\gamma_i) = \frac{1}{1+\gamma_i}$$
$$\pi(\gamma_i) = \exp\left(-\frac{1}{2}\left(\frac{\eta_i}{\tau}\right)^2 \gamma_i\right),$$

and, hence, we sample

$$u_i \sim U(0, (1+\gamma_i)^{-1})$$
$$\gamma_i \sim \text{Exp}\left(\frac{1}{2}\left(\frac{\eta_i}{\tau}\right)^2\right) I(\gamma < (1-u_i)/u_i),$$

where $I(\cdot)$ indicates the truncation region and Exp is the exponential distribution. After sampling γ_i we convert back to λ_i with $\lambda_i = \gamma_i^{-\frac{1}{2}}$.

Sampling the topic indicators $\mathbf{z}|\boldsymbol{\Phi}, \boldsymbol{\eta}, \mathbf{z}^{\neg i}, \mathbf{X}$

The final step of the sampler is to derive the conditionals for the topic indicators $\mathbf{z}$. We first remove all parameters that do not depend on $\mathbf{z}$ from the joint posterior.

$$p(z_{i,d} = k|\Phi, \eta, \mathbf{z}^{\neg i}, \mathbf{x}_d) \propto p(\mathbf{w}_d|\mathbf{z}_d, \Phi) \cdot p(\mathbf{z}_d|\theta_d, \Phi) \cdot p(\mathbf{a}_d|\eta, \bar{\mathbf{z}}_d, \mathbf{x}_d),$$

where $\prod_{d\in D}[p(\mathbf{w}_d|\mathbf{z}_d, \Phi)] \cdot p(\mathbf{z}|\Theta, \Phi)$ is the standard LDA posterior. Due to the conditional conjugacy of the Dirichlet prior for the multinomial distribution, we can integrate out either Θ or both Θ and Φ. Integrating out only Θ results in either the partially collapsed sampler for parallel sampling over documents

$$p(z_{i,d} = k|\Phi, \eta, \mathbf{z}^{\neg i}, \mathbf{X}) \propto \phi_{k,v} \cdot \left(\mathbf{M}_{d,k}^{\neg i} + \alpha\right) \cdot p(\mathbf{a}_d|\eta, \bar{\mathbf{z}}_{d,k}^{\neg i}, \mathbf{x}_d),$$

where we also need to sample Φ. This can be done in parallel over topics as $\phi_k \sim \text{Dir}(\beta + \mathbf{N}_k)$.

Alternatively, we can integrate out both Θ and Φ and use the sequential collapsed sampler

$$p(z_{i,d} = k|\Phi, \eta, \mathbf{z}^{\neg i}, \mathbf{X}) \propto \frac{\mathbf{N}_{k,v_{n_d}}^{\neg i} + \beta}{\sum_j^V \left[\mathbf{N}_{k,j}^{\neg i} + \beta\right]} \cdot \left(\mathbf{M}_{d,k}^{\neg i} + \alpha\right) \cdot p(\mathbf{a}_d|\eta, \bar{\mathbf{z}}_{d,k}^{\neg i}, \mathbf{x}_d),$$

Since the latent variables of a Diagonal Orthant probit model are conditionally independent given the regression parameters Johndrow et al. (2013, Section 1.3) we have that

$$p(\mathbf{a}_d|\eta, \mathbf{z}_d^{\neg i}, \mathbf{x}_d) = \prod_{l=1}^{L} p(a_{d,l}|\eta_l, \bar{\mathbf{z}}_{d,k}^{\neg i}, \mathbf{x}_d),$$

where $p(a_{d,l}|\eta_l, \bar{\mathbf{z}}_{d,k}^{\neg i}, \mathbf{x}_d)$ is the density of the $\mathcal{N}((\bar{\mathbf{z}}_d^{\neg i}\ \mathbf{x}_{d,k})\eta_l^{\mathsf{T}}, 1)$ distribution, so

$$\begin{aligned}
p(a_{d,l}|\eta_l, \bar{\mathbf{z}}_{d,k}^{\neg i}, \mathbf{x}_d) &\propto \exp\left(-\frac{1}{2}\left(-2(\bar{\mathbf{z}}_d\,\mathbf{x}_d)\eta^{\mathsf{T}}\mathbf{a}_d^{\mathsf{T}} + (\bar{\mathbf{z}}_d\,\mathbf{x}_d)\eta^{\mathsf{T}}\eta(\bar{\mathbf{z}}_d\,\mathbf{x}_d)^{\mathsf{T}}\right)\right)\\
&= \exp\left(-\frac{1}{2}\sum^{L}\left[-2\frac{a_{d,l}}{N_d}\eta_{l,k} + 2\left((\bar{\mathbf{z}}_d^{\neg i}\ \mathbf{x}_d)\eta_l^{\mathsf{T}}\right)\frac{1}{N_d}\eta_{l,k} + \left(\frac{1}{N_d}\eta_{l,k}\right)^2\right]\right)\\
&= \exp\left(-\frac{1}{2}\sum^{L}\left[-2\frac{\eta_{l,k}}{N_d}\left(a_{d,l} - (\bar{\mathbf{z}}_d^{\neg i}\ \mathbf{x}_d)\eta_l^{\mathsf{T}}\right) + \left(\frac{\eta_{l,k}}{N_d}\right)^2\right]\right).
\end{aligned}$$

Appendix B: Efficient updating of supervision effects

One of the more important aspects of the sampler is that we need to update the supervised addition to full conditional posterior of $z_{i,d}$, $g_{d,k}^{\neg i}$. Observe that this should be computed before starting to sample each topic indicator, per document, and for all k.

$$g_{d,k} = \exp\left(-\frac{1}{2}\sum_{l}^{L}\left[-2\frac{\eta_{l,k}}{N_d}\left(a_{d,l} - (\bar{\mathbf{z}}_d\,\mathbf{x}_d)\eta_l^{\mathsf{T}}\right) + \left(\frac{\eta_{l,k}}{N_d}\right)^2\right]\right),$$

for all $k = 1, \dots K$, $d = 1, \dots D$ and $i = 1, \dots N_d$.

To sample a topic indicator we first need to compute the supervised effect for each k when the topic indicator z_i has been removed as

$$g_{d,k}^{\neg i} = \exp\left(-\frac{1}{2}\sum_{l}^{L}\left[-2\frac{\eta_{l,k}}{N_d}\left(a_{d,l} - (\bar{\mathbf{z}}_d^{\neg i}\ \mathbf{x}_d)\eta_l^{\mathsf{T}}\right) + \left(\frac{\eta_{l,k}}{N_d}\right)^2\right]\right).$$

To draw a new topic indicator we then use

$$p(z_{i,d} = k|\cdot) \propto \phi_{k,v} \cdot \left(\mathbf{M}_{d,k}^{\neg i} + \alpha\right) \cdot g_{d,k}^{\neg i}.$$

Once we have calculated $g_{d,k}$, we would like to efficiently add and withdraw a topic indicator from these values, since sampling the topic indicators is done once per token and iteration of the sampler, and the number of tokens can be very large. In the following way we can calculate the relation between $g_{d,k}$ and $g_{d,k}^{\neg i}$.

$$g_{d,k}^{\neg i} = \exp\left(-\frac{1}{2}\sum_{l}^{L}\left[-2\frac{\eta_{l,k}}{N_d}\left(a_{d,l} - (\bar{\mathbf{z}}_d^{\neg i}\ \mathbf{x}_d)\eta_l^{\mathsf{T}}\right) + \left(\frac{\eta_{l,k}}{N_d}\right)^2\right]\right)$$

$$= \exp\left(-\frac{1}{2}\sum_{l}^{L}\left[-2\frac{\eta_{l,k}}{N_d}\left(a_{d,l} - [(\bar{\mathbf{z}}_d\,\mathbf{x}_d)]\,\eta_l^{\mathsf{T}} + \eta_{l,z_{i,d}}/N_d\right) + \left(\frac{\eta_{l,k}}{N_d}\right)^2\right]\right)$$

$$= \exp\left(-\frac{1}{2}\sum_{l}^{L}\left[-2\frac{\eta_{l,k}}{N_d}\left(a_{d,l} - [(\bar{\mathbf{z}}_d\,\mathbf{x}_d)]\,\eta_l^{\mathsf{T}}\right) + \left(\frac{\eta_{l,k}}{N_d}\right)^2\right] + \sum_{l}^{L}\frac{\eta_{l,k}\eta_{l,z_{i,d}}}{N_d^2}\right)$$

$$= g_{d,k}\cdot\exp\left(\frac{1}{N_d^2}\sum_{l}^{L}\eta_{l,k}\eta_{l,z_{i,d}}\right),$$

$$https://www.overleaf.com/project/5a2c2856e2804c1b13b23cd2$$

and therefore

$$g_{k,d} = g_{d,k}^{\neg i}\cdot\exp\left(-\frac{1}{N_d^2}\sum_{l}^{L}\eta_{l,k}\eta_{l,z_{i,d}}\right),$$

where $z_{i,d}$ is the topic indicator at position i. As can be seen, to update $\mathbf{g}_d$ we need to loop over the $\mathbf{g}_d$ vector and update it element-wise when adding and removing a topic indicator $z_{i,d}$. As is shown below, it is possible to update this $\mathbf{g}_d$ vector on the fly when sampling each new token, based on

$$\begin{aligned} g_{d,k}^{\neg i} &= g_{k,d}\cdot\exp\left(\frac{1}{N_d^2}\sum_{l}^{L}\eta_{l,k}\eta_{l,z_{i,d}}\right)\\ &= g_{d,k}^{\neg(i-1)}\cdot\exp\left(-\frac{1}{N_d^2}\sum_{l}^{L}\eta_{l,k}\eta_{l,z_{(i-1),d}}\right)\cdot\exp\left(\frac{1}{N_d^2}\sum_{l}^{L}\eta_{l,k}\eta_{l,z_{i,d}}\right)\\ &= g_{d,k}^{\neg(i-1)}\cdot\exp\left(\frac{1}{N_d^2}\left[\sum_{l}^{L}\eta_{l,k}\eta_{l,z_{i,d}} - \sum_{l}^{L}\eta_{l,k}\eta_{l,z_{(i-1),d}}\right]\right), \end{aligned}$$

where $z_{i,d}$ is the topic indicator at position i, $z_{(i-1),d}$ is the previous topic indicator and $g_{d,k}^{\neg(i-1)}$ is the supervised effect for the previous topic indicator. In addition, the expression $\sum_l^L \eta_{l,k}\eta_{l,z_{i,d}}$ can be pre-calculated during each iteration further reducing the (amortized) complexity of the sampler.

References

Ahmed A, Xing EP (2010) Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In: Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 1140–1150

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc 88(422):669–679

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Carvalho C, Polson N, Scott J (2010) The horseshoe estimator for sparse signals. Biometrika 97:465–480

Castillo I, Schmidt-Hieber J, Van der Vaart A (2015) Bayesian linear regression with sparse priors. Ann Stat 43(5):1986–2018

Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM (2009) Reading tea leaves: how humans interpret topic models. In: Advances in neural information processing systems, pp 288–296

Chemudugunta C, Smyth P, Steyvers M (2007) Modeling general and specific aspects of documents with a probabilistic topic model. In: Advances in neural information processing systems, pp 241–248

Damlen P, Wakefield J, Walker S (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. J R Stat Soc Ser B (Stat Methodol) 61(2):331–344

Geladi P, Kowalski BR (1986) Partial least-squares regression: a tutorial. Anal Chim Acta 185:1–17

Griffiths TL, Steyvers M, Blei DM, Tenenbaum JB (2005) Integrating topics and syntax. In: Advances in neural information processing systems, pp 537–544

Imai K, van Dyk DA (2005) A Bayesian analysis of the multinomial probit model using marginal data augmentation. J Econom 124(2):311–334

Jameel S, Lam W, Bing L (2015) Supervised topic models with word order structure for document classification and retrieval learning. Inf Retr J 18(4):283–330

Jiang Q, Zhu J, Sun M, Xing EP (2012) Monte Carlo methods for maximum margin supervised topic models. In: Advances in neural information processing systems, pp 1592–1600

Johndrow J, Dunson D, Lum K (2013) Diagonal orthant multinomial probit models. In: Proceedings of the sixteenth international conference on artificial intelligence and statistics, pp 29–38

Jonsson L, Broman D, Magnusson M, Sandahl K, Villani M, Eldh S (2016) Automatic localization of bugs to faulty components in large scale software systems using Bayesian classification. In: 2016 IEEE international conference on software quality, reliability and security (QRS). IEEE, pp 423–430

Li X, Ouyang J, Zhou X, Lu Y, Liu Y (2015) Supervised labeled latent Dirichlet allocation for document categorization. Appl Intell 42(3):581–593

Magnusson M, Jonsson L, Villani M, Broman D (2018) Sparse partially collapsed mcmc for parallel inference in topic models. J Comput Graph Stat 27(2):449–463

McAuliffe JD, Blei DM (2008) Supervised topic models. In: Advances in neural information processing systems, pp 121–128

Mimno D, McCallum A (2012) Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. arXiv preprint arXiv:1206.3278

Mimno D, Wallach HM, Talley E, Leenders M, McCallum A (2011) Optimizing semantic coherence in topic models. In: Proceedings of the 2011 conference on empirical methods in natural language processing. association for computational linguistics, pp 262–272

Mullen L (2016) tokenizers: a consistent interface to tokenize natural language text. R package version 0.1.4

Nalenz M, Villani M (2018) Tree ensembles with rule structured horseshoe regularization. Ann Appl Stat 12(4):2379–2408

Newman D, Asuncion A, Smyth P, Welling M (2009) Distributed algorithms for topic models. J Mach Learn Res 10(Aug):1801–1828

Parnin C, Orso A (2011) Are automated debugging techniques actually helping programmers? In: Proceedings of the 2011 international symposium on software testing and analysis. ACM, pp 199–209

Perotte AJ, Wood F, Elhadad N, Bartlett N (2011) Hierarchically supervised latent Dirichlet allocation. In: Advances in neural information processing systems, pp 2609–2617

Polson NG, Scott JG, Windle J (2013) Bayesian inference for logistic models using Pólya-gamma latent variables. J Am Stat Assoc 108(504):1339–1349

Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. In: Proceedings of the 20th conference on uncertainty in artificial intelligence. AUAI Press, pp 487–494

Rubin TN, Chambers A, Smyth P, Steyvers M (2012) Statistical topic models for multi-label document classification. Mach Learn 88(1–2):157–208

Sandhaus E (2008) The New York Times annotated corpus LDC2008T19. Linguistic Data Consortium, Philadelphia

Scott JG (2010) Parameter expansion in local-shrinkage models. arXiv preprint arXiv:1010.5265

Wallach HM, Mimno DM, McCallum A (2009) Rethinking LDA: why priors matter. In: Advances in neural information processing systems, pp 1973–1981

Zheng X, Yu Y, Xing EP (2015) Linear time samplers for supervised topic models using compositional proposals. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1523–1532

Zhu J, Ahmed A, Xing EP (2012) MedLDA: maximum margin supervised topic models. J Mach Learn Res 13(1):2237–2278

Zhu J, Zheng X, Zhang B (2013) Improved Bayesian logistic supervised topic models with data augmentation. In: Proceedings of the 51st annual meeting of the association for computational linguistics, vol 1, pp 187–195

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.