2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Deep Topic Modeling by Multilayer Bootstrap Network and Lasso

Jianyu Wang
School of Marine Science and Technology
Northwestern Polytechnical University
Xi'an, China
Email: alexwang96@mail.nwpu.edu.cn

Xiao-Lei Zhang
School of Marine Science and Technology
Northwestern Polytechnical University
Xi'an, China
Email: xiaolei.zhang@nwpu.edu.cn

***Abstract*—Topic modeling is widely studied for the dimension reduction and analysis of documents. However, it is formulated as a difficult optimization problem. Current approximate solutions also suffer from inaccurate model- or data-assumptions. To deal with the above problems, we propose a polynomial-time deep topic model with no model and data assumptions. Specifically, we first apply multilayer bootstrap network (MBN), which is an unsupervised deep model, to reduce the dimension of documents, and then use the low-dimensional data representations or their clustering results as the target of supervised Lasso for topic word discovery. To our knowledge, this is the first time that MBN and Lasso are applied to unsupervised topic modeling. Experimental comparison results with five representative topic models on the 20-newsgroups and TDT2 corpora illustrate the effectiveness of the proposed algorithm.**

I. Introduction

Topic modeling is an unsupervised method that learns latent structures and salient features from document collections. It is originally formulated as a hierarchical generative model: a document is generated from a mixture of topics, and a word in the document is generated by first choosing a topic from a document-specific distribution, and then choosing the word from the topic-specific distribution. The main difficulty of topic modeling is the optimization problem, which is NP-hard in the worst case due to the intractability of the posterior inference. Existing methods aim to find approximate solutions to the difficult optimization problem, which falls into the framework of matrix factorization.

Matrix factorization based topic modeling maps documents into a low-dimensional semantic space by decomposing the documents into a weighted combination of a set of topic distributions: $\mathbf{D} \approx \mathbf{CW}$ where $\mathbf{D}(:, d)$ represents the d-th document which is a column vector over a set of words with a vocabulary size of v, $\mathbf{C}(:, g)$ denotes the g-th topic which is a probability mass function over the vocabulary, and $W(g, d)$ denotes the probability of the g-th topic in the d-th document. Existing methods for the matrix decomposition can be categorized to two classes in general—probabilistic models [1]–[4] and nonnegative matrix factorizations (NMF) [5]–[7].

A seminal work of probabilistic models is latent Dirichlet allocation (LDA) [1]. It assumes that each document is a sample from a multinomial distribution whose parameters are generated from $\mathbf{CW}(:, d)$. Each column of $\mathbf{C}$ and $\mathbf{W}$ also represent multinomial distributions independently drawn from Dirichlet distributions. It adopts Kullback-Leibler divergence to measure the distance between $\mathbf{D}$ and $\mathbf{CW}$, since the posterior distributions $p(\mathbf{W}|\mathbf{D})$ and $p(\mathbf{C}|\mathbf{D})$ are coupled. Later on, many models followed the above framework, such as hierarchical Dirichlet process [8], [9] and Laplacian probabilistic semantic indexing [10]. However, the model assumptions, such as the multinomial distribution, may not be always accurate.

The validness of NMF comes from the fact that the matrices $\mathbf{C}$ and $\mathbf{W}$ should be nonnegative. The objective function of NMF is generally as follows:

$$(\mathbf{C}, \mathbf{W}) = \arg \min_{\mathbf{C} \geq \mathbf{0}; \mathbf{W} \geq \mathbf{0}} \| \mathbf{D} - \mathbf{CW} \|_F^2 \tag{1}$$

An important weakness of this formulation is that there is no guarantee that the solutions of $\mathbf{C}$ and $\mathbf{W}$ are unique [11]. To solve the identifiability problem, many NMF methods adopted an *anchor word* assumption, which assumes that every topic has a characteristic anchor word that does not appear in the other topics [12]. However, this assumption may not always hold in practice. Recently, an anchor-free NMF based on the second-order statistics of documents [13] has been proposed, which significantly improved the performance of NMF methods. Another problem of NMF is that it is formulated as a shallow learning method, which may not capture the nonlinearity of documents.

Motivated by the above problems, this paper proposes a deep topic model (DTM), which learns a deep representation of the documents, i.e. $f(\mathbf{D})$, and the topic-word matrix $\mathbf{C}$ separately, under the assumption that if each of the components is good enough, then the overall performance can be boosted. Specifically, we apply multilayer bootstrap networks (MBN) [14] to learn a document-topic projection $f(\cdot)$ first, and then apply Lasso [15] to learn $\mathbf{C}$ given $f(\mathbf{D})$. MBN is a simple nonparametric deep model for unsupervised dimension reduction, which overcomes the problems of model assumptions, shallow learning, and anchor word assumption. Given the output of MBN, the topic modeling becomes a simple supervised regression problem. We employ Lasso for this problem. Empirical results on the TDT2 and 20-newsgroups corpora illustrate the effectiveness of the proposed algorithm.

II. Deep topic modeling

A. Object function

The objective of DTM is defined as

$$\min_{f(\cdot),\mathbf{C}} \frac{1}{2}\|\mathbf{C}f(\mathbf{D}) - \mathbf{D}\|_F^2 + \lambda\Omega(\mathbf{C}) \tag{2}$$

where $f(\cdot)$ is an unsupervised deep model containing multiple layers of nonlinear transforms, $\Omega(\cdot)$ is a regularizer, and λ is a regularization hyperparameter. We optimize (2) in two steps. The first step learns $f(\mathbf{D})$ by MBN, which outputs the document-topic matrix $\mathbf{W}$. The second step learns the topic-word matrix $\mathbf{C}$ by Lasso, given $\mathbf{W} = f(\mathbf{D})$. The overall DTM algorithm is shown in Fig. 1.

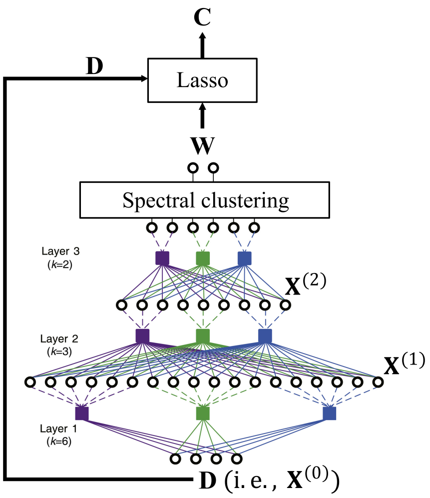

Fig. 1: Deep topic model.

B. Multilayer bootstrap network

The network structure of MBN is shown in Fig. 1. It is a deep dimensionality reduction algorithm optimized by random resampling of data and one-nearest-neighbor optimization [14]. It consists of L gradually narrowed hidden layers from bottom-up. Each hidden layer consists of V k-centroids clusterings ($V > 100$), where parameter k at the l-th layer is denoted by k_l, $l = 1, \ldots, L$. Each k_l-centroids clustering has k_l output units, each of which indicates one cluster. The output layer is the linear-kernel-based spectral clustering [16]. It outputs $\mathbf{W}$, which is used as the input of the Lasso component.

MBN is trained layer-by-layer from bottom-up. To train the l-th layer, we simply need to focus on training each k_l-centroids clustering as follows:

- **Random sampling of input**. The first step randomly selects k_l data points from $\mathbf{X}^{(l-1)} = [\mathbf{x}_1^{(l-1)}, \ldots, \mathbf{x}_N^{(l-1)}]$ as the k_l centroids of the clustering, where N is the size of the corpus. If $l = 1$, then $\mathbf{X}^{(l-1)} = \mathbf{D}$.
- **One-nearest-neighbor learning**. The second step assigns any input $\mathbf{x}^{(l-1)}$ to one of the k_l clusters and outputs a k_l-dimensional indicator vector $\mathbf{h} = [h_1, \ldots, h_{k_l}]^T$, which is a one-hot sparse vector.

The output units of all k_l-centroids clusterings are concatenated as the input of their upper layer, i.e. $\mathbf{x}^{(l)} = [\mathbf{h}_1^T, \ldots, \mathbf{h}_V^T]^T$. From the above description, we can see that MBN does not make any model or data assumptions.

Note that the parameter setting $\{k_l\}_{l=1}^L$ is important to maintain the tree property of MBN. In practice, it obeys the following criterion:

$$k_1 = \lfloor N/2 \rfloor, \quad k_l = \lfloor \delta k_{l-1} \rfloor \tag{3}$$

$$k_L \approx \begin{cases} \lceil \frac{N_Z}{N_z} \rceil, & \text{if } \mathbf{D} \text{ is strongly class imbalanced} \\ 1.5c, & \text{otherwise} \end{cases} \tag{4}$$

where $\delta \in (0, 1)$ is a user defined hyperparameter with 0.5 as its default, c is the number of topics, N_Z and N_z are the numbers of the documents belonging to the largest and smallest topics respectively. δ controls the network structure. (4) guarantees that at least one data point is sampled from each of the topics in probability. In other words, it ensures that the random samples at the top hidden layer is an effective model.

C. Lasso

Substituting the output of MBN, i.e. $\mathbf{W}$, to (2) derives:

$$\min_{\mathbf{C}} \frac{1}{2}\|\mathbf{CW} - \mathbf{D}\|_F^2 + \lambda\Omega(\mathbf{C}) \tag{5}$$

(5) is a typical regularized regression problem [17]. Many regression models can be applied to (5). Here we choose Lasso, given its strong ability on the feature selection and prediction problems for high-dimensional data. Specifically, we use Lasso to calculate the conditional probability distribution of each word over the topics [15], which is formulated as the follow problem:

$$\min_{\mathbf{C}(i,:)} \frac{1}{2} \parallel \mathbf{C}(i,:)\mathbf{W} - \mathbf{D}(i,:) \parallel_2^2 + \lambda \left\|\mathbf{C}(i,:)\right\|_1 \tag{6}$$

where $i = 1, \ldots, v$ is the index of the i-th word. We adopt the *alternating direction method of multipliers* (ADMM) [18] solver to solve problem (6).[1]

III. Related work

It is known that the main difficulty of hierarchical probabilistic topic models is the high computation on the inference problem of the hidden variables. Topic models based on deep variational auto-encoders overcome the difficulty. They generally can be decomposed into two modules: an inference network $q(\mathbf{h}|\mathbf{D}(:,d))$ which compresses the documents into continuous hidden vectors $\mathbf{h}$ by deep neural networks, and a generative model $p(\mathbf{D}(:,d)|\mathbf{h}) = \prod_{v=1}^V p(\mathbf{D}(v,d)|\mathbf{h})$ which reconstructs the documents by generating the words independently from $\mathbf{h}$ [19] via restricted Boltzmann machines, sigmoid belief networks, Dirichlet processes, etc [20], where $\mathbf{D}(v,d)$ is the vth word of the document $\mathbf{D}(:,d)$. They maximise the

[1] https://github.com/foges/pogs

evidence lower-bound of the joint likelihood of the documents and hidden variables:

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{h}|\mathbf{D}(:,d))}\left[\sum_{v=1}^{V}\log p(\mathbf{D}(v,d)|\mathbf{h})\right] - D_{KL}[q(\mathbf{h}|\mathbf{D}(:,d))\|p(\mathbf{h})] \tag{7}$$

where $D_{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence between two distributions, $\mathbb{E}_{q(\mathbf{h}|\mathbf{D}(:,d))}(\cdot)$ is the expectation operator over $q(\mathbf{h}|\mathbf{D}(:,d))$, and $p(\mathbf{h})$ is a prior for $\mathbf{h}$. The above models integrate the power of neural networks into the inference of the probabilistic topic models, which not only helps the probabilistic topic models scalable to big datasets but also speeds up the convergence of the probabilistic topic models significantly. However, the prior assumption of $\mathbf{h}$ may not always hold, and moreover, the inference network faces a problem of component collapsing [21] which is a kind of bad local optima that is particularly endemic to auto-encoding variational Bayes and similar methods. On the contrary, the proposed method not only is able to generate deep representations of the documents but also does not suffer the aforementioned weaknesses.

IV. Experiments

A. Experimental settings

We conducted experiments on the 20-newsrgoups and the top 30 largest topics of the NIST Topic Detection and Tracking (TDT2) corpora. 20-Newsgroups consists of 18,846 documents with a vocabulary size of 26,214. The subset of TDT2 consists of 9,394 documents with a vocabulary size is 36,771. For each corpus, we randomly sampled $c = 5, 10, 15, 20$ topics respectively, and reported the average results over 50 Monte-Carlo runs. The indices of the 50 runs on TDT2 are the same as those at http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html. We used TF-IDF as the feature. We used cosine similarity to measure the similarity of two documents in the TF-IDF space.

For the proposed DTM, we set the hyperparameters of MBN and Lasso to their default values, i.e. $V = 400$, $\delta = 0.5$, and $\lambda = 1/3$.[2] We compared DTM to the following five representative topic modeling methods:

- **LDA** [1]: it is a seminal probabilistic model based on multinomial and Dirichlet distributions.
- **Locally-consistent topic modeling (LTM)** [22]: it extends the probabilistic latent semantic indexing algorithm [23] by incorporating cosine similarity kernel to model the local manifold structure of documents.
- **Successive nonnegative projection (SNPA)** [24]: it is an NMF method. It does not require the matrix $\mathbf{W}$ to be full rank, which makes it more robust to noise than traditional NMF methods.
- **Anchor-free correlated topic modeling (AchorFree)** [13]: it is an NMF method. It does not have the anchor-word assumption, which makes it behave much better than traditional NMF methods.

[2]The default value of λ is in the implementation of the ADMM algorithm.

- **Deep Poisson Factor Modeling (DPFM)** [20]: it is a deep learning based topic model built on the Dirichlet process. We set its DNN to a depth of two hidden layers, and set the number of the hidden units of the two hidden layers to c and $\lceil c/2 \rceil$ respectively for its best performance. We used the output from the first hidden layer for clustering.

We further compared MBN with a cosine-similarity-kernel-based spectral clustering (SC) algorithm [16], and compared DTM with SC+Lasso, for evaluating the effects of MBN on performance.

B. Evaluation Metrics

We evaluated the comparison results in terms of *clustering accuracy* (ACC), *coherence* (Coh.) [25], and *similarity count* (SimC.). Coherence evaluates the quality of a single mined topic. It is calculated by

$$\mathrm{Coh}(\nu) = \sum_{v_1, v_2 \in \nu} \log \frac{\mathrm{freq}(v_1, v_2) + \varepsilon}{\mathrm{freq}(v_2)} \tag{8}$$

where v_1 and v_2 denote two words in the vocabulary ν, $freq(v_1, v_2)$ denotes the number of the documents where v_1 and v_2 co-appear, $freq(v_2)$ denotes the number of the documents containing v_2, and $\varepsilon = 0.01$ is used to prevent the logarithm operator from zero. The higher the clustering accuracy or coherence score is, the better the topic model is. Because the coherence measurement does not evaluate the redundancy of a topic, we use the similarity count to measure the similarity between topics. For each topic, the similarity count is obtained simply by adding up the overlapped words of the topics within the leading c words. The lower the similarity count score is, the better the topic model is.

C. Results

Table I shows the comparison results on the 20-newsgroups corpus. From the table, we see that DTM achieves higher clustering accuracy than the other algorithms. For example, DTM achieves more than 5% absolute clustering accuracy improvement over the runner-up method LTM when $c = 10$, and 1% higher in other cases. In addition, the single-topic quality of the topics mined by DTM ranks the fourth in terms of coherence. The overlaps between the topics mined by DTM ranks the second in terms of similarity count.

Table II shows the results on the TDT2 corpus. From the table, we can see that DTM obtains the best performance in terms of clustering accuracy and similarity count evaluation metrics. For example, the clustering accuracy produced by DTM is over 4% absolutely higher than that of the runner-up method when mining 5 topics, and over 13% higher than the latter when $c = 10$.

Table III shows the results on the Reuters-21578 corpus. From the table, we can see that DTM obtains the best performance in terms of clustering accuracy and similarity count. For instance, the clustering accuracy produced by DTM is over 7% absolutely higher than that of the runner-up method when $c = 3$. We further averaged all six ranking lists in Tables

TABLE I: Comparison results on 20-newsgroups.

Metric	Model	c=3	c=4	c=5	c=6	c=7	c=8	c=9	c=10	c=15	c=20	rank
ACC	LDA	0.81	0.73	0.70	0.66	0.65	0.62	0.60	0.59	0.52	0.49	5.9
	LTM	0.90	0.83	0.84	0.82	0.79	0.77	0.72	0.71	0.63	0.60	2.1
	SNPA	0.43	0.33	0.29	0.26	0.24	0.22	0.21	0.21	0.18	0.15	7.0
	AnchorFree	0.88	0.84	0.76	0.71	0.71	0.69	0.66	0.64	0.52	0.45	4.6
	SC	0.84	0.81	0.79	0.76	0.74	0.71	0.69	0.67	0.58	0.52	3.2
	DPFM	0.84	0.81	0.79	0.76	0.74	0.71	0.69	0.67	0.58	0.52	4.2
	DTM	0.91	0.89	0.87	0.86	0.83	0.81	0.77	0.76	0.68	0.64	1.0
Coh.	LDA	-603.3	-634.1	-651.7	-678.7	-686.5	-702.2	-716.6	-729.3	-762.8	-759.1	3.0
	LTM	-636.5	-677.5	-704.4	-753.5	-741.0	-778.8	-790.3	-799.5	-854.7	-855.9	5.2
	SNPA	-801.3	-792.6	-819.3	-863.5	-856.6	-873.9	-907.8	-911.9	-1055.7	-901.7	7.0
	AnchorFree	-572.9	-573.3	-565.9	-538.8	-544.2	-566.8	-562.4	-571.9	-575.6	-596.1	1.2
	SC	-670.3	-689.8	-712.8	-742.6	-751.1	-742.1	-775.2	-774.8	-831.5	-890.2	5.3
	DPFM	-534.4	-585.5	-562.9	-588.6	-587.2	-592.8	-605.6	-616.2	-640.4	-676.1	1.8
	DTM	-681.1	-681.4	-696.6	-729.9	-731.1	-736.2	-760.4	-768.4	-823.2	-899.8	4.5
SimC.	LDA	9.4	14.6	22.3	33.1	37.2	42.5	50.5	66.4	116.2	196.0	4.0
	LTM	0.4	1.0	1.6	2.3	4.3	5.2	6.4	9.2	19.6	26.0	1.1
	SNPA	5.5	12.2	13.2	19.7	22.3	24.1	28.1	23.6	21.5	15.0	2.7
	AnchorFree	10.4	21.3	32.2	53.1	76.1	112.1	142.0	195.8	598.8	1235.0	5.4
	SC	21.6	38.4	64.2	95.7	124.1	160.7	193.1	224.8	365.6	496.1	5.8
	DPFM	21.6	38.4	64.2	95.7	124.1	160.7	193.1	224.8	365.6	496.1	6.8
	DTM	1.7	3.8	4.3	7.2	9.8	13.8	16.6	20.0	39.4	54.8	2.2

I, II and III. The overall rank of the 6 comparison methods from the number one to number six is DTM, AnchorFree, LTM, SC, SNPA, LDA, and DPFM, respectively.

Table IV shows the top 20 topic words discovered by AnchorFree and DTM respectively when mining a corpus of 5 topics in TDT2. From the table, we can see that DTM produces more discriminative topic words than AnchorFree. Specifically, DTM does not produce overlapping words, while AnchorFree produces 23 overlapping words among the 100 topic words. The topic words of the second and fifth topics produced by AnchorFree have an overlap of over 50%. Some informative words discovered by DTM, such as the words related to anti-government activities or violence in the fifth topic, were not detected by AnchorFree. The above phenomenon is observed in other experiments too. We conjecture that the additional conditional assumptions made by AnchorFree, such as consecutive words being persistently drawn from the same topic, might affect the topic characterization. On the contrary, the proposed DTM not only avoids making additional assumptions but also does not suffer the weaknesses of deep neural networks.

D. *Effects of hyperparameters on performance*

We study V and δ independently on 20-newsgroups, TDT2 and Reuters-21578. When we study a hyperparameter, we tune it in a range, leaving the other hyperparameters to their default values. The experimental results are shown in Fig. 2. From Figs. 2a, 2c and 2e, we see that enlarging V increases the accuracy of DTM steadily, and the performance of DTM becomes stable when $V > 100$. However, increasing V enlarges the computational complexity of DTM as well. To balance the accuracy and computational complexity, setting $V = 400$ is reasonable. From Figs. 2b, 2d and 2f, we observe that, although the performance is relatively sensitive to hyperparameter δ, the hyperparameter has a stable interval around the default value 0.5. To conclude, setting $\delta = 0.5$ is safe for DTM.

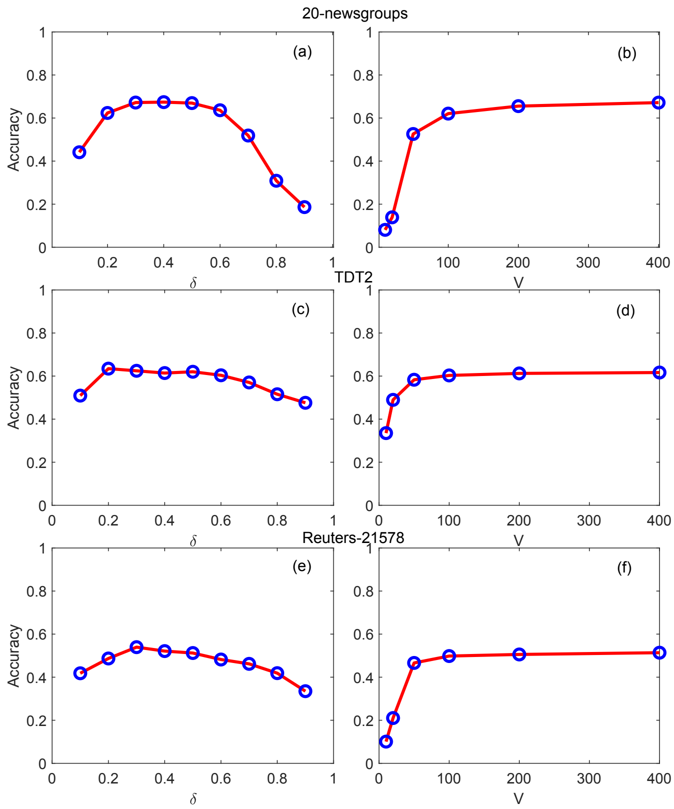

Fig. 2: Effect of hyperparameters V and δ on performance.

TABLE II: Comparison results on TDT2.

Metric	Model	c=3	c=4	c=5	c=6	c=7	c=8	c=9	c=10	c=15	c=20	c=25	rank
ACC	LDA	0.79	0.74	0.70	0.68	0.66	0.62	0.65	0.64	0.59	0.61	0.61	5.8
	LTM	0.99	0.95	0.94	0.92	0.86	0.84	0.81	0.77	0.69	0.65	0.63	2.7
	SNPA	0.79	0.73	0.70	0.64	0.61	0.58	0.58	0.56	0.47	0.46	0.44	7.0
	AnchorFree	0.97	0.95	0.92	0.91	0.90	0.87	0.87	0.85	0.80	0.77	0.74	2.3
	SC	0.92	0.82	0.79	0.75	0.70	0.68	0.70	0.67	0.63	0.58	0.59	5.0
	DPFM	0.88	0.82	0.80	0.79	0.75	0.72	0.75	0.73	0.68	0.68	0.65	4.0
	DTM	0.97	0.99	0.98	0.98	0.94	0.94	0.94	0.90	0.85	0.82	0.76	1.2
Coh.	LDA	-427.4	-510.3	-509.8	-546.0	-543.6	-565.3	-570.7	-574.4	-617.9	-642.5	-666.1	4.2
	LTM	-678.4	-660.6	-634.3	-626.2	-597.0	-594.9	-594.8	-597.6	-579.3	-616.1	-635.0	4.8
	SNPA	-613.9	-592.5	-611.0	-642.8	-646.1	-657.7	-655.4	-668.1	-660.3	-679.5	-686.6	5.7
	AnchorFree	-419.1	-430.8	-407.0	-428.8	-397.8	-445.8	-418.1	-422.3	-433.0	-458.3	-469.5	1.5
	SC	-353.9	-429.0	-441.6	-468.2	-507.4	-492.9	-488.6	-517.6	-542.9	-566.7	-584.3	2.4
	DPFM	-952.1	-888.2	-803.9	-831.7	-731.3	-704.8	-755.2	-715.7	-676.8	-627.0	-588.1	6.4
	DTM	-360.1	-358.7	-402.1	-389.6	-430.0	-469.2	-508.9	-505.1	-623.9	-684.0	-706.6	3.0
SimC.	LDA	2.8	5.3	8.0	11.9	16.1	21.1	25.5	30.5	65.1	104.8	147.2	4.4
	LTM	24.4	25.3	24.7	23.1	25.4	23.9	24.3	23.3	23.3	20.8	22.8	4.1
	SNPA	16.1	24.0	29.4	44.1	53.5	58.8	74.5	74.8	189.4	271.5	450.1	5.8
	AnchorFree	4.1	2.2	4.9	6.6	4.5	8.8	9.9	13.5	40.8	79.7	132.3	2.6
	SC	0.2	2.4	3.6	8.9	12.2	13.2	19.3	25.4	44.1	91.8	123.2	3.0
	DPFM	50.1	54.1	112.2	113.8	191.4	192.3	285.1	287.8	690.2	1056.2	1741.3	7.0
	DTM	0.5	0.1	0.2	0.4	1.0	1.0	1.1	2.2	5.9	10.0	18.2	1.1

TABLE III: Comparison results on Reuters-21578.

Metric	Model	c=3	c=4	c=5	c=6	c=7	c=8	c=9	c=10	c=15	c=20	c=25	rank
ACC	LDA	0.63	0.57	0.53	0.51	0.46	0.44	0.41	0.42	0.35	0.33	0.34	6.9
	LTM	0.74	0.69	0.62	0.58	0.58	0.57	0.55	0.53	0.42	0.36	0.36	4.5
	SNPA	0.79	0.73	0.70	0.64	0.61	0.58	0.58	0.56	0.47	0.46	0.44	2.7
	AnchorFree	0.79	0.73	0.65	0.64	0.65	0.61	0.59	0.58	0.52	0.53	0.47	2.1
	SC	0.66	0.62	0.60	0.58	0.52	0.46	0.43	0.44	0.37	0.34	0.34	6.1
	DPFM	0.72	0.66	0.62	0.62	0.55	0.55	0.51	0.54	0.51	0.46	0.42	4.5
	DTM	0.86	0.77	0.72	0.68	0.69	0.65	0.62	0.61	0.52	0.49	0.47	1.3
Coh.	LDA	-674.1	-677.2	-686.3	-715.2	-705.9	-762.9	-776.8	-776.5	-847.7	-903.4	-902.7	3.7
	LTM	-943.6	-952.3	-942.7	-947.6	-940.7	-967.2	-975.1	-945.3	-959.2	-955.9	-933.0	6.0
	SNPA	-613.9	-592.5	-611.0	-642.8	-646.1	-657.7	-655.4	-668.1	-660.3	-679.5	-686.6	1.2
	AnchorFree	-827.3	-744.0	-771.6	-699.5	-684.5	-722.7	-711.0	-703.6	-685.3	-678.4	-667.8	2.9
	SC	-644.6	-619.2	-657.5	-691.7	-664.8	-696.2	-692.2	-690.9	-720.2	-744.4	-763.2	2.3
	DPFM	-996.3	-1017.1	-1045.2	-1046.7	-982.4	-901.2	-858.3	-911.3	-950.8	-911.1	-905.4	5.9
	DTM	-808.3	-812.6	-828.3	-847.9	-918.7	-976.2	-1002.4	-975.8	-1030.0	-1102.3	-1099.3	6.0
SimC.	LDA	230.8	218.3	223.4	228.0	221.3	277.8	332.5	276.2	209.5	222.6	202.9	6.1
	LTM	45.1	39.6	38.8	40.6	41.7	47.0	55.4	51.2	46.2	49.4	48.9	3.2
	SNPA	16.1	24.0	29.4	44.1	53.5	58.8	74.5	74.8	189.4	271.5	450.1	4.7
	AnchorFree	7.3	12.0	16.9	19.6	33.4	61.6	69.8	86.0	126.7	226.0	339.7	3.7
	SC	3.3	7.7	10.4	15.5	25.4	41.5	57.2	63.1	137.7	252.9	312.2	2.8
	DPFM	49.7	51.2	104.9	109.7	190.5	189.8	289.2	287.5	658.4	1000.5	1607.3	6.4
	DTM	0.7	1.8	2.3	3.9	4.7	7.2	8.3	10.6	25.2	37.6	50.7	1.1

V. Conclusion

In this paper, we have proposed a deep topic model based on MBN and Lasso. The novelty of DTM lies in the following three respects. First, we extended the linear matrix factorization problem to its nonlinear case. Second, we estimated the topic-document matrix and word-topic matrix separately by MBN and Lasso independently, which simplifies the optimization problem of (2). At last, we applied MBN and Lasso to the unsupervised topic modeling for the first time. Particularly, MBN, as an unsupervised deep model, overcomes the weaknesses of the model assumptions, anchor word assumption, and shallow learning, which accounts for the advantage of DTM over the 5 representative comparison methods. Experimental results on 20-newsgroups, TDT2 and Reuters-21578 have demonstrated the effectiveness of the proposed method.

Acknowledgement

We thank Dr.Kejun Huang for sharing the source code of the AnchorFree algorithm.

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2018AAA0102200, in part by National Science Foundation of China under Grant No. 61761146001, 61831019, 61671381, in part by the Project of the Science, Technology, and Innovation Commission of Shenzhen Municipality under grant No. JCYJ20170815161820095, and in part by the Open Research Project of the State Key Laboratory of Media Convergence and Communication, Communication University of China, China under Grant No. SKLMCC2020KF009.

TABLE IV: Topic words discovered by DRTM and AnchorFree on a 5-topic subset of TDT2 corpus. The topic words in bold denotes overlapped words between topics.

AnchorFree					DRTM				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
netanyahu	**asian**	bowl	tornadoes	**economic**	netanyahu	asian	bowl	florida	nigeria
israeli	**asia**	super	florida	**indonesia**	israeli	percent	super	tornadoes	abacha
israel	**economic**	broncos	central	**asian**	israel	indonesia	broncos	tornado	military
palestinian	**financial**	denver	storms	**financial**	palestinian	asia	denver	storms	police
peace	**percent**	packers	ripped	imf	peace	economy	packers	killed	nigerian
arafat	**economy**	bay	victims	**economy**	albright	financial	green	victims	opposition
palestinians	market	green	tornado	**crisis**	arafat	market	game	damage	nigerias
albright	stock	football	homes	**asia**	palestinians	stock	bay	homes	anti
benjamin	**crisis**	game	killed	monetary	talks	economic	football	ripped	elections
west	markets	san	people	**currency**	west	billion	elway	nino	arrested
talks	stocks	elway	damage	**billion**	benjamin	crisis	san	el	lagos
bank	**currency**	diego	twisters	fund	madeleine	imf	team	weather	democracy
prime	prices	xxxii	nino	**percent**	london	japan	sports	twisters	sani
london	dollar	nfl	el	international	ross	spkr	diego	storm	syscivilиантem
minister	investors	quarterback	deadly	government	withdrawal	currency	coach	rain	protest
yasser	index	sports	storm	**bank**	process	markets	play	stories	protests
ross	**billion**	play	counties	korea	prime	dollar	win	deadly	presidential
withdrawal	**bank**	yards	weather	south	yasser	south	teams	struck	abachas
madeleine	growth	favre	funerals	indonesian	secretary	government	season	residents	violent
13	**indonesia**	pittsburgh	toll	suharto	13	prices	fans	california	nigerians

REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[2] Tomoharu Iwata, Tsutomu Hirao, and Naonori Ueda, "Unsupervised cluster matching via probabilistic latent variable models.," in *AAAI*, 2013.

[3] Tomoharu Iwata, Tsutomu Hirao, and Naonori Ueda, "Topic models for unsupervised cluster matching," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 4, pp. 786–795, 2018.

[4] Hosein Azarbonyad, Mostafa Dehghani, Tom Kenter, Maarten Marx, Jaap Kamps, and Maarten de Rijke, "Hitr: Hierarchical topic model re-estimation for measuring topical diversity of documents," *IEEE Transactions on Knowledge and Data Engineering*, 2018, To appear.

[5] Nicolas Gillis and Stephen A Vavasis, "Fast and robust recursive algorithmsfor separable nonnegative matrix factorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 4, pp. 698–714, 2014.

[6] Junyu Xuan, Jie Lu, Guangquan Zhang, Richard Yi Da Xu, and Xiangfeng Luo, "Doubly nonparametric sparse nonnegative matrix factorization based on dependent indian buffet processes," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 5, pp. 1835–1849, 2018.

[7] Jesx00FAs Bobadilla, Rodolfo Bojorque, Antonio Hernando Esteban, and Remigio Hurtado, "Recommender systems clustering using bayesian non negative matrix factorization," *IEEE Access*, vol. 6, pp. 3549–3564, 2018.

[8] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei, "Sharing clusters among related groups: Hierarchical dirichlet processes," in *NIPS*, 2005, pp. 1385–1392.

[9] Weifeng Li, Junming Yin, and Hsinchsun Chen, "Supervised topic modeling using hierarchical dirichlet process-based inverse regression: Experiments on e-commerce applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1192–1205, 2018.

[10] Deng Cai, Qiaozhu Mei, Jiawei Han, and Chengxiang Zhai, "Modeling hidden topics on document manifold," in *CIKM*. ACM, 2008, pp. 911–920.

[11] Wei Xu, Xin Liu, and Yihong Gong, "Document clustering based on non-negative matrix factorization," in *ACM SIGIR*. ACM, 2003, pp. 267–273.

[12] Kejun Huang, Xiao Fu, and Nikolaos D Sidiropoulos, "Anchor-free correlated topic modeling: Identifiability and algorithm," in *NIPS*, 2016, pp. 1786–1794.

[13] Xiao Fu, Kejun Huang, Nicholas D Sidiropoulos, Qingjiang Shi, and Mingyi Hong, "Anchor-free correlated topic modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[14] Xiao-Lei Zhang, "Multilayer bootstrap networks," *Neural Networks*, vol. 103, pp. 29–43, 2018.

[15] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[16] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm," in *NIPS*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds., pp. 849–856. MIT Press, 2002.

[17] Makoto Yamada, Takeuchi Koh, Tomoharu Iwata, John Shawe-Taylor, and Samuel Kaski, "Localized Lasso for High-Dimensional Regression," in *AISTATS*, Aarti Singh and Jerry Zhu, Eds., Fort Lauderdale, FL, USA, 20–22 Apr 2017, vol. 54 of *Proceedings of Machine Learning Research*, pp. 325–333, PMLR.

[18] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," 2010.

[19] Yishu Miao, Lei Yu, and Phil Blunsom, "Neural variational inference for text processing," in *International conference on machine learning*, 2016, pp. 1727–1736.

[20] Ricardo Henao, Zhe Gan, James Lu, and Lawrence Carin, "Deep poisson factor modeling," in *Advances in Neural Information Processing Systems*, 2015, pp. 2800–2808.

[21] Akash Srivastava and Charles Sutton, "Autoencoding variational inference for topic models," *arXiv preprint arXiv:1703.01488*, 2017.

[22] Deng Cai, Xuanhui Wang, and Xiaofei He, "Probabilistic dyadic data analysis with local and global consistency," in *ICML*, New York, NY, USA, 2009, ICML '09, pp. 105–112, ACM.

[23] Thomas Hofmann, "Probabilistic latent semantic indexing," in *ACM SIGIR Forum*. ACM, 2017, vol. 51, pp. 211–218.

[24] Nicolas Gillis, "Successive nonnegative projection algorithm for robust nonnegative blind source separation," *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 1420–1450, 2014.

[25] John M Neuhaus and Jack D Kalbfleisch, "Between-and within-cluster covariate effects in the analysis of clustered data," *Biometrics*, pp. 638–645, 1998.