How to Supervise Topic Models

Cheng Zhang Hedvig Kjellström

Computer Vision and Active Perception Lab, Centre for Autonomous Systems
KTH Royal Institute of Technology,Stockholm, Sweden
{chengz,hedvig}@kth.se

Abstract. Supervised topic models are important machine learning tools which have been widely used in computer vision as well as in other domains. However, there is a gap in the understanding of the supervision impact on the model. In this paper, we present a thorough analysis on the behaviour of supervised topic models using Supervised Latent Dirichlet Allocation (SLDA) and propose two factorized supervised topic models, which factorize the topics into signal and noise. Experimental results on both synthetic data and real-world data for computer vision tasks show that supervision need to be boosted to be effective and factorized topic models are able to enhance the performance.

Keywords: Topic Modeling, SLDA, LDA, Factorized Supervised Topic Models

1 Introduction

Topic modelling, as one of the most important machine learning tools, has been successfully applied to in computer vision [8, 5, 22, 24, 11, 15], as well as other domains. It is a type of generative latent structure model that represent the underlying structure of data as topics. Hence, it has advantages on handling missing data and reasoning the data structure, which are desired properties in many computer vision tasks.

In many applications, not least in computer vision, the learning task is often to estimate a label from a piece of data. Hence, supervised topic models have drawn a lot of attention. Although several supervised topic models have been proposed [2, 8, 12, 27], very little work has been done to study the impact of supervision on the latent representation itself. In this paper, we will perform such a study, analysing the behaviour of one type of supervised topic model, Supervised Latent Dirichlet Allocation (SLDA) [2, 22], and propose a number of enhancements that could potentially improve the performance.

Supervised topic models are especially important for computer vision since classification is one of the most common tasks in this domain. Popular supervised models include: the above mentioned SLDA [2, 22], which models the joint likelihood of the class label and the observed data in a principled Bayesian framework; Labeled LDA (LLDA)[1] [8], which optimizes the hyperparameter for each class, but no direct dependence between the class label and the observed data is modeled; Discriminative LDA

This research has been supported by the Swedish Research Council (VR)

[1] LLDA indicates the model from [8] which is deigned for natural scene classification. The other popular model termed LLDA is from [17] and is designed for using multiple tags rather than class label.

(DiscLDA) [12], which models the conditional likelihood of the data on the class label through matrix transformation; Max-Entropy Discrimination LDA (MedLDA) [27, 28], which utilizes max-margin principle to learn the topic space using regularised Bayesian inference. Among these models, SLDA is the most popular one, since it is the most principled and straightforward Bayesian framework. Hence, we will focus on SLDA based topic models in this paper.

Very few studies have been done on understanding the behaviour of topic models, although various new topic models have been proposed every year. Semantic consistency of learned topics is studied [6] for classic unsupervised topic models which are pLSI [10], LDA [3] and Correlated Topic Model (CTM) [1]; and the behaviour of LDA with respect to the size of observed data (length of documents and number of documents) has been studied recently [19]. However, there is still a big gap in the understanding of the behaviour of supervised topic models. In computer vision, similar classification results have been achieved using standard LDA with a separate SVM for classification, as with SLDA [15], which raise the question of how effectively the topic space in an SLDA model is adapted to the class labels. How much impact the supervision has in the model has been discussed [18], but never been studied. In this paper, we will address this question and present analysis on the behaviour using SLDA and an adaptation, Power SLDA (P-SLDA), where the effect of the class label is boosted.

A latent representation that is able to capture the difference among data from different classes is the key to achieve good classification performance. The goal of using supervised topic models is to learn a better latent representation of the data that is suitable for the given task. Intuitively, only part of the information from the data is relevant for the classification task. For example, given the task to classify mugs from books, the shape of the object is relevant, which is the signal, and the pattern printed on the objects is not relevant, which is the noise. Classic topic models model the entire data together. Hence, the performance suffers when the data have low signal-noise ratio using classic topic models. Several works for different applications have considered this problem and allow the model to have different strategies to handle noise [11, 21, 25, 16]. Three of these, [11, 21, 16], are designed for specific (non-classification) tasks, while the fourth, [25], is heuristic in the sense that it introduces an entropic regularizer. In this paper, we propose two variations of SLDA which are probabilistically principled framework. To explore a better way to supervise topic models, these proposed models will be studied together with SLDA and P-SLDA on how they can influence the learning of topics.

We summarize our contributions as follows:

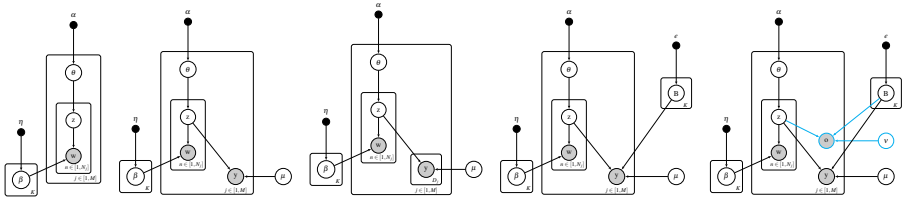
1. *A thorough analysis of the supervision effect of SLDA compared to LDA are presented.*
   Experimental results shows that the impact of supervision is limited on the learning of the latent space compared to the LDA due to the imbalance of the model.
2. *Power SLDA (P-SLDA) which maps the class label to higher dimension to boost supervision is contracted for further analysis on supervision behaviours.*
   Clear impact can be observed with boosted supervision, however, the benefit of supervision with P-SLDA is data dependent.
3. *Two novel factorized topic models are designed to learn better latent representation for classification tasks and provide better interpretation of topics.*

Experimental results show that these factorized models are able to factorize topics into signal and noise and are more robust compared to SLDA and P-SLDA.

The paper is organised as follows: all the models that are involved in the paper are described in Section 2; experimental evaluation and analysis are presented in Section 3; finally, we conclude the paper in Section 4.

2 Methods

In this section, we will firstly present all the models that are used in this paper. Then, we will briefly present the inference algorithms and classification methods. The derivation of inference algorithms and implementation details will be presented in the supplementary document.



(a) LDA (b) SLDA (c) P-SLDA (d) NUF-SLDA (e) NCF-SLDA

Fig. 1. Graphical representation of topic models studied in this paper. M indicates the number of documents; N_j indicates the number of words in the document j; K indicates the number of topics; α and β are hyper-parameters; w indicates the observed words; y indicates the class label;$\theta \sim Dir(\alpha)$ is the topic distribution for each document; $z \sim Mult(\theta)$ is the topic assignment for each word; $\beta \sim Dir(\eta)$ indicates the topics which are distributed over words.

2.1 Models

Topic models encode latent structure as topics, which assume that each piece of information is composed of latent topics. LDA [3], shown in Figure 1(a), is the cornerstone of topic models which was originally applied in information retrieval. LDA assumes a generative process where each document is modeled as a distribution over topics, and each topic is modeled as a distribution over words. LDA is the basic framework which has been evolved and applied for different tasks and all these models are called topic models. They can be applied on different types of data. For example, in computer vision, a document can be an image and a word can be a visual word from the bag-of-words representation. In this paper, we concentrate on supervised topic models that can be applied for classification tasks which is extremely important for computer vision applications. In this section, we will present all the models in the evolutionary order. Supervision is designed to be a part of the model for all the models but the first model, which learn the topics and classification parameters in separate stages.

LDAC LDA [3] is an unsupervised model. To perform classification tasks, the simplest way is to use the topics that are learned from LDA and apply an additional classifier on the topics. We call it LDA for Classification (LDAC) in this paper. Hence, LDAC is the same on modeling the topics as LDA. LDA/LDAC can be present as:

$$p(w,z,\theta,\beta|\alpha,\eta,\mu) = p(\beta|\eta)\prod_{j=1}^{M}\Big(p(\theta_j|\alpha)\prod_{n=1}^{N}\big(p(w_{jn}|z_{jn},\beta)p(z_{jn}|\theta_j)\big)\Big). \quad (1)$$

The graphical representation is shown in Figure 1 (a). A standard softmax regression is used for the classification tasks for a fair comparison in this paper. Note that the training of LDA and the softmax regression are done in separate steps. The label information is not involved in the learning of the topics.

SLDA For classification tasks, a topic representation that leads to better separation between classes is preferred. Hence, a supervised model is preferred for classification tasks. SLDA [2, 22], shown in Figure 1 (b), is the most straightforward and the most commonly used supervised topic modelling framework. Compared to LDA, the supervision is modelled as a response [2] to the topic assignments of each document. Hence, the topics are used to generate both the words and the label. SLDA models the words and the label jointly, which means that the inference will optimize the joint likelihood of the words and the labels. SLDA can be represented as:

$$p(w,z,\theta,\beta,y|\alpha,\eta,\mu) = p(\beta|\eta)\prod_{j=1}^{M}\Big(p(\theta_j|\alpha)p(y_j|z_{1:N},\mu)\prod_{n=1}^{N}\big(p(w_{jn}|z_{jn},\beta)p(z_{jn}|\theta_j)\big)\Big), \tag{2}$$

where y_j is a one-dimensional label and $p(y_j = l|z_j,\mu) = \frac{\exp(\mu_l^T(\bar{z}_j))}{\sum_{c=1}^{C}\exp(\mu_c^T(\bar{z}_j))}$, where $\bar{z}_j = \Big(\frac{1}{N}\sum_{n=1}^{N}\mathfrak{z}_{jn}\Big)$, in which $\mathfrak{z}_{jn}$ is the vector representation of the topic assignment indicator z_{jn}.

P-SLDA Topic assignments are used to explain both the words and the label using SLDA. However, for a document, the words are N_j-dimensional and the class label y_j is one-dimensional. It is not balanced between these two views due to different dimensionality. Hence, the supervision may not be sufficient in the model. Power-SLDA (P-SLDA), shown in Figure 1 (c), is a model that we construct to study the effectiveness of supervision with SLDA. It is a variation of SLDA. Compared to SLDA, in which the response is drawn only once for each document, P-SLDA draw the response D_j times. Hence, P-SLDA allows the label to be mapped to D_j dimensional. By varying D_j, we can study how much the supervision influence the learning of the topics. P-SLDA can be presented as:

$$p(w,z,\theta,\beta,y|\alpha,\eta,\mu) = p(\beta|\eta)\prod_{j=1}^{M}\Big(p(\theta_j|\alpha)p(y_j|z_{1:N},\mu)^{D_j}\prod_{n=1}^{N}\big(p(w_{jn}|z_{jn},\beta)p(z_{jn}|\theta_j)\big)\Big). \tag{3}$$

Comparing to SLDA, the difference lies in the power index D_j on $p(y_j|z_{1:N},\mu)$. Since documents may have different length, we define $D_j = \frac{N_j}{s}$, where s is the scaling parameter. When $s = N_j$, PSLDA becomes SLDA. When $s = 1$, the label is mapped to the same dimension as the words $D_j = N_j$.

NUF-SLDA and P-NUF-SLDA SLDA and P-SLDA model all the data together as many other topic modelling framework based on SLDA. However, the data are noisy, and the noise in the data may be inconsistent with the label which will cause poor performance when the data has low signal-noise ratio. The concept, which use factorized

[2] The response can be any type with generalized linear model [2]. In this paper, we mainly focus on the case when the response is the class label [22].

representation for information that can be shared among different views and information that cannot be shared between different views, has been applied to different frameworks with a long history [20, 4, 7]. We will adopt the same concept for supervised topic models. In NUF-SLDA, we assume that only part of the topics should be shared between the observed words and the label, which are used for generating the words and generating the label ; and the other part of the model is only used to model the rest of words which are not relevant for the classification task. We call the shared topics as signal topics and the ones not shared as noise topics. As the graphical representation of the model shown in Figure 1 (d), we introduce a signal-noise indicator B in the SLDA model which indicates whether the topic is used to model signal or noise, where $B \sim Bern(e)$. Comparing to SLDA, the main difference is that the class label y only respond to the topics which are indicated as signal (with $B_k = 1$). NUFSLDA can be represented as:

$$p(w,z,\theta,\beta,y|\alpha,\eta,\mu) = p(\beta|\eta)p(B|e)\prod_{j=1}^{M}\Big(p(\theta_j|\alpha)p(y_j|z_{1:N},\mu,B)\prod_{n=1}^{N}\big(p(w_{jn}|z_{jn},\beta)p(z_{jn}|\theta_j)\big)\Big). \tag{4}$$

Differently from SLDA, the softmax regression in the NUF-SLDA is defined by [3]

$$p(y_j = l|z_j,B,\mu) = \frac{\exp(\mu_l^T(\bar{z}_j \otimes B))}{\sum_{c=1}^{C}\exp(\mu_c^T(\bar{z}_j \otimes B))}. \tag{5}$$

Similarly to P-SLDA, P-NUF-SLDA with boosted supervision can be constructed as:

$$p(w,z,\theta,\beta,y|\alpha,\eta,\mu) = p(\beta|\eta)p(B|e)\prod_{j=1}^{M}\Big(p(\theta_j|\alpha)p(y_j|z_{1:N},\mu,B)^{D_j}\prod_{n=1}^{N}\big(p(w_{jn}|z_{jn},\beta)p(z_{jn}|\theta_j)\big)\Big). \tag{6}$$

NCF-SLDA and P-NCF-SLDA The previous model, NUF-SLDA, factorizes signal and noise. In this section, we adjust the model to constrain the noise to be *structured*, which share the same assumption as in[25]. Compared to NUF-SLDA, the key difference is that the noise part responds to a noise class. With this constraint, all the noise has the same label, hence, it is the structured noise. The graphic representation of NCF-SLDA is shown in Figure 1 (e). The noise response variable o is introduced, which is marked cyan in Figure 1 (e). NCF-SLDA can be represented as:

$$p(w,z,\theta,\beta,y|\alpha,\eta,\mu) = p(\beta|\eta)p(B|e)\prod_{j=1}^{M}\Big(p(\theta_j|\alpha)p(y_j|z_{1:N},\mu,B)p(o|z_{1:N},B,\nu)\prod_{n=1}^{N}\big(p(w_{jn}|z_{jn},\beta)p(z_{jn}|\theta_j)\big)\Big). \tag{7}$$

The additional noise response term is modeled as $p(o_j|z,B,\nu) = \frac{\exp(\sum_{k=1}^{K}\nu_k\bar{z}_{jk}(1-B_k))}{\exp(\sum_{k=1}^{K}\nu_k\bar{z}_{jk}(1-B_k))+1}$.

Similarly, P-NCF-SLDA can be presented as:

$$p(w,z,\theta,\beta,y|\alpha,\eta,\mu) = p(\beta|\eta)p(B|e)\prod_{j=1}^{M}\Big(p(\theta_j|\alpha)p(y_j|z_{1:N},\mu,B)^{D_j}p(o|z_{1:N},B,\nu)\prod_{n=1}^{N}\big(p(w_{jn}|z_{jn},\beta)p(z_{jn}|\theta_j)\big)\Big). \tag{8}$$

2.2 Inference

Variational inference and sampling based methods are the two main classes of methods that are generally used in the inference of topic models. Variational inference is known

[3] "$\otimes$" is used to indicate the element product

for its fast convergence and it is easy to adapt batch variational inference to an online setting [9, 23, 26]. In this work, we will use the standard batch mean field variational inference for all the models in this paper. Fully factorized variational distribution is used as [3, 22, 2]. Derivation details of variational inference for NUF-SLDA and NCF-SLDA are presented in the supplementary document.

2.3 Classification

For classification, we would like to estimate $p(y_j|z_j,\mu)$ or $p(y_j|z_j,B,\mu)$ for the test document j. The estimated label is the one with highest probability. In this case, the variational approximation for the true posterior is used. Hence, the prediction rule for LDAC, SLDA and PSLDA is:

$$\hat{y}_j = argmax_{l\in\{1,\ldots,C\}} \mathop{\mathbb{E}}_{q}[\mu_l^T \bar{z}] = argmax_{l\in\{1,\ldots,C\}} \mu_{y_j}^T \Big((\frac{1}{N}\sum_{n=1}^{N} \phi_{jn}) \Big), \tag{9}$$

and the prediction rule for both NUF-SLDA and NCF-SLDA is:

$$\hat{y}_j = argmax_{l\in\{1,\ldots,C\}} \mathop{\mathbb{E}}_{q}[\mu_l^T (\bar{z}\otimes B)] = argmax_{l\in\{1,\ldots,C\}} \mu_{y_j}^T \Big((\frac{1}{N}\sum_{n=1}^{N} \phi_{jn}) \otimes f \Big). \tag{10}$$

Note that μ is learned during the inference of the model and is able to affect the learning of topic assignment for all models but LDAC. In LDAC, μ is learned as a separate parameter where all the topics are learned using LDA.

3 Experiments

In this section, we will present our experimental results and discussion on these results. The experiments are carried out on three datasets: synthetic dataset[4], KTH video dataset and natural scene image dataset. The synthetic dataset is constructed to analysis the behaviour of the model in a controlled manner. The other two datasets are real world datasets, among which KTH dataset present a low signal- noise ratio case and the natural scene dataset present a high signal-noise ratio case. For each dataset, experiments are performed to evaluate two aspects: the effectiveness of the supervised topic models; and the performance of factorized topic models . The experimental results are ordered by datasets and the discussion will be presented in the end of this section.

3.1 Classification on Synthetic Dataset

We construct a synthetic dataset to study the behaviour of the models in a controlled manner. Figure 2 shows how synthetic data is generated. In the experiment, 200 documents for each class are generated for training and 40 documents for each class are generated for testing, which yields 1000 training and 200 testing documents in total.

[4] The synthetic dataset and our implementation for all the novel models will be published.

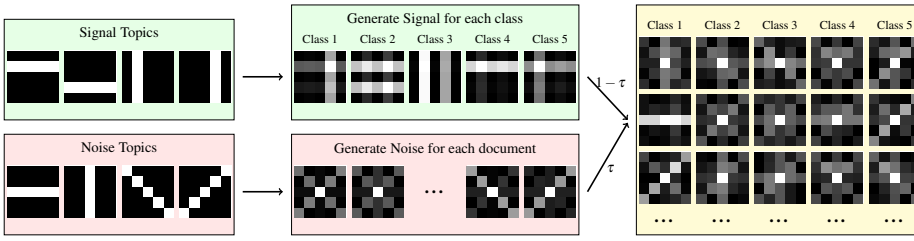
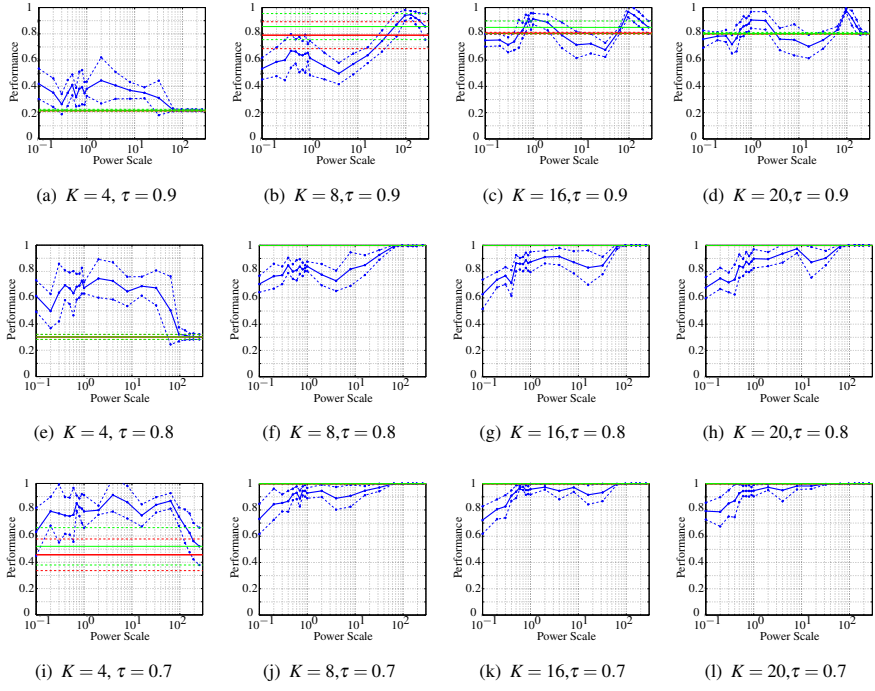


Fig. 2. The generation of the synthetic data. Eight topics are set first where four of them are defined as signal topics and the other four are defined as noise topics. Signal for five classes which are convex combination of the signal topics are generated. Then we add noise, which are random convex combination of noise topics, to generate the dataset. For each document/image, the noise is generated independently. The noise level is control by the parameter τ. The final document/image is generated by $(1-\tau) \times Signal + \tau \times Noise$. The noise level $\tau = 0.8$ is used for the example documents above.

Supervision effectiveness Firstly, we compare the classification performance on synthetic datasets using LDAC, SLDA and P-SLDA with different power scales s. Figure 3 shows the performance of these models with different number of topics and different noise levels. Hyperparameters $\alpha = 0.5$ and $\eta = 0.1$ are used in these experiments. All experiments are run 9 times with different random seeds for initialization. The mean and standard deviation are reported. LDAC and SLDA have similar performance over all different settings, although better performance is expected from SLDA over LDAC since the learning process is supervised. By boosting the supervision using P-SLDA, clear change of performance can be observed. This shows that the supervision is not effective using SLDA on learning of the latent space. Figure 3(a) (e) (i) show the performance with $K = 4$ with different noise levels. We can see that the improvement on classification results is significant through all different noise levels when the number of topics is small. Figure 3 (a) (b), (c), (d), show that the the performance can be clearly improved with all different number of topics when the noise level is high ($\tau = 0.9$). However, the lower-right plots in Figure 3 show different levels of drop in performance where the number of topics are large and the data is less noisy.

To further understand the phenomenon in Figure 3, we visualize the topics that are learned with different models. We present two typical cases using the noise level $\tau = 0.8$ to analyze the reason for the performance change. Figure 4 shows the topics learned with different models when $K = 4$, which corresponds to Figure 3 (e). As expected, LDA/LDAC only learn the topics to represent noise, since noise is dominant in the data. Topics learned using SLDA are almost the same as the topics that is learned with LDA, which shows the way to model the class label in SLDA is not effective to supervise the model to learn a better representation for classification. This also explains that SLDA has similar performance as LDAC, since the learned latent structures are similar. By mapping the class label to D dimension using P-SLDA, we can observe that the learned topics start to differ from LDA. As shown in Figure 4(c) and (d), the larger the D is, the more impact the supervision has on the model. Since the topics are used to explain both signal and noise in P-SLDA and there are limited topics, the learned topics become mixed with signal and noise even with boosted supervision. However, P-SLDA is still able to catch the signal compared to LDA, hence, the performance is improved in this case.



(a) $K = 4,\ \tau = 0.9$ (b) $K = 8, \tau = 0.9$ (c) $K = 16, \tau = 0.9$ (d) $K = 20, \tau = 0.9$

(e) $K = 4,\ \tau = 0.8$ (f) $K = 8, \tau = 0.8$ (g) $K = 16, \tau = 0.8$ (h) $K = 20, \tau = 0.8$

(i) $K = 4,\ \tau = 0.7$ (j) $K = 8, \tau = 0.7$ (k) $K = 16, \tau = 0.7$ (l) $K = 20, \tau = 0.7$

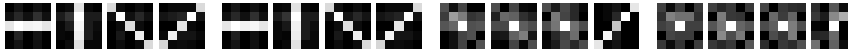
Fig. 3. Performance Evaluation for LDAC, SLDA and P-SLDA with difference power scale s under different number of topics K and different noise level τ. All experiments are run 9 times over different random seeding. The mean is represented using solid line and the standard deviation is represented using dashed line. The blue curve in these plots shows the performance of P-SLDA with different label dimension D. The x axis is the supervision power scale s which is plotted in the log scale ranging from 0.1 to 256, which indicates that D range from 2560 to 1 from left to right. While $s = N = 256$, P-SLDA becomes SLDA. The performance of SLDA is marked with green dot and the performance of LDAC is plotted in red.

Figure 5 shows the topics learned with different models when $K = 16$, which correspond to Figure 3(g). Topics learned with SLDA and LDA appear to be the same as in the previous case. However, both signal and noise are captured when the topics space is large through all the models, which explains the good performance by both SLDA and LDAC in Figure 3 (g). By boosting the supervision, the learned topics start to change. However, the changes are minor compared to the previous case when $K = 4$. When we boost the supervision in an extreme case, $D = 10N$, the learned topics start to break into small fragments, which is due to over-fitting and causes the performance drop.

To further study the supervision effectiveness, we observe the likelihood behaviour during the learning process. Recall the Evidence Lower BOund (ELBO) of SLDA is

$$\begin{aligned}\mathscr{L}_{SLDA} = &\mathop{\mathbb{E}}_{q}[\log p(w|z,\beta)] + \mathop{\mathbb{E}}_{q}[\log p(z|\theta)] + \mathop{\mathbb{E}}_{q}[\log p(\theta|\alpha)] + \mathop{\mathbb{E}}_{q}[\log p(\beta|\eta)] \\ &+ \mathop{\mathbb{E}}_{q}[\log p(y|z,\mu)] - \mathop{\mathbb{E}}_{q}[\log q(\beta)] - \mathop{\mathbb{E}}_{q}[\log q(z)] - \mathop{\mathbb{E}}_{q}[\log q(\theta)].\end{aligned} \tag{11}$$

The same form holds for P-SLDA. The difference lies in the dimension of y. In SLDA, y is 1-dimensional, whereas in P-SLDA, y is D-dimensional. The value of each item over each iteration is plotted in Figure 6 for different cases. It is clear that $\mathbb{E}_q[\log p(z|\theta)]$ becomes higher and $\mathbb{E}_q[\log q(z)]$ becomes lower while the supervision is



(a) LDA (b) SLDA (c) P-SLDA D=8 (d) P-SLDA D: 256

Fig. 4. The learned topics ($\hat{\beta}$) from different models with number of topics $K = 4$ and $\tau = 0.8$. (a) The topics learned using LDA (b) The topics learned with SLDA (c) The topics learned with PSLDA with power scale $s = 32$ which is $D = 8$ (d) The topics learned with P-SLDA with power scale $s = 1$ which is $D = N = 256$

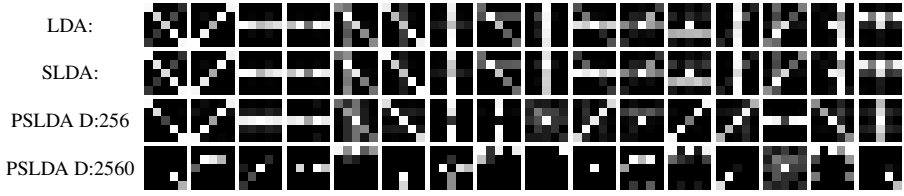


Fig. 5. The learned topics ($\hat{\beta}$) from different models with number of topics $K = 16$ and $\tau = 0.8$.

boosted. $\mathbb{E}_q[\log p(w|z,\beta)]$ drops slightly as well with boosted supervision. The topic assignments z are used to explain both the words and the label in a document. By boosting the supervision, the label will get better explained by the topic assignments, however, the cost is that the words get less well explained by the topic assignments, which is confirmed by the drop of $\mathbb{E}_q[\log p(w|z,\beta)]$. The drop of the entropy term $\mathbb{E}_q[\log q(z)]$ shows that the topic assignment distribution becomes more sparse with boosted supervision. This is caused by the fact that different topics are used to explain different classes and topics become less shared among different classes. This shows the tradeoff between the use of latent space to explain the words and the label, when data are noisy.

3.2 Factorized Models

In this part, we evaluate factorized topic models NUF-SLDA, P-NUF-SLDA, NCF-SLDA and P-NCF-SLDA. The same experimental setting is used as in the previous section. Due to space limitation, we only show results with $\tau = 0.8$ in this part. Firstly, whether the factorized model is able to learn the correct factorization is evaluated. Figure 7 shows the learned topics with $K = 8$ using NUF-SLDA and NCF-LDA. Both models are able to correctly factorize the topics.

The performances of these factorized models compared with P-SLDA are shown in Figure 8. $e = 0.2$ is used through all these experiments. We can see that P-NUF-SLDA is more robust compared to P-SLDA and P-NCF-SLDA, when there is sufficient amount of topics. P-SLDA could achieve better performance when the number of topics is small.Because factorized model separate the topic space to signal part and noise part. When the number of topics are not sufficient, factorizing the model makes the number of topics describing the signal even smaller.

3.3 Video Action Classification

We use threes action classes: boxing, hand clapping and hand waving from KTH action dataset [13, 11] for the action classification experiment. Intuitively, only the human

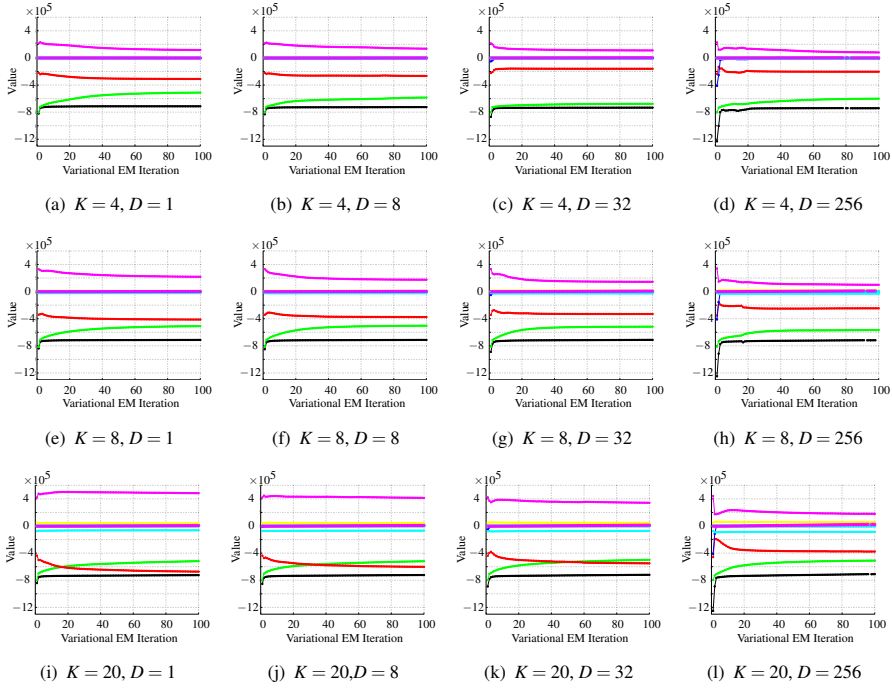


Fig. 6. Likelihood Analysis for supervision effectiveness. Legend: $-*-$ $\mathscr{L}_{P-SLDA}$; $-*-$ $\mathbb{E}_q[\log p(w|z,\beta)]$; $-*-$ $\mathbb{E}_q[\log p(z|\theta)]$; $-*-$ $\mathbb{E}_q[\log p(\theta|\alpha)]$; $-*-$ $\mathbb{E}_q[\log p(y|z,\mu)]$; $-o-$ $\mathbb{E}_q[\log p(\beta|\eta)]$; $-*-$ $-\mathbb{E}_q[\log q(\theta)]$;$-*-$ $-\mathbb{E}_q[\log q(z)]$; $-o-$ $\mathbb{E}_q[\log q(\beta)]$

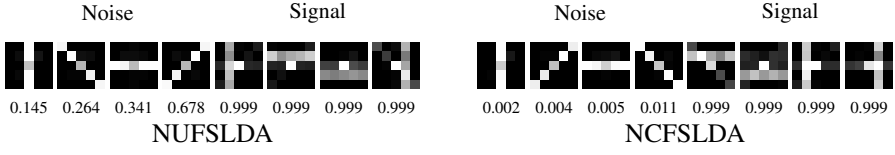
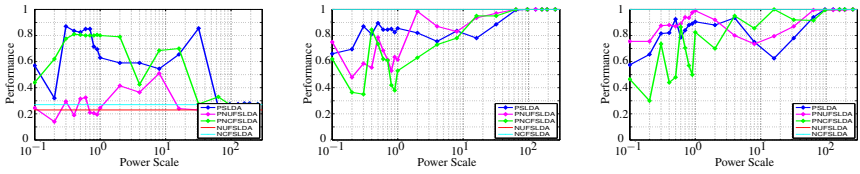


Fig. 7. Learned topics by NUF-SLDA and NCF-SLDA with $\tau = 0.8$. $\mathbb{E}_q[B]$ is marked below each topic which indicates the confidence for the topic to be signal.

movement is the signal and all the variances in the background and video shooting settings are noise. Hence, this is a real world dataset where signal-noise ratio is low. There are around 100 video clips for each action of which 80% are randomly selected for training and the remaining 20% are for testing. Bag-of-STIP [13] is use to represente visual words in each video clip which is treated as a document. $\alpha = 0.1$, $\eta = 0.1$ are used through all the experiments using this dataset.

Supervision Effectiveness Figure 9 shows the performance of LDAC, SLDA and P-SLDA with different power scale s, which are consistent with the one using synthetic dataset. When the number of topics is small, boosting the supervision can improve the performance significantly as in Figure 9 (a), (b), (c). When the number of topics is more than sufficient, boosting the supervision may disturb the classification performance. The result is not only interesting for understanding the supervised effectiveness, it is



(a) $K = 4$ (b) $K = 8$ (c) $K = 16$

Fig. 8. Performance comparison of factorized topic models for synthtic dataset with $\tau = 0.8$.

also significant from a application perspective. The number of topics is essential for computational complexity in the inference. By boosting the supervision, using a small number of topics will be able to achieve similar level of performance as using a large number of topics, but the computation time will be significantly reduced.

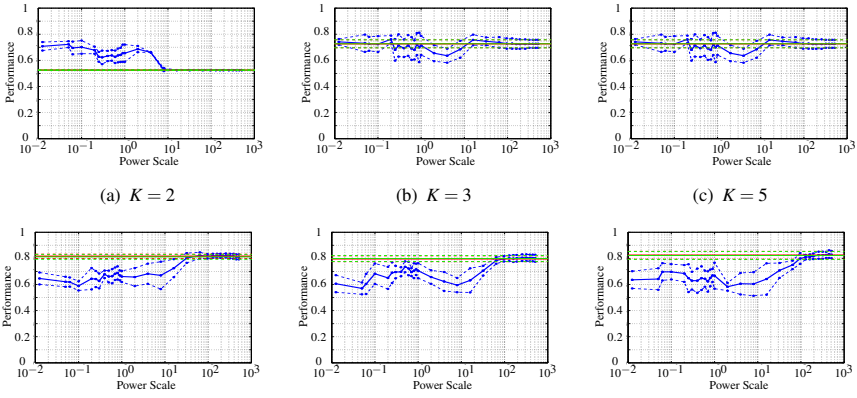
Factorized Models Performances of factorized models are evaluated and compared in Figure 10. This dataset has low signal-noise ratio, hence $e = 0.3$ is used through all the experiments in this part. We can see that P-NUF-SLDA is more robust to the change of supervision level and it has potential to overperform P-SLDA as in Figure 10 (c) (d). However, since it uses less topics to models the signal. The performance maybe worse when the number of topics is not enough to factorize. NCF-SLDA is in general not as robust with boosted supervision, we believe that it is caused by that the structured noise assumption is too strong and the boosting effect is doubled in P-NCF-SLDA with the additional noise label.

3.4 Natural Scene Classification

Four classes of natural scene images are used in this experiment as [8, 26]. Intuitively, all the information from natural scene is useful to judge the scene category. Hence, this is a real world dataset where the signal-noise ratio is high. There are more than 300 images per class of which 80% of the data are randomly selected for training and the remaining 20% for testing. Bag-of-SIFT [14] is used to represent visual words in each image which is treated as a document. $\alpha = 0.1$, $\eta = 0.1$ are used through all the experiments in this section.

Supervision Effectiveness Figure 11 shows the performance of LDAC, SLDA and P-SLDA with different power scale s for natural scene classification. LDAC and SLDA have the same performance with different number of topics as previous experiments. Differently from the previous experiments, the classification performance does not change as much by boosting the supervision. The performance gets worse when the supervision is boosted too much, which is caused by overfitting. This shows that when the signal-noise ratio is high, the optimum for both unsupervised model and supervised model are similar, since the label is consistent with the words. A little improvement can still be observed by boosting the supervision as in Figure 11 (a), since the data is not ideal.

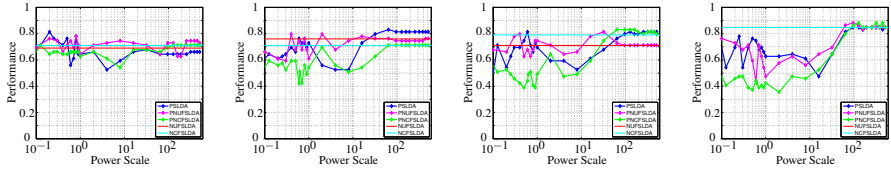
Factorized Models Figure 12 shows the performance of factorized models on the natural scene classification task. Consistent with the previous experiments, the performance



(a) $K = 2$ (b) $K = 3$ (c) $K = 5$

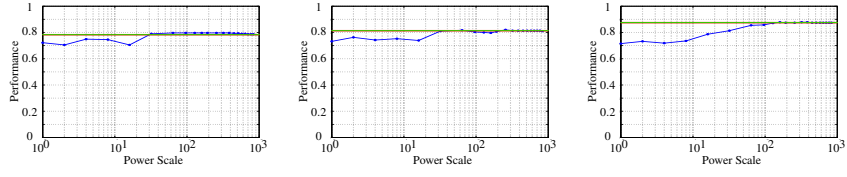
(d) $K = 10$ (e) $K = 20$ (f) $K = 40$

Fig. 9. Action classification performance. The x axis is the power scale s and the y axis shows the classification performance. All the experiments are repeated 8 times with different random seeds for initialization. The mean, solid line, and the standard deviation, dashed line, are shown in the plot. P-SLDA with different power scale is plotted with the blue curve. LDAC is marked with the green line and SLDA is marked with the red line.



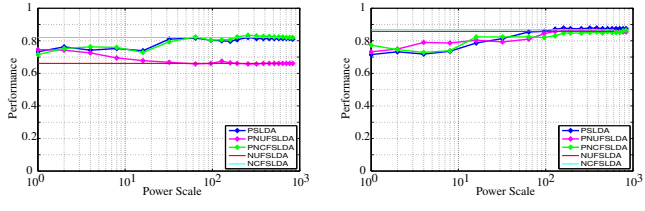
(a) $K = 5$ (b) $K = 10$ (c) $K = 20$ (d) $K = 30$

Fig. 10. Performance comparison of factorized topic models.



(a) $K = 5$ (b) $K = 10$ (c) $K = 30$

Fig. 11. Scene classification performance. P-SLDA with different power scale is plotted with the blue curve. LDAC is marked with the green line and SLDA is marked with the red line.



(a) $K = 10$ (b) $K = 30$

Fig. 12. Performance comparison of factorized topic models for natural scene classification. The x axis is the power scale s (ranging from 10^0 to 10^3) and the y axis shows the performance (ranging from 0 to 1).

of P-NUF-SLDA is more robust with respect to the boosting of the supervision and shows better performance when the label is mapped to high dimension compared to P-SLDA. P-NUF-SLDA does not perform as good as P-SLDA when the number of topics is small. Because with factorisation, only around half of the topics are used to model the signal which is not enough when the total number of the topics is small. P-NCF-SLDA perform on par with P-SLDA. All the models are more robust with this dataset since the data has high signal-noise ratio.

3.5 Discussion

To sum up the experiments with three different datasets, we will present a discussion in three points. Firstly, all the experiments above show that SLDA and LDAC have similar performance through all different settings. Further analysis with the synthetic data shows that the topics learned by SLDA and LDA are similar. This confirms that supervision on LDA using SLDA is not effective on learning of topics. Secondly, P-SLDA is able to boost the supervision, which makes the supervision affect the learning of topics. Experiments on different settings show that whether boosting the supervision can be beneficial is highly dependent on the data and the parameter setting. When the data is noisy (low signal-noise ratio), as in the first two experiments, boosting supervision is able to increase the performance, especially when the number of topics is small. When the data is informative (high signal-noise ratio), boosting the supervision is not able to clearly affect the classification performance since the label and words information are consistent. Over boosting the supervision can harm the performance since it makes the model biased towards the label and causes overfitting. Thirdly, factorized models are able to recognize the signal topics and noise topics correctly, which improves the interpratation of the learned topics. They also have more robust performance with the boosting of the supervision.

4 Conclusions

In this paper, we have presented a thorough study on the behaviour of supervision on topic models, which fills the gap in the understanding of supervised topic models; and we have proposed two types of alternative factorized supervised topic models which improve the interpretation of topics and enhance the model performance. Variational inference has been used and fully derived for the proposed models. All the models have been evaluated with both synthetic data and real world data. We conclude in the study that: supervision is not effective using SLDA on the learning of the topics; balancing the model using P-SLDA can boost the supervision, which provide further improvements in case of noisy data; factorized models can increase the performance robustness.

We will continue our research in two directions. Firstly, we will analyze and compare a wider range of supervised topic models, such as DiscLDA [12] and MedLDA [27, 28], to have a deeper insight on the behaviours of all different supervised topic models. Secondly, we will continue working on factorized topic models, since most models can not deal with noise sufficiently well. We will both apply the factorization on different supervised topic modeling framework and use more effective factorization prior.

References

1. D. M. Blei and J. Lafferty. Correlated topic models. In *NIPS*, 2006.

2. D. M. Blei and J. D. McAuliffe. Supervised topic models, *arxiv:1003.0783*, 2010.
3. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
4. M. W. Browne. The maximum-likelihood solution in inter-battery factor analysis. *British Journal of Mathematical and Statistical Psychology*, 32(1):75–86, August 2011.
5. L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV*, 2007.
6. J. Chang, J.Boyd-Graber, C. Wang, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, 2009.
7. A. Damianou, H. C. E kand M. Titsias, and N. D. Lawrence. Manifold Relevance Determination. In *ICML*, pages 145–152, 2012.
8. L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.
9. M. D. Hoffman, D. M. Blei, and F. Bach. Online learning for latent Dirichlet allocation. In *NIPS*, 2010.
10. T. Hofmann. Probabilistic latent semantic analysis. In *ACM SIGIR*, 1999.
11. T. M. Hospedales, S. G. Gong, and T. Xiang. Learning tags from unsegemented videos of multiple human actions. 2011.
12. S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*, 2008.
13. I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *ECCV*, 2004.
14. D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.
15. Z. Niu, G. Hua, X. Gao, and Q. Tian. Semi-supervised relational topic model for weakly annotated image recognition in social media. In *CVPR*, 2014.
16. M. Rabinovich and D. M. Blei. The inverse regression topic model. In *ICML*, 2014.
17. D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Conference on Empirical Methods in Natural Language Processing*, 2009.
18. N. Rasiwasia and N. Vasconcelos. Latent dirichlet allocation models for image classification. *PAMI*, 35(11):2665–2679, 2013.
19. J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang. Understanding the limiting factors of topic modeling via posteriror contraction analysis. In *ICML*, 2014.
20. L. R. Tucker. An Inter-Battery Method of Factory Analysis. *Psychometrika*, 23, June 1958.
21. C. Wang and D. Blei. Collaborative topic modeling for recommending scientific articles. In *ACM SIGKDD*, 2011.
22. C. Wang, D. M. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.
23. C. Wang, J. Paisley, and D. Blei. Online variational inference for the hierarchical Dirichlet process. In *AISTATS*, 2011.
24. D. Weinshall, G. Levi, and D. Hanukaev. Latent Dirichlet allocation topic model with soft assignment of descriptors to words. In *ICML*, 2013.
25. C. Zhang, C. H. Ek, A. Damianou, and H. Kjellström. Factorized topic models. In *International Conference on Learning Representations*, 2013.
26. C. Zhang, C. H. Ek, X. Gratal, F. T. Pokorny, and H. Kjellström. Supervised hierarchical Dirichlet processes with variational inference. In *ICCV workshop on Inference for probabilistic graphical models*, 2013.
27. J. Zhu, A. Ahmed, and E. P. Xing. Medlda: Maximum margin supervised topic models for regression and classification. In *ICML*, 2009.
28. J. Zhu, N. Chen, H. Perkins, and B. Zhang. Gibbs max-margin supervised topic models with fast sampling algorithms. In *ICML*, 2013.