

---

# Factorized Topic Models

---

**Cheng Zhang**  
CVAP/CAS, KTH  
Stockholm, Sweden  
chengz@kth.se

**Carl Henrik Ek**  
CVAP/CAS, KTH  
Stockholm, Sweden  
chek@kth.se

**Andreas Damianou**  
University of Sheffield  
Sheffield, UK  
andreas.damianou@sheffield.ac.uk

**Hedvig Kjellström**  
CVAP/CAS, KTH  
Stockholm, Sweden  
hedvig@kth.se

## Abstract

In this paper we present a modification to a latent topic model, which makes the model exploit supervision to produce a factorized representation of the observed data. The structured parameterization separately encodes variance that is shared between classes from variance that is private to each class by the introduction of a new prior over the topic space. The approach allows for a more efficient inference and provides an intuitive interpretation of the data in terms of an informative signal together with structured noise. The factorized representation is shown to enhance inference performance for image, text, and video classification.

## 1 Introduction

Representing data in terms of latent variables is an important tool in many applications. A generative latent variable model provides a parameterization that encodes the variations in the observed data, relating them to an underlying representation, e.g., a set of classes, using some kind of mapping. It is important to note that any modeling task is inherently ill-conditioned as there exists an infinite number of combinations of mappings and parameterizations that could have generated the data. To that end, we choose different models, based on different assumptions and preferences that will induce different representations, motivated by how well they fit the data and for what purpose we wish to use the representation.

Inference in generative models meets difficulties if the variations in the observed data are not representative of the variations in the underlying state to be inferred. As an example, consider a visual animal classifier, trained with, e.g., SIFT [20] features extracted from training images of horses, cows and cats with a variation of fur texture. The task is now to classify an image of a spotted horse. Based on the features, which will mostly pick up the fur texture, the classifier will be unsure of the class, since there are spotted horses, cows and cats in the training data. The core of the problem is that fur texture is a weak cue to animal class given this data: Horses, cows and cats can all be red, spotted, brown, black and grey. Shape is on the other hand a strong cue to distinguish between these classes. However, the visual features will mostly capture texture information – the shape information (signal) is “hidden” among the significantly richer texture information (structured noise) making up the dominant part of the variation in the data.

In this paper we address this issue by explicitly factorizing the data into a structured noise part, whose variations are shared between all classes, and a signal part, whose variations are characteristic of a certain class. For our purposes, it is very useful to think about data as composed of *topics*. Probabilistic topic models [23, 11, 4, 2] model a data example as a collection of *words* (in the case of images, *visual words*), each sampled from a latent distribution of *topics*. The topics can be thought of as different aspects of the data – a topic model trained with the data in our animal example above might model one topic for shape and another for fur texture, and a certain data instance is modeled as a combination of a certain shape and a certain texture.

Our approach is to *encourage the topics to assume either a very high correlation or a very low correlation with class*. The class can then be inferred using only the class-specific topics, while the

---

This research has been supported by the EU through TOMSY, IST-FP7- Collaborative Project-270436, and the Swedish Research Council (VR).

shared topics are used to *explain away* the aspects of the data that are not interesting to this particular inference problem. We present a variant of a Latent Dirichlet Allocation (LDA) [2] model which is able to model the signal and structured noise separately from the data. This new model is trained using a factorizing prior, which partitions the topic space into a private signal part and a shared noise part. The model is described in Section 3.

Experiments in Section 4 show that the proposed model outperforms both the standard LDA and a supervised variant, SLDA [3], on classification of images, text, and video. Furthermore, the explicit noise model increases the sparsity of the topic representation. This is encouraging for two reasons: firstly, it indicates that the factorized LDA model is a better model of class compared to the unrestricted LDA; enabling better performance on any inference or data synthesis task. Secondly, it enables a more economical data representation in terms of storage and computation; crucial for applications with very large data sets. The factorization method can be applied to other topic models as well, and the sparse factorized topic representation is beneficial not only for classification, as shown here, but also for synthesis [5], ambiguity modeling [7], and domain transfer [21].

## 2 Related Work

In this section we will create a context for the model that we are about to propose by relating it to factorized latent variable models in general and topic models in specific. Providing a complete review of either is beyond the scope of this paper, why here we will focus on only the most relevant subset of work needed to motivate the model.

The motivation for learning a latent variable model is to exploit the structure of the new representation to perform tasks such as synthesis or prediction of novel data, or to ease an association task such as classification. For continuous observations, several classic algorithms such as Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) can be interpreted as latent variable models [1, 16, 17, 27]. Another modeling scenario is when observations are provided in the form of collections of discrete entities. An example is text data where a document consists of a collection of words. One approach to encode such data is using a latent representation that groups words in terms of topics. Several approaches for automatically learning topics from data have been suggested in the literature. A first proposal of a generative topic model was Probabilistic Latent Semantic Indexing (pLSI) [11]. The model represents each document as a mixture of topics. The next important development in terms of a Bayesian version of pLSI by adding a prior to the mixture weights. This was done by the adaptation of a Dirichlet layer and referred to as Latent Dirichlet Allocation (LDA) [4].

Central to the work presented in this paper is a specific latent structure simultaneously proposed by several authors [7, 13, 14, 18]. Given multiple observation modalities of a single underlying state, the purpose of these models is to learn a representation that separately encodes the modality-independent variance from the modality-dependent. The latent representation is factorized such that the modality-independent and modality-dependent are encoded in separate subspaces [6]. This factorization has an intuitive interpretation in that the private space encodes variations that exists in only one modality and does therefore encode variations representing the ambiguities between the modalities [7].

In this paper we will exploit a similar type of factorization within a topic model, but instead of exploiting correlations between observation modalities, we employ a single observation modality and a class label associated with each observation. In specific, our approach will encourage a factorization relating to class, such that the topics will be split into those encoding within-class variations from those that encode between-class variations. Such a factorization becomes interesting for inferring the class label from unseen data; the class-shared topics can be considered as representing “structured noise” while only the private class topics contain the relevant for class inference.

However, it is not easy to directly transfer the above factorization, formulated *between* modalities and described for continuous data, to topic models, which are inherently discrete. Results have been presented [12, 25, 28] for the case of two conditionally independent observation modalities, addressing the image and text cross-modal multimedia retrieval problem with topic representation. In [12] a model that can be seen as a Markov random field of LDA topic models is presented. The topic distribution of each topic model affect the underlying topic spaces of other topic models, connected to that model through the Markov random field. Further, in [25] CCA is applied to the topic space of the text data, which in turn has been learned from LDA and the image feature space.

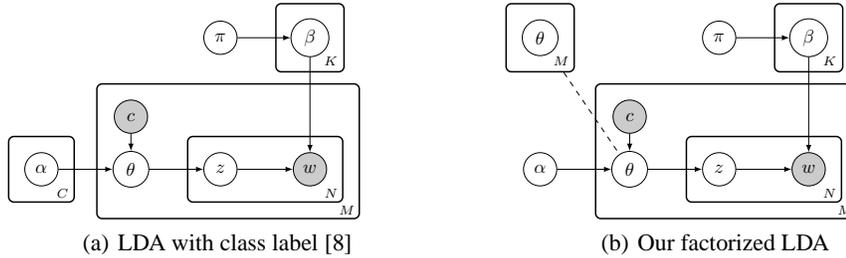


Figure 1: Graphic representation of LDA structures. The notation in (b) is adopted from Jia et al. [12].

LDA and CCA are used as two separate steps. Differently, [28] instead use a Hierarchical Dirichlet Process (HDP) based method which has a complexity selection property. It takes the topics that only describe variance in only one modality as the private space, which explains away the information that cannot be matched between different modalities. This is an extension of [22] to multi-modalities, hence it can not be generalized to other topic models, such as LDA or pLSI and it can not be used to model the private and shared information with only one modality.

Differently from [12, 28, 25], which need to model the shared topics and private topics in the joint topic space across different observation modalities, our factorization takes place over one modality across different classes, where the structured noise is modeled in the class-shared topics and the signal is modeled in class-private topics. Furthermore, and importantly, our approach is flexible and can be easily transferred to any type of topic model. Our choice of LDA stems from the fact that it has previously been successfully applied for a large range of data and has desirable sparsity properties that makes for an efficient model.

Topic models, and the LDA model in specific, are motivated by the benefit of representations that are sparse in terms of the distribution of topics for each document. In addition to this, the model we are about to present aims to encourage a specific structure of the topic themselves. This notion is not new and have been proposed by several other authors. In [9] the topics are represented as combinations of a small number of latent components as such leading to a more compact model. In [29] the each topic is constrained by the words in the vocabulary. However, none of these models aim to learn a topic structure that is related to class.

### 3 Model

As described in the introduction, we add factorization to a model that describes variations of data in terms of a set of latent topics. We seek a structured representation that encodes topics containing within-class, or class private, variations separately from those containing variations that are shared between the classes. We apply our factorization framework to an adaptation of LDA, which incorporates additional class information to recover such a factorized latent space. In this section, the traditional LDA model [4] is first revisited, followed by the description of our factorized topic model.

#### 3.1 LDA Revisited

Formally a document  $\mathbf{w}$  consist of a collection of words  $\mathbf{w} = [w_1, \dots, w_N]$  from a vocabulary indexed by  $\{1, \dots, V\}$ . Within a topic model each document of  $N$  words is described as a mixture of  $K$  topics such that each word is associated with a specific topic:  $\mathbf{z} = [z_1, \dots, z_N]$ , where  $z_n \in \{1, \dots, K\}$ . The mixture is defined as

$$p(\mathbf{w}|\mathbf{z}, \beta) = \prod_{n=1}^N \sum_{k=1}^K p(w_n|z_n, \beta_k) \quad (1)$$

where  $\beta_k$  is the distribution over the vocabulary for topic  $k$ . The novelty, and the reason for the success, of the LDA model is how the topics  $\mathbf{z}$  and the topic vocabulary  $\beta$  are constructed within the framework. The underpinning intuition is that the topics should present a compact representation with  $K \ll N$ , and that the structure of the topics should be sparse such to achieve a robust and interpretable model. Assuming the topics  $\mathbf{z}$  to be governed by a multinomial distribution,  $\mathbf{z} \sim \text{Multi}(\mathbf{z}|\theta)$ , sparsity can be achieved by choosing the parameters  $\theta$  as governed by a Dirichlet distribution,  $\theta \sim \text{Dir}(\theta|\alpha)$ . By the same motivation a Dirichlet prior is placed over the topic-vocabulary distribution  $\beta \sim \text{Dir}(\beta|\pi)$ . As the Dirichlet is conjugate to the multinomial distribution,

the marginal likelihood can be reached analytically by combining the likelihood with the prior and performing the integration,

$$p(\mathbf{w}|\alpha, \pi) = \int p(\boldsymbol{\theta}|\alpha) \left( \prod_{n=1}^N \left[ \sum_{z_n} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}_{z_n}) \right] p(\boldsymbol{\beta}|\pi) \right) d\boldsymbol{\theta} \quad (2)$$

from which the parameters of the model can be learned.

One way of incorporating class information within the LDA framework was suggested in [8] where the use of a class dependent topic distribution was proposed. This was implemented by using the class variable  $c$  as a “switch”;  $p(\boldsymbol{\theta}|\alpha, c) = \prod_{j=1}^C \text{Dir}(\boldsymbol{\theta}|\alpha_j)^{\delta_{cj}}$  where  $\delta_{ij}$  is the Kronecker delta function. Using this model the class can be inferred for a new document  $\mathbf{w}^*$  through a maximum likelihood procedure  $\hat{c} = \arg \max_c p(\mathbf{w}^*|\alpha, \pi, c)$  [8].

In this paper we take inspiration from the work presented in [8]. However, we choose to incorporate the class information in a slightly different manner. In specific, we use a factorizing prior over the topic distribution, which firstly encourages sparsity, and secondly introduces a preference for a class conditioned structure, such that separate topics encode within-class variations and between-class variations in the data. Thus, the model we will propose have a stronger class dependency compared to [8]. We will now proceed to describe and motivate the relevance of this class dependency.

### 3.2 Factorized Topic Model

As motivated in Section 1, our idea is to separate the topic space into two parts, where the class-private part explains the class-dependent information (signal) and the shared part explains the class-independent information (structured noise). To achieve this we introduce an additional prior  $p(\boldsymbol{\theta})$  to the model presented in [8]. This will encourage a factorized structure such that the  $K$  topics can be “softly” split into  $K_p$  class-private topics and  $K_s$  shared topics where  $K_p + K_s = K$ . The advantage of such a structured topic space is that it will be more compact than a regular model; all aspects of the data that correlate with class will be pushed into the class-private part of the topic space. Since the other, class-shared, part of the topic space will then only contain noise, the class of a new document  $\mathbf{w}^*$  will in effect be inferred using only the class-private part. Further, in our model, we will use the same sparsity prior  $\alpha$  over the topics for all classes. This removes the additional flexibility of allowing a different topic sparsity for each class – which can be relevant in certain special cases — but the gain is a more robust model with fewer free parameters, requiring less training data.

In the following, let  $\boldsymbol{\theta}^{\text{class}}$  be the topic distributions of all classes, obtained by marginalizing  $\theta$  over class. Its rows are defined as  $\boldsymbol{\theta}_c^{\text{class}} \propto \sum_{m=1}^M \boldsymbol{\theta}_m \delta_{c_m c}$ , where  $c_m$  indicates the class label of the  $m^{\text{th}}$  document and  $\delta_{ij}$  the Kronecker delta function. Examples of  $\boldsymbol{\theta}^{\text{class}}$  distributions can be seen in Figures 3, 4, 5, and 6.

Intuitively, the private topics would concentrate to a certain class in  $\boldsymbol{\theta}^{\text{class}}$ , while the shared topics would be more spread among all classes (more uniformly distributed over a column in  $\boldsymbol{\theta}^{\text{class}}$ ). Information entropy, widely used in different fields [19, 24], provides a good measurement of this property. In this case, we employ an entropy-like measure  $H(k)$  over class for each topic  $k$ :

$$H(k) = \frac{1}{\log(1/C)} \sum_{c=1}^C \frac{\theta_{c,k}^{\text{class}}}{\sum_{\xi=1}^C \theta_{\xi,k}^{\text{class}}} \log\left(\frac{\theta_{c,k}^{\text{class}}}{\sum_{\xi=1}^C \theta_{\xi,k}^{\text{class}}}\right) \quad (3)$$

where  $\theta_{c,k}^{\text{class}}$  is the element in row  $c$  and column  $k$  of  $\boldsymbol{\theta}^{\text{class}}$ .  $H(k) \in [0, 1]$ , 0 if all the probability in the topic  $k$  is concentrated to one class, 1 if all classes are equally probable to contain the topic  $k$ .

To split the topics into a private and shared part, we wish the prior  $p(\boldsymbol{\theta}|\kappa)$  to encourage topics  $k$  to either have a low  $H(k)$  (be very class-specific) or high  $H(k)$  (be very class-unspecific). Hence, we introduce a function as:

$$A(k) = H(k)^2 - H(k) + 1. \quad (4)$$

The prior is defined as:

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^K A(k). \quad (5)$$

This prior thus treats each column of  $\boldsymbol{\theta}^{\text{class}}$  independently. With the additional prior (Figure 1(b)), the generative model becomes:

$$p(\mathbf{w}|\alpha, \pi, c) = \int p(\boldsymbol{\theta}|\alpha, c)p(\boldsymbol{\theta}) \left( \prod_{n=1}^N \left[ \sum_{z_n} p(z_n|\boldsymbol{\theta})p(w_n|z_n, \boldsymbol{\beta}) \right] p(\boldsymbol{\beta}|\pi) \right) d\boldsymbol{\theta}. \quad (6)$$

**Learning.** We use Gibbs sampling for learning the parameters of the model, more specifically, collapsed Gibbs sampling [10] in the same manner as [12]. The factorizing prior presents itself in the learning as an additional factor in the objective function over  $z$ , compared to the original LDA model. It should be noted that the factorizing prior in Equation 5 is independent of the type of learning procedure – the model in Equation 6 can also be trained using, e.g., a variational method.

When training the model, the topics are initialized randomly, which means that they all have a  $H$  close to 1. During Gibbs sampling, it would be very unlikely to find a topic with low  $H$ , given the bimodality of  $A$  in Equation (4). To address this problem, we introduce an “auto-annealing” procedure, where  $A$  is replaced with a dynamic cooling function starting off by encouraging low  $H$  only, and gradually encouraging high  $H$  more and more, as the average  $H$  decreases (i.e., when some topics have found a class-specific state). Hence,  $A$  is changed to a dynamic function

$$A(k) = H(k)^2 - 2\hat{H}H(k) + 1 \quad (7)$$

where the average  $H$ ,  $\hat{H} = \sum_{k=1}^K H(k)/K$ , is used as an annealing parameter in the function. As with other annealing procedures, the “auto-annealing” procedure means that the factorizing prior  $p(\boldsymbol{\theta})$  changes in each step of the iterative learning procedure. In a normal annealing procedure, this change would be actuated by changing the annealing parameter. Here,  $\hat{H}$  can be thought of as an autonomous annealing parameter since it converges automatically to a value reflecting the fraction of the class-dependent versus class-independent variation in the data. For example, the text data set (Figure 4) has a lower  $\hat{H}$  than the natural scene dataset (Figure 5).

**Segmenting the topic space.** When the model have been trained we can evaluate the structure of the learned topic space by computing  $H(k)$  for each topic  $k$ . We consider topics with low  $H$  as class-dependent while topics with high  $H$  are considered as independent. As such the topic space can be “softly” segmented and interpreted in a class conditioned manner. As an example, the words building up the shared topics can be considered as *stop words*. In text processing, there is usually some standard stop words list, which can be used to pre-process the text. However, these stop words are predefined, for example, “the”, “at” etc. However, they sometimes also provide class-relevant information, for example, some topics are more location dependent or have more nouns. On the other hand, there are words, like “learning”, “performance” etc, which do not carry much information in, say, a machine learning conference corpus. In our model, we automatically learn the real stop words for the given domain. Furthermore, while it is easy to predefine the stop words in text data, this problem becomes much more challenging in computer vision applications. The “stop-visual-words” are ill-defined and much less intuitive to find, why an algorithm which automatically learns them, such as the one we propose, is very beneficial. We would like to emphasize again that there is still only one topic space; no hard splitting or removal of topics is done, neither for learning, nor for inference.

## 4 Experiments

The proposed model is evaluated on four different classification tasks, and compared to two baselines consisting of a regular LDA model with class label [8], and a model with stronger class-supervision in the topic learning, SLDA [3].

### 4.1 Object Classification

We first demonstrate how the factorization works using a toy dataset. The dataset, shown in Figure 2, is constructed to have a very high degree of structured noise. There are four object classes: bulb, car, duck, and mug. All 8 instances of a certain class have the same shape and image location. However, there is a very high intra-class variability in foreground and background texture. Furthermore, all four classes contain the exact same foreground-background texture combinations. Thus, the texture (which will dominate the variation among features from any visual extractor) can be regarded as structured noise, while the true signal relates to shape. The properties of this dataset can also be found to some extent in natural images: most realistic image and object classes display large intra-class appearance variation, and different classes share appearance aspects. Furthermore,



Figure 2: All the instances in the toy object dataset.

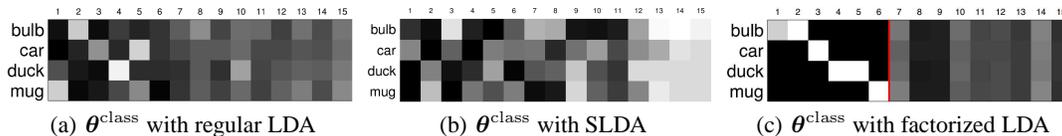


Figure 3: Toy object dataset. (a) Regular LDA topic distribution marginalized over class  $\theta^{\text{class}}$ , topics sorted in ascending order of class-specificity. (b) SLDA topic distribution marginalized over class  $\theta^{\text{class}}$ , topics sorted in ascending order of class-specificity. (c) Factorized LDA topic distribution marginalized over class  $\theta^{\text{class}}$ , topics sorted in ascending order of class-specificity, red line indicating partition between  $\theta^p$  and  $\theta^s$ .

the backgrounds in natural scenes are often complex and varying, introducing even more variation among training data for a class.

SIFT features on two different scales are densely extracted from all images, and a 64-word vocabulary is learned in which all SIFT features are represented. Thus, each image is represented by a bag of visual words in this vocabulary.

The experiment is performed in a hold-one-out manner, where each image in turn is classified using a model trained on the other 31 images. In the following, we will by “regular LDA” mean the regular LDA with upstream supervision presented in [8], but trained using Gibbs sampling in the same way as our model, with the same value of  $\alpha$  for all documents. With “SLDA”, we mean the more strongly supervised LDA variant with downstream supervision presented in [3], implemented by Blei et al.

Our proposed factorized LDA, as well as regular LDA and SLDA, are trained with 15 topics,  $\alpha = 0.1$  and  $\pi = 0.2$ . The classification performance for each class is found by averaging over the performances for the 8 images of that class.

It should be noted that the test image always will have a texture that is different from the training images of that class. However, the same texture can be found in other classes. A classifier that tries to explain all variation in the data in terms of class variation will therefore have difficulties in modeling this data set; a regular LDA or SLDA model trained with this data will be forced to represent texture as well as shape in the same topics, since the Dirichlet prior will promote topic sparsity. Thus, very few topics will purely represent one class, as shown in Figures 3(a) and 3(b).

However, our model, which explicitly factorizes the topics into those private to a certain class and those shared between all classes, will allow the relevant shape variation to be represented separately from the texture variation, which will just confuse the classification in this case. Figure 3(c) shows the factorized topic distribution; it is clear that the topics in  $\theta^p$  are private to a certain class, while the noise topics in  $\theta^s$  are shared equally over all classes; all the structured noise has thus been pushed into  $\theta^s$ . Thus, even though the full topic space is used for classification, it is effectively only based on  $\theta^p$ , while the shared topics  $\theta^s$  (right of the red line in Figure 3(c)) are effectively disregarded in the classification since they appear with equal probability in all classes.

As expected, the explicit noise model greatly improves classification on this dataset: the factorized LDA reaches 81.25%, while a regular LDA reaches a classification rate of 34.38%, only slightly

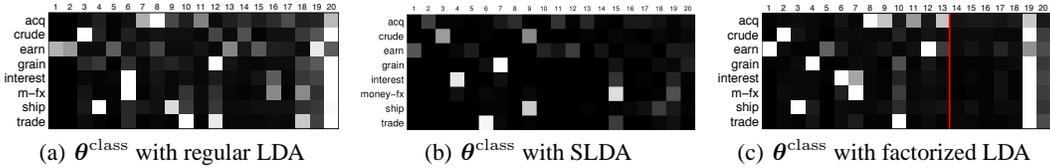


Figure 4: Reuters 21578 R8 dataset. (a) Regular LDA topic distribution marginalized over class  $\theta^{\text{class}}$ , topics sorted in ascending order of class-specificity. (b) SLDA topic distribution marginalized over class  $\theta^{\text{class}}$ , topics sorted in ascending order of class-specificity. (c) Factorized LDA topic distribution marginalized over class  $\theta^{\text{class}}$ , topics sorted in ascending order of class-specificity, red line indicating partition between  $\theta^p$  and  $\theta^s$ .

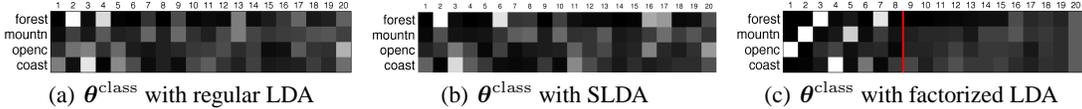


Figure 5: Natural scene dataset. (a) Regular LDA topic distribution marginalized over class  $\theta^{\text{class}}$ , topics sorted in ascending order of class-specificity. (b) SLDA topic distribution marginalized over class  $\theta^{\text{class}}$ , topics sorted in ascending order of class-specificity. (c) Factorized LDA topic distribution marginalized over class  $\theta^{\text{class}}$ , topics sorted in ascending order of class-specificity, red line indicating partition between  $\theta^p$  and  $\theta^s$ .

above chance, and SLDA who is forced by the stronger supervision to represent all variation (where texture is dominating) in terms of class achieves a result of 0% since the texture of the test image is not present in the training data of the same class.

## 4.2 Text Classification

We now evaluate the proposed model in a realistic text classification scenario. We use the standard R8 training and testing set from the Reuters 21578 dataset [26], which contains 5485 training documents and 2189 testing documents. The all-terms version of the data is used since we want to illustrate how our model deals with noise.

The regular LDA, SLDA and factorized LDA models are trained with 20 topics, and parameter settings  $\alpha = 0.5$  and  $\pi = 0.1$ . The topic distributions are shown in Figures 4(a), 4(b), and 4(c). The factorized class-private topic distribution  $\theta^p$  (left of the red line in Figure 4(c)) is noticeably cleaner than the regular distribution  $\theta^{\text{class}}$  (Figure 4(a)). In the factorized LDA, only  $\theta^p$  contributes to the classification, while the shared topics  $\theta^s$  (right of the red line in Figure 4(c)) are effectively disregarded since they appear with equal probability in all classes. The topics of the SLDA model are sparser (Figure 4(b)), but all topics are forced to be class-specific by the stronger supervision.

There is a significant classification improvement using the factorized topic space, from 74.63% with regular LDA and 63.75% with SLDA to 83.91% with factorized LDA.

## 4.3 Scene Classification

We also evaluate the proposed model on a challenging natural scene dataset used in [8]. There are four classes: forest, mountain, open country and coast, with 100 training images and 50 test images per class. From each image, SIFT features on two different scales are densely extracted, and labeled according to a 192-word vocabulary learned from the features, as in [8].

The regular LDA, SLDA, and factorized LDA models are trained with 20 topics, and parameter settings  $\alpha = 0.5$  and  $\pi = 0.1$ . Figures 5(a), 5(b), and 5(c) show the respective topic distributions; notably, the class-specific topic space  $\theta^p$  effectively used for classification in our factorized LDA only contains 8 topics, while 12 topics ( $\theta^s$ ) are devoted to modeling structured noise. Thus, the factorized representation is notably sparser than a regular LDA representation, which gives the opportunity to save both storage space and computation time during classification – an important factor to take into account for large datasets.

In addition to rendering a notably sparser data representation, the factorized LDA reaches a marginally higher performance rate than with a regular LDA and SLDA: 84.50% for our model compared to 80.50% for the regular LDA and 84.00% for SLDA. All performances are slightly better than the original implementation of the regular LDA [8], which reaches 76.0%.

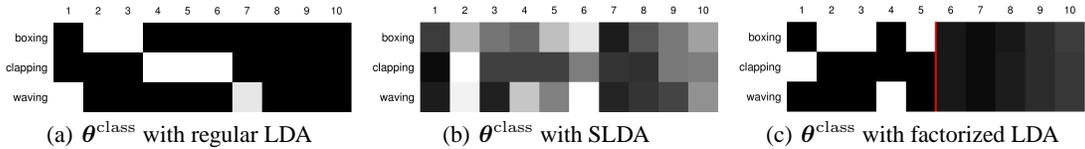


Figure 6: Action dataset. (a) Regular LDA topic distribution marginalized over class  $\theta^{\text{class}}$ , topics sorted in ascending order of class-specificity. (b) SLDA topic distribution marginalized over class  $\theta^{\text{class}}$ , topics sorted in ascending order of class-specificity. (c) Factorized LDA topic distribution marginalized over class  $\theta^{\text{class}}$ , topics sorted in ascending order of class-specificity, red line indicating partition between  $\theta^p$  and  $\theta^s$ .

#### 4.4 Action Classification

We proceed to evaluate the methods on a dataset with more variation independent of class. The dataset consists of three actions from the KTH Action dataset [15]: boxing, handclapping and handwaving. There are 100 short video sequences of each action, which show 25 different people performing the action, recorded in four shooting conditions (zooming and panning of camera, different background). The shooting condition has large influence on the motion in the video, as each zoom or panning motion adds global motion to the video and backgrounds contribute to the motion features as well. However, the variation in shooting condition is not at all correlated with action class in the dataset. Just as in the toy experiment above (but now in a more realistic setting), a large proportion of the data variation is thus independent of the action class. Due to the low signal-to-noise ratio, a topic model without factorization will have difficulties capturing the aspects of data relevant for discriminating activity class.

The experiment was performed by separating out from the training data all the 25 images of an action filmed with a certain shooting condition. The topic models were then trained with all other data, and evaluated with the 25 removed images. Hence, the certain combination of action and camera condition in the test data was not present in the training data. This was done for all actions in turn, and the result was averaged over actions.

STIP features [15] were extracted from all sequences and clustered into a vocabulary of 128 spatio-temporal words. This representation was used to train the regular LDA, SLDA and factorized LDA models with 10 topics,  $\alpha = 0.1$  and  $\pi = 0.1$ .

Figure 6 shows the topic distributions corresponding to these three models. We can see that Factorized LDA is able to model the class-dependent information (left of the red line) and the class-independent information (right of the red line), which makes it be able to archive better performance in noisy data. For the regular LDA, although the topics are not shared, however, it models all the information and assigned that to different classes with new topics which made the topics themselves became noisy. So does SLDA which models the "noise" as the useful topics.

Factorized LDA gives an accuracy of 65.22%, which is far better than both regular LDA, 38%, and SLDA, 51.33%. This confirms that the findings of the toy experiment above applies to realistic settings as well. Confusion matrices are shown in Figures 6(a), 6(b), and 6(c) respectively.

### 5 Conclusions

We present a factorized latent topic model, which explicitly represents aspects of the data which are not correlated with model state. Specifically, we train an LDA class model with an additional factorizing prior, which encourages topics to either be very class-specific or evenly shared among classes. The topic space  $\theta$  is thus partitioned into one part  $\theta^p$  whose topics are private to certain classes, and another part  $\theta^s$  with topics shared between classes. Only  $\theta^p$  contributes effectively to classification.

Experiments show the factorized LDA model to give consistently better classification performance and sparser topic representations than both a regular LDA model [8] and SLDA [3]. Sparse representations are advantageous for large datasets since they save storage space and computation time during classification.

Future work includes investigating the effect of this factorization prior on other topic models, such as HDP, and to integrate the prior into models with multiple data views, such as in [12, 28, 30].

## References

- [1] F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical report, 2005.
- [2] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
- [3] D. M. Blei and J. D. McAuliffe. Supervised topic models, *arxiv:1003.0783*, 2010.
- [4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [5] A. Damianou, C. H. Ek, M. Titsias, and N. D. Lawrence. Manifold Relevance Determination. In *ICML*, 2012.
- [6] C. H. Ek. *Shared Gaussian Process Latent Variable Models*, PhD Thesis, 2009.
- [7] C. H. Ek, J. Rihan, P. H. S. Torr, G. Rogez, and N. D. Lawrence. Ambiguity modeling in latent spaces. *Machine Learning for Multimodal Interaction*, 2008.
- [8] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.
- [9] M. R. Gormley, M. Dredze, B. Van Durme, and J. Eisner. Shared components topic models. In *NAACL*, 2012.
- [10] G. Heinrich. Parameter estimation for text analysis. Technical report, 2005.
- [11] T. Hofmann. Probabilistic latent semantic analysis. In *UAI*, 1999.
- [12] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *ICCV*, 2011.
- [13] A. Klami and S. Kaski. Generative models that discover dependencies between data sets. In *IEEE Workshop on Machine Learning for Signal Processing*, 2006.
- [14] A. Klami and S. Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72(1-3):39–46, 2008.
- [15] I. Laptev. On space-time interest points. *IJCV*, 64(2/3):107–123, 2005.
- [16] N.D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.
- [17] G. Leen and C. Fyfe. A Gaussian process latent variable model formulation of canonical correlation analysis. In *European Symposium on Artificial Neural Networks*, 2006.
- [18] G. Leen and C. Fyfe. Learning shared and separate features from two related data sets using GPLVM’s. In *Learning from Multiple Sources Workshop, Neural Information Processing*, 2008.
- [19] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1):259–289, 2008.
- [20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
- [21] R. Navaratnam, A. W. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *ICCV*, 2007.
- [22] J. Paisley, C. Wang, and D. Blei. The discrete infinite logistic normal distribution. *arXiv preprint arXiv:1103.4789*, 2011.
- [23] C. H. Papadimitriou, P. Raghavan, and H. Tamaki. Latent semantic indexing: a probabilistic analysis. *Journal of Computer and System Sciences*, (61):217–235, 2000.
- [24] J. P.W. Pluim, J.B. A. Maintz, and M. A. Viergever. Interpolation artefacts in mutual information-based image registration. *CVIU*, 77(2):1077–3142, 2000.
- [25] N. Rasiwasia, J. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *MMM*, 2010.
- [26] Reuters-21578. <http://csmining.org/index.php/r52-and-r8-of-reuters-21578.html>.
- [27] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.
- [28] S. Virtanen, Y. Jia, A. Klami, and T. Darrell. Factorized multi-modal topic model. In *UAI*, 2012.
- [29] C. Wang and D. M. Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *NIPS*, 2009.
- [30] C. Wang, D. M. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.