Pattern Recognition 77 (2018) 226–236

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Supervising topic models with Gaussian processes

Melih Kandemir[a,*], Taygun Kekeç[b], Reyyan Yeniterzi[a]

[a] Özyeğin University, Istanbul, Turkey
[b] Delft University of Technology, Pattern Recognition Laboratory, Delft, Netherlands

ARTICLE INFO

Article history:
Received 24 November 2016
Revised 8 December 2017
Accepted 30 December 2017
Available online 30 December 2017

Keywords:
Latent Dirichlet allocation
Nonparametric Bayesian inference
Gaussian processes
Variational inference
Supervised topic models

ABSTRACT

Topic modeling is a powerful approach for modeling data represented as high-dimensional histograms. While the high dimensionality of such input data is extremely beneficial in unsupervised applications including language modeling and text data exploration, it introduces difficulties in cases where class information is available to boost up prediction performance. Feeding such input directly to a classifier suffers from the curse of dimensionality. Performing dimensionality reduction and classification disjointly, on the other hand, cannot enjoy optimal performance due to information loss in the gap between these two steps unaware of each other. Existing supervised topic models introduced as a remedy to such scenarios have thus far incorporated only linear classifiers in order to keep inference tractable, causing a dramatical sacrifice from expressive power. In this paper, we propose the first Bayesian construction to perform topic modeling and non-linear classification jointly. We use the well-known Latent Dirichlet Allocation (LDA) for topic modeling and sparse Gaussian processes for non-linear classification. We combine these two components by a latent variable encoding the empirical topic distribution of each document in the corpus. We achieve a novel variational inference scheme by adapting ideas from the newly emerging deep Gaussian processes into the realm of topic modeling. We demonstrate that our model outperforms other existing approaches such as: (i) disjoint LDA and non-linear classification, (ii) joint LDA and linear classification, (iii) joint non-LDA linear subspace modeling and linear classification, and (iv) non-linear classification without topic modeling, in three benchmark data sets from two real-world applications: text categorization and image tagging.

1. Introduction

Developments in computing and communication technologies transformed the entire globe into a unified information society. We are exposed to a huge mass of data on a daily basis, which we are supposed to make sense of. A major portion of this data fits into the classical document-word-vocabulary conceptualization. A corpus consists of a number of documents and each document a number of words chosen from a vocabulary common to all documents in the corpus. The first step in analysis of such corpora is to represent the documents by a set of features. A natural feature set for a document is a vector of the frequencies of words in the vocabulary. This feature set provides a condensed and intuitive representation of a document at the expense of losing the word order information, hence it is named as the *bag-of-words (BoW)* representation. Nevertheless, it still maintains rich enough a portion of the statistical properties of the corpus for a myriad of analysis purposes.

A primary analysis task on a document corpus is to discover document groups in a clustering fashion in order to reveal relevant properties of data for a particular goal. Applying the idea of clustering to corpus analysis brings up the *topic* concept. Statistical topic models [1] are developed as analysis tools to infer topics that can summarize a corpus using probabilistic modeling principles. Probably the most widespread statistical topic model is Latent Dirichlet Allocation (LDA) [2], which is a Bayesian model that describes a document by its complete generative process. A generic corpus-level prior topic distribution generates the unique topic distribution of each document. Each word in this document is then generated by first sampling its topic from this distribution and then sampling the vocabulary element from the word distribution of the chosen topic. Useful applications of LDA include voting behavior analysis of large populations [3], modeling evolutions of documents in time [4], image segmentation [5], and gene expression profiling [6].

From a pure data analysis perspective, LDA can be thought as a dimensionality reduction tool for high-dimensional count data. Apart from being used for data exploration, inferred topic distributions can also be used as features to describe patterns. These

* Corresponding author.
E-mail addresses: melih.kandemir@ozyegin.edu.tr (M. Kandemir), I.T.Kekec@tudelft.nl (T. Kekeç), reyyan.yeniterzi@ozyegin.edu.tr (R. Yeniterzi).

https://doi.org/10.1016/j.patcog.2017.12.019
0031-3203/

features can then be fed into a classifier. However, multiple studies have shown that it is indeed possible to extend LDA to perform dimensionality reduction and classification jointly [7,8,10], which results in a consistent improvement in prediction performance. The reason for this improvement is that when jointly supervised, LDA is able to leverage the effect of hidden features which help with the discrimination of documents into previously seen categories. The common weakness of these earlier attempts to supervise LDA has been that they are satisfied with learning linear decision boundaries on topic distributions for the sake of feasible inference, severely limiting the expressive power of the resultant model.

In this paper, we introduce for the first time a way to supervise LDA with a *non-linear* predictor. Similarly to [11], we represent a document by its empirical topic activation frequency and feed this latent variable into a Gaussian process (GP) as input. We approximate the intractable posterior of the resultant model by a novel and efficient variational inference scheme. This scheme imports useful ideas from the advances in inference of GPs with stochastic input [12] and Deep GPs [13]. We show that our construction straightforwardly extends to the multilabel prediction setting as well.

Our model is applicable to any supervised data set that represents its instances as histograms. In a proof-of-concept study, we show how its usability could go far beyond plain bag-of-words or bag-of-visual-words representations. We convert a text corpus into the word vector format, cluster the word vectors, and calculate how many words from each cluster exist within each document. Consequently we use our GP-supervised topic model to jointly extract topics in the corpus and classify its documents. This way, we show at the conceptual level that the word-level interpretability of word vector representations, the power of neural nets in text modeling, and the document-level interpretability of topic models can be put together in an effective document classification model.

We evaluate our resultant model on three data sets from two challenging real-world applications: (i) free-form text categorization, and (ii) image tagging. Our model proves to be more accurate in label prediction than four key baselines that had set the state of the art prior to this work: (i) disjoint LDA and non-linear classification, (ii) joint LDA and linear classification, (iii) joint linear (non-LDA) subspace modeling and linear classification, and (iv) non-linear classification without LDA. The source code of our model is publicly available.[1]

2. Related work

Supervised topic modeling approaches mainly differ in the form of explaining how supervision is being generated. Depending on the application, this supervision can represent class labels, response probabilities or meta-data that is available during the collection of each document. For instance, meta-data can represent review ratings, author identity, spatial location of the document in a graph, publication venue or document timestamps. Unfortunately, traditional unsupervised topic models such as PLSA and LDA are designed only to use the discrete bag-of-words representation and cannot exploit such meta-data. Upon availability of such information, supervised topic modeling follows two main lines of thought. Downstream topic modeling approaches assume that document topics are the causes of responses. This means given some bag-of-words observations for each document, topics generate the responses. For example, sLDA [11] and our approach fall into this category. Upstream topic models instead, explain topic generation based on conditioning on the meta-data. Given the meta-data, the underlying document topics are the effects. This way of modeling is especially effective in predicting the unknown author of documents using mixed-membership models [14]. Different authors will have a different command of vocabulary, and their topics of interest are likely to differ. Dirichlet-Multinomial regression model of [15] also falls into this second stream, and generates topics from the meta-data.

In the seminal sLDA (Supervised LDA) [11], topics generate the reviewer ratings by learning the parameters of a generalized linear model (GLM) with an appropriate link function. MedLDA (Maximum Entropy Discrimination Latent Dirichlet Allocation) [8,16] model is yet another supervised topic model which bases on deriving a max-margin principle for the topical data under the variational inference framework. Their problem yields an entropy-regularized posterior distribution of the support vector coefficients. Our model differs from the MedLDA in utilizing a kernelized learner: a Gaussian process [17].

Simultaneous image representation and annotation tasks benefit significantly from supervision of LDA. Under a different taxonomy, these models can also be classified into generative and discriminative approaches. DiscLDA is such a discriminative topic modeling method [10] while sLDA is a generative method. Our approach falls into the class of generative models since it explains the generation of both words and responses. Other work includes various topic models such as the inverse regression topic model of Rabinovic [18] that considers both meta-data and distortion vectors effect the topics in context. Shi et al. [19] present a weakly-supervised object localization approach based on joint topic modeling to locate objects in an image. Alternatively, Zhang et al. [20] propose a downstream modeling approach that draws the class labels from the average words using a softmax probability rule. The downside of this approach is that variational inference becomes difficult due to the nonlinear operation. In contrast, we draw them from a Gaussian process. The work of [21] proposes topic modeling for the autoregressive setting. Niu et al. [22] proposes SS-RTM (Semi Supervised Relational Topic Model) combining multi-modal tagging as a semi-supervised problem. For each class category, one binary problem is formed to distinguish between positive and negative tagged documents. For a comprehensive review of various aspects of supervised topic models, the reader is referred to [23].

There exist few approaches to reconcile LDA and GPs [24,35]. This line of work employs Gaussian processes to generate topic distributions of documents on a non-linear latent manifold as proposed by Lawrence et al. [25]. Kernel Topic Models (KTM) [24] replace the Dirichlet prior of the topic distribution in LDA by a GP. The multinomial topic distribution is parameterized by a softmax function and the parameters of the softmax function are assigned a GP as a prior. The Gaussian Process Topic Model (GPTM) [35] employs GPs as hyperpriors, rather than as priors, on the prior of the topic distribution. The GP output linearly determines the mean of a normal prior applied on the topic distribution. Both of these models resort to GPs to capture non-linear correlations across topics. They are *unsupervised* density estimators. In this paper, we use GPs for a different and novel purpose: *supervision* of document categories. As opposed to GPs being priors or hyperpriors on the topic distributions in KTM and GPTM, in our model, the topic distributions are priors to the inputs of the GPs that predict the document label. Hence, although they look in close conceptual relation, such unsupervised models are not competitors for our supervised approach.

Lastly, there exist adaptations of LDA to document classification problems as disjoint feature extractors. The spherical topic model learns a unit vector for each document, as opposed to the word histogram of standard LDA. Such a representation enriches the feature space that can be modeled by LDA, improving classification

[1] https://github.com/melihkandemir/gpstm.

performance [9]. This approach differs from all supervised topic models, including ours, in that it treats topic modeling and classification as two disjoint steps. Supervised topic models perform end-to-end training.

Apart from the context of topic modeling, prior art spans multiple studies about theoretical foundations of machine learning with GPs. For instance, Kemmler et al. [53] adapt GPs to the one-class classification setup, where observations from only the negative class are available at training time. Another study by Huber et al. [54] sparsifies GPs by a recursive formulation that leads to a learning algorithm trainable in streaming data scenarios. Thanks to their high expressive power, GPs also prove useful predictors in appealing real-world applications, such as action recognition [55] and human gait analysis [52].

In addition to GPs, topic models also have a large spectrum of real-world applications. Hu et al. [48], which introduces a Gaussian-LDA model for audio retrieval, which is a continuous data processing task. The model skips the vector quantization step in order to avoid the possible loss of information and directly models the topics as a Gaussian distribution over the continuous features. For the video classification task, Hou et al. [47] propose multilayer multi-view topic model (MLMV-LDA) which integrates high-level representations from several different views using topic modelling techniques. For the task of person re-identification from videos, Liu et al. [50] propose Attribute-Restricted Latent Topic Models (ARLTM) which make use of the high-level human-specific knowledge, such as human-provided semantic information independent from the pose and the camera view. In a video retrieval related study [46], the authors extend the unsupervised symmetric pLSA approach with a supervised model and apply relevance feedback for deducing a probabilistic latent topic ranking function.

Last but not least, methodological studies on topic models include [51], which combines generative and discriminative modeling approaches by constructing a dependency between topic distributions of neighboring documents, and Labeled Latent Dirichlet Allocation (L-LDA) [49] which extends LDA with a prior distribution on label-word matchings.

3. The Gaussian Process Supervised Topic Model

We are given a set of independent and identically distributed (i.i.d.) documents $\boldsymbol{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D\}$ and a vocabulary of V words. Each document d is characterized by the histogram $\mathbf{w}_d \in \mathbb{R}^V$ of the appearances of the vocabulary items. Differently from the unsupervised exploratory analysis approaches, in our setting each document also has a corresponding R-dimensional vector of outputs $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_D\}$ with $\mathbf{y}_d \in \mathbb{R}^R$. We choose these outputs to be real-valued both to keep our model generic and to simplify the calculations. We nevertheless perform experiments only on data sets with categorical outputs corresponding to document tags.

Our aim is to devise a topic model that learns a mapping from the word histograms $\mathbf{w}_d$ to outputs $\mathbf{y}_d$. We start by modeling the topic generation process, for which we directly adopt the well-known Latent Dirichlet Allocation (LDA) approach

$$\begin{aligned} p(\boldsymbol{\theta}_d) &= Dir(\boldsymbol{\theta}_d|\boldsymbol{\alpha}), && d = 1, \cdots, D, \\ p(\mathbf{z}_n^d|\boldsymbol{\theta}_d) &= Mult(\mathbf{z}_n^d|\boldsymbol{\theta}_d), && n = 1, \cdots, N_d, \\ p(\mathbf{w}_n^d|\mathbf{z}_n^d, \boldsymbol{\beta}) &= Mult(\mathbf{w}_n^d|\boldsymbol{\beta}_{\mathbf{z}_n^d}), && n = 1, \cdots, N_d, \end{aligned}$$

and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K]$ where the indices d, k, and n run over documents in the corpus, topics, and words within a document, respectively. The scalar N_d denotes the total number of words in a document. The functions $Dir(\cdot|\cdot)$ and $Mult(\cdot|\cdot)$ denote a Dirichlet density and a multinomial mass function, respectively, both defined on the first argument and parameterized by the second. LDA starts the document generation process by sampling the topic distribution of each document d via density $p(\boldsymbol{\theta}_d)$ which is then used to determine the topic assignment $\mathbf{z}_n^d$ of word n of document d. Here, $\mathbf{z}_n^d$ is a 1-of-K coded vector, which has zeros in all entries and 1 on the entry corresponding to the active topic for word n of document d. The vector $\boldsymbol{\alpha}$ is a hyperparameter for $p(\boldsymbol{\theta}_n)$, which determines the functional form of the topic distribution of a generic document. Lastly, the 1-of-V coded word $\mathbf{w}_n^d$ is sampled from the word distribution $\boldsymbol{\beta}_k$ of the topic encoded by $\mathbf{z}_n^d$ and expressed in short hand as $\boldsymbol{\beta}_{\mathbf{z}_n^d}$.

Our key contribution is a novel way to supervise the topic model. The standard LDA has been used exhaustively in extracting exploratory information from various types of corpora, where target patterns are not known. However, there also exist ecologically-valid scenarios where it is possible to assign documents to predefined classes. It is more desirable for such scenarios to benefit from available document tags to enhance prediction accuracy. The two seminal attempts that address this problem are sLDA [11] and MedLDA [8]. Both of these models learn the topic distribution of a document and *linearly* classify the document from this distribution. The main challenge of this approach comes from the fact that LDA includes the inputs of these classifiers (i.e. topic distributions) as latent variables. Benefiting from the advances in approximate inference of Gaussian processes [12], we introduce the first method that can supervise LDA with a jointly-trained *non-linear* label predictor. We choose a GP as our predictor to achieve non-linearity.

The straightforward way to join the LDA latent variables with GP would be to feed the topic distribution variable $\boldsymbol{\theta}_d$ of document d into the GP as input. This approach introduces many complications to the inference scheme, which originate from two facts: (i) Dirichlet and normal distributions are not conjugates, (ii) the Dirichlet distribution is bound with the constraint that its target variable has to sum up to one. This same problem has also been encountered in sLDA [11] and has been overcome by binding LDA and the predictor through the topic assignment variable $\mathbf{z}_n^d$. More specifically, the topic distribution of document d is approximated by its inferred empirical mean of its topic activation frequencies:

$$\boldsymbol{\theta}_d \approx \frac{1}{N_d} \sum_{n=1} \mathbf{z}_n^d.$$

We adopt the same intuition and first extend LDA by

$$p(\mathbf{c}_d|\mathbf{Z}_d) = \mathcal{N}\left(\mathbf{c}_d \middle| \frac{1}{N_d} \sum_{n=1}^{N_d} \boldsymbol{z}_{k,n}^d,\ \zeta^{-1}\boldsymbol{I}\right),$$

where ζ is noise precision and $\mathbf{Z}_d = [\mathbf{z}_1^d, \cdots, \mathbf{z}_{N_d}^d]$. We refer to this new variable as the *latent document feature vector*, as it acts as a feature vector that characterizes the document for the subsequent prediction task. Finally, we feed this set of latent document feature vectors $\mathbf{C} = [\mathbf{c}_1, \cdots, \mathbf{c}_D]$ into a GP regressor as input:

$$\begin{aligned} p(\mathbf{f}_r|\mathbf{C}) &= \mathcal{N}(\mathbf{f}_r|\mathbf{0}, \mathbf{K}_{\mathbf{CC}}), \qquad r = 1, \cdots, R, \\ p(\mathbf{y}_r|\mathbf{f}_r) &= \mathcal{N}(\mathbf{y}_r|\mathbf{f}_r, \kappa^{-1}\boldsymbol{I}), \qquad r = 1, \cdots, R, \end{aligned}$$

where r denotes the output dimension r and $\mathbf{K}_{\mathbf{CC}} \in \mathbb{R}^{D \times D}$ is the GP covariance matrix, each entry of which contains the similarity of a pair of documents for a predetermined similarity metric (i.e. a kernel function) $\mathbf{K}_{\mathbf{CC}}[d, d = k(\mathbf{c}_d, \mathbf{c}_d')$, and κ is noise covariance. To enjoy ultimate non-linearity we choose the Radial Basis Function (RBF) as our kernel function:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2\sigma^2}||\mathbf{x} - \mathbf{x}'||_2^2\right),$$

where σ^2 is the kernel bandwidth. Put together, we refer to this model as the *Gaussian Process Supervised Topic Model (GPSTM)*. The relationship of its constituent latent variables is depicted in Fig. 1.

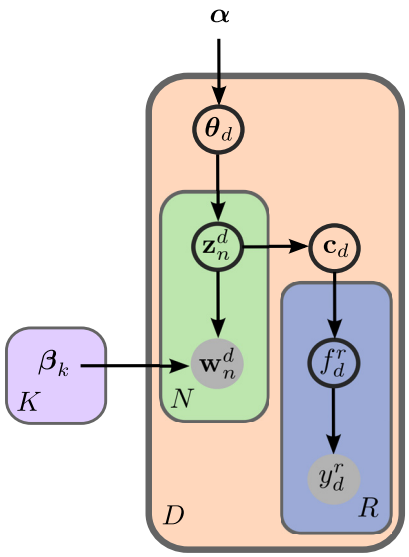

Fig. 1. The plate diagram of the proposed Gaussian Process Supervised Topic Model (GPSTM). The plain LDA on the left-hand side (green plate) is connected to a Gaussian process on the right (blue plate) through a latent document feature vector $\mathbf{c}_d$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1. Inference

Let $\mathcal{W} = \{\mathbf{W}_1, \cdots, \mathbf{W}_D\}$ denote the collection of the 1-of-V coded word occurrences $\mathbf{W}_d$ of each document d in the corpus. In order to perform learning, we aim to infer the posterior

$$p(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{C}, \mathbf{F}, \mathbf{U}|\mathcal{W}, \mathbf{Y}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{D} p(\boldsymbol{\theta}_d)\Big[\prod_{n=1}^{N_d} p(\mathbf{z}_n^d|\boldsymbol{\theta}_d)p(\mathbf{w}_n^d|\mathbf{z}_n^d, \boldsymbol{\beta})\Big]p(\mathbf{c}_d|\mathbf{Z}_d) \times \prod_{r=1}^{R}\Big\{p(\mathbf{u}_r|\mathbf{G})p(\mathbf{f}_r|\mathbf{u}_r, \mathbf{C}, \mathbf{G})p(\mathbf{y}_r|\mathbf{f}_r)\Big\}p(\mathcal{W}, \mathbf{Y}; \boldsymbol{\alpha}, \boldsymbol{\beta})^{-1}, \quad (1)$$

where $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_D]$ and $\mathbf{Z} = [\mathbf{Z}_1, \cdots, \mathbf{Z}_D]$. Above, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are model hyperparameters. The evidence term $p(\mathcal{W}, \mathbf{Y}; \boldsymbol{\beta})$ dividing the joint density is obtained by integrating out the latent variables $\boldsymbol{\theta}$, $\mathbf{Z}$, $\mathbf{C}$, $\mathbf{f}$ in the numerator, which is not tractable. Hence, we need to resort to approximate Bayesian inference. As there exist well-established and scalable inference techniques for both LDA [26] and GP [27] within the framework of variational inference, we also adopt this framework and inherit the existing know-how wherever possible.

The point that brings LDA and GP together introduces an unpleasant term for variational inference. The GP takes $\mathbf{C}$ as input, which follows a normal distribution. Eventually, this leads to having to take the expectation of the inverse of the GP covariance matrix $\mathbf{K}_{\mathbf{CC}}$ with respect to a normal distribution, which is both intractable and excessively costly for sampling (i.e. each sample would involve inversion of a $D \times D$ matrix. As a workaround, we decompose the GP in our model following the Fully Independent Training Conditional (FITC) approximation used earlier in Bayesian Gaussian Process Latent Variable Models [12] and later on set a basis for Deep GPs [13]

$$p(\mathbf{u}_r|\mathbf{G}) = \mathcal{N}(\mathbf{u}_r|\mathbf{0}, \mathbf{K}_{\mathbf{GG}}), \quad (2)$$

$$p(\mathbf{f}_r|\mathbf{X}, \mathbf{u}_r, \mathbf{C}) = \mathcal{N}(\mathbf{f}_r|\mathbf{K}_{\mathbf{GC}}^T\mathbf{K}_{\mathbf{GG}}^{-1}\mathbf{u}_r, \mathbf{K}_{\mathbf{CC}} - \mathbf{K}_{\mathbf{GC}}^T\mathbf{K}_{\mathbf{GG}}^{-1}\mathbf{K}_{\mathbf{GC}}), \quad (3)$$

$$p(\mathbf{y}_r|\mathbf{f}_r) = \mathcal{N}(\mathbf{y}_r|\mathbf{f}_r, \kappa^{-1}\mathbf{I}), \quad (4)$$

where $\mathbf{G} = [\mathbf{g}_1, \cdots, \mathbf{g}_P]$ is a pseudo input data set, called the *inducing point set*, consistin.g of P inducing points, $\mathbf{u}_r$ is the vector of its corresponding pseudo targets, called the *inducing outputs*, and $\mathbf{K}_{\mathbf{GC}}$ is the matrix of the kernel responses between $\mathbf{G}$ and $\mathbf{C}$. Here, Eq. (2) applies a GP prior on the inducing point set. As this is an imaginary data set whose entries appear only as hyperparameters, its size can be manually tuned. Common practice is to choose P to be dramatically smaller than the data size (D in our case). Eq. (3) is the standard posterior predictive density of a GP regressor. Essentially, it *predicts* the noise-free versions of real observations ($\mathbf{f}_r$) from the inducing points. Finally, Eq. (4) injects white noise on the observations as in the full GP regression case.

We approximate the intractable posterior in Eq. (1) by a distribution which factorizes as

$$Q = \prod_{d=1}^{D}\Big[q(\boldsymbol{\theta}_d)\Big[\prod_{n=1}^{N_d} q(\mathbf{z}_n^d)\Big]q(\mathbf{c}_d)\Big]\prod_{r=1}^{R}\Big[q(\mathbf{u}_r)p(\mathbf{f}_r|\mathbf{u}_r, \mathbf{C})\Big].$$

Variational inference approximates the posterior by minimizing its backward Kullback-Leibler (KL) divergence with respect to this factor distribution

$$\underset{Q}{\text{argmin}} \;\; \mathbb{KL}\Big[Q\Big|\Big|p(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{C}, \mathbf{f}|\mathcal{W}, \mathbf{Y}; \boldsymbol{\alpha}, \boldsymbol{\beta})\Big].$$

Rearranging the terms, we can show that minimizing this KL divergence is equivalent to maximizing the Evidence Lower Bound (ELBO), which reads

$$\mathcal{L} = \int\int\int \sum_{\mathbf{Z}} Q \log \frac{p(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{C}, \mathbf{F}, \mathbf{U}|\mathcal{W}, \mathbf{Y}; \boldsymbol{\alpha}, \boldsymbol{\beta})}{Q} d\boldsymbol{\theta} d\mathbf{C} d\mathbf{F} d\mathbf{U}.$$

Placing the terms and taking the integrals, it turns out to be possible to update the factors $q(\mathbf{u}_r)$, $q(\boldsymbol{\theta}_d)$, and $q(\mathbf{z}_n^d)$ in closed form, using the rule

$$q \leftarrow \mathbb{E}_{Q\setminus q}[\log p(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{C}, \mathbf{F}, \mathbf{U}; \boldsymbol{\alpha}, \boldsymbol{\beta})],$$

where q is any of the factors listed above and $Q\setminus q$ denotes all factors in Q except q. Among these factors, $q(\boldsymbol{\theta}_d)$ follows the standard LDA update rule

$$q(\boldsymbol{\theta}_d) \leftarrow Dir(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d), \quad (5)$$

where

$$\boldsymbol{\gamma}_d = [\gamma_1^d/Z_\gamma, \cdots, \gamma_K^d/Z_\gamma]$$

with

$$\gamma_k^d \leftarrow \alpha_k + \sum_{n=1}^{N_b} q(\mathbf{z}_n^d = k).$$

and $Z_\gamma = \sum_{k=1}^{K}\gamma_k^d$. Here $q(\mathbf{z}_n^d = k)$ denotes the expectation of factor $q(\mathbf{z}_n^d)$ to choose topic k, which will be identified below in the subsequent update rules. The update rule for $q(\mathbf{z}_n^d = k)$ follows similar lines to [7] and reads

$$q(\mathbf{z}_n^d) \leftarrow Mult(\mathbf{z}_n^d|\boldsymbol{\phi}_n^d), \quad (6)$$

where $\boldsymbol{\phi}_n^d = [\phi_{n,1}^d/Z_z, \cdots, \phi_{n,K}^d/Z_z]$ with

$$\phi_{n,k}^d \leftarrow \beta_{k,v_n}\exp\{\Psi(\gamma_k^d)\}\exp\Big\{\frac{\zeta}{N_d}\mathbb{E}_{q(\mathbf{c}_d)}[\mathbf{c}_d] - \frac{\zeta}{2N_d^2}\Big[\sum_{j\neq n}\phi_{j,k}^d + 1\Big]\Big\},$$

where $\mathbb{E}_{q(\mathbf{c}_d)}[\mathbf{c}_d]$ will be determined later below. Here, $Z_z = \sum_{k=1}^{K}\phi_{n,k}^d$ and β_{k,v_n} is the probability of vocabulary element v corresponding to word n of document d for topic k and $\Psi(\cdot)$ is the digamma function. The final closed-form update is

$$q(\mathbf{u}_r) \leftarrow \mathcal{N}(\mathbf{u}_r|\mathbf{m}_r, \mathbf{S}_r),$$

where

$$\mathbf{S}_r \leftarrow \left[\mathbf{K}_{\mathbf{GG}}^{-1} + \kappa \mathbf{K}_{\mathbf{GG}}^{-1}\mathbb{E}_{q(\mathbf{C})}[\mathbf{K}_{\mathbf{GC}}\mathbf{K}_{\mathbf{GC}}^T]\mathbf{K}_{\mathbf{GG}}^{-1}\right]^{-1},$$

$$\mathbf{m}_r \leftarrow \kappa \mathbf{S}_r \mathbf{K}_{\mathbf{GG}}^{-1}\mathbb{E}_{q(\mathbf{C})}[\mathbf{K}_{\mathbf{GC}}]\mathbf{y}_r.$$

Following the common practice, we learn $\boldsymbol{\beta}$ by Type II Maximum Likelihood, which involves maximization of the marginal likelihood with respect to the hyperparameter of interest. While the marginal likelihood of our model is not tractable, the ELBO serves as an approximation to it. Hence, we maximize ELBO with respect to $\boldsymbol{\beta}$:

$$\underset{\boldsymbol{\beta}}{\text{argmax}} \quad \log p(\mathcal{W}, \mathbf{Y}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \geq \underset{\boldsymbol{\beta}}{\text{argmax}} \quad \mathcal{L},$$
$$s.t. \sum_{v=1}^{V} \beta_k^v = 1, \qquad k = 1, \cdots, K.$$

Setting the gradient of the ELBO with respect to $\boldsymbol{\beta}$ to zero, also incorporating the constraints on the topic word distributions to sum up to unity via Lagrange multipliers, we attain the standard update rule for the unsupervised LDA:

$$\beta_k^v \leftarrow \sum_{d=1}^{D}\sum_{n=1}^{N_d} \phi_{n,k}^d / Z_\beta,$$

where $Z_\beta = \sum_{v=1}^{V} \beta_k^v$.

Lastly, we need to update $q(\mathbf{c}_d)$, which does not have a closed-form solution. We impose this factor to follow a normal distribution for convenience:

$$q(\mathbf{c}_d) = \mathcal{N}(\mathbf{c}_d | \mathbf{a}_d, \mathbf{H}_d).$$

Then we take the gradient of the resultant ELBO with respect to its mean

$$\nabla_{\mathbf{a}_d}\mathcal{L} = -\zeta \mathbf{a}_d + \frac{\zeta}{N_d}\sum_{n}^{N_d} \boldsymbol{\phi}_{n,d} + \kappa \sum_{r=1}^{R}\sum_{d=1}^{D} (y_r^d) \frac{\partial \mathbb{E}_{q(\mathbf{c}_d)}[\mathbf{k}_{\mathbf{Gc}_d}]}{\partial \mathbf{a}_d}\mathbf{K}_{\mathbf{GG}}^{-1}$$
$$-\frac{\kappa}{2}\sum_{d=1}^{D} tr\left(\left[\frac{\partial \mathbb{E}_{q(\mathbf{c}_d)}[\mathbf{k}_{\mathbf{Gc}_d}\mathbf{k}_{\mathbf{Gc}_d}^T] + \mathbf{I}\right]}{\partial \mathbf{a}_d}\mathbf{K}_{\mathbf{GG}}^{-1}\left(\sum_{r=1}^{R}[\mathbf{m}_r\mathbf{m}_r^T + \boldsymbol{S}_r]\right)\mathbf{K}_{\mathbf{GG}}^{-1}\right)$$

and use it to traverse towards a local optimum by gradient-descent steps

$$\mathbf{a}_d^{(t+1)} \leftarrow \mathbf{a}_d^{(t)} + \eta \nabla_{\mathbf{a}_d}\mathcal{L},$$

where η is the step size or learning rate and the superscript (t) denotes the iteration count. In this update, the expectations with respect to kernel matrices can be analytically calculated, as suggested by Titsias et al. [12] for Bayesian GPLVMs, which is the key virtue that enables tractable inference of deep GPs [13]:

$$\mathbb{E}_{q(\mathbf{c}_d)}[k(\mathbf{g}_p, \mathbf{c}_d)]$$
$$= |\sigma^{-2}\mathbf{H}_d + \mathbf{I}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{g}_p - \mathbf{a}_d)^T(\sigma^2\mathbf{I} + \mathbf{H}_d)^{-1}(\mathbf{g}_p - \mathbf{a}_d)\right\},$$

$$\mathbb{E}_{q(\mathbf{c}_d)}[k(\mathbf{g}_p, \mathbf{c}_d)k(\mathbf{g}_{p'}, \mathbf{c}_d)]$$
$$= |2\sigma^{-2}\mathbf{H}_d + \mathbf{I}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{r}^T(2\sigma^2\mathbf{I} + \mathbf{H}_d)^{-1}\mathbf{r}\right\}.$$

where $\mathbf{r} = (\mathbf{g}_p + \mathbf{g}_{p'})/2 - \mathbf{a}_d$. Finally, the gradients of these terms with respect to $\mathbf{a}_d$ are

$$\frac{\partial \mathbb{E}_{q(\mathbf{c}_d)}[k(\mathbf{g}_p, \mathbf{c}_d)]}{\partial \mathbf{a}_d} = \mathbb{E}_{q(\mathbf{c}_d)}[k(\mathbf{g}_p, \mathbf{c}_d)](\sigma^2\mathbf{I} + \mathbf{H}_d)^{-1}(\mathbf{g}_p - \mathbf{a}_d).$$

and

$$\frac{\partial \mathbb{E}_{q(\mathbf{c}_d)}[k(\mathbf{g}_p, \mathbf{c}_d)k(\mathbf{g}_{p'}, \mathbf{c}_d)]}{\partial \mathbf{a}_d}$$
$$= \mathbb{E}_{q(\mathbf{c}_d)}[k(\mathbf{g}_p, \mathbf{c}_d)k(\mathbf{g}_{p'}, \mathbf{c}_d)](2\sigma^2\mathbf{I} + \mathbf{H}_d)^{-1}\mathbf{r}.$$

Although the gradient of $\mathcal{L}$ with respect to $\mathbf{H}_d$ is also tractable, we fix it to a spherical covariance for the sake of faster inference: $\mathbf{H}_d = 0.1\mathbf{I}$. Due to similar reasons, we set $\alpha_d = 1$ and the noise precisions ζ and κ to 10.

3.2. Prediction

For a new set of D^* documents $\mathcal{W}^* = \{\mathbf{W}_1^*, \cdots, \mathbf{W}_{D^*}^*\}$, we are interested in predicting the corresponding labels $\mathbf{Y}^* = [\mathbf{y}_1^*, \cdots, \mathbf{y}_R^*]$. For an output dimension r, the ideal posterior predictive density reads

$$p(\mathbf{y}_r^* | \mathcal{W}^*, \mathcal{W}, \mathbf{Y})$$
$$= \int\int\int p(\mathbf{y}_r^* | \mathbf{f}_r^*) p(\mathbf{f}_r^* | \mathbf{u}_r, \mathbf{C}^*) p(\mathbf{u}_r | \mathcal{W}, \mathbf{Y})$$
$$\times \left[\int p(\mathbf{C}^* | \mathbf{Z}^*)\left\{\int p(\mathbf{Z}^*, \boldsymbol{\theta}^* | \mathcal{W}^*; \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta}^*\right\} d\mathbf{Z}^*\right] d\mathbf{C}^* d\mathbf{u}_r d\mathbf{f}_r^*.$$

In this formula, $p(\mathbf{Z}^*, \boldsymbol{\theta}^* | \mathcal{W}^*; \boldsymbol{\alpha}, \boldsymbol{\beta})$ corresponds to the posterior density for the plain unsupervised LDA on the new $\mathcal{W}^*$. As this term does not have an analytical solution, identically to the plain LDA, we need to approximate it by variational inference also in prediction time. Specifically, we learn an approximate posterior $q(\mathbf{Z}^*, \boldsymbol{\theta}^*) = q(\mathbf{Z}^*)q(\boldsymbol{\theta}^*)$ by iterating between the update rules given in Eqs. (5) and (6) on $\mathcal{W}^*$. Note here that $\boldsymbol{\beta}$ comes directly from training. Hence, in prediction time, we use the topic word distributions that are learned from the *supervised* training data to encode discriminative properties of document classes into word distributions of topics. Placing the approximate posterior of LDA into the posterior predictive density results in

$$p(\mathbf{y}_r^* | \mathcal{W}^*, \mathcal{W}, \mathbf{Y})$$
$$\approx \int\int\int p(\mathbf{y}_r^* | \mathbf{f}_r^*) p(\mathbf{f}_r^* | \mathbf{u}_r, \mathbf{C}^*) p(\mathbf{u}_r | \mathcal{W}, \mathbf{Y})$$
$$\times \left[\int p(\mathbf{C}^* | \mathbf{Z}^*)\left\{q(\mathbf{Z}^*)\int q(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^*\right\} d\mathbf{Z}^*\right] d\mathbf{C}^* d\mathbf{u}_r d\mathbf{f}_r^*.$$

Here, the integral $\int q(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^*$ vanishes to unity, as $q(\boldsymbol{\theta}^*)$ is a normalized density function. We approximate the subsequent intractable integral by its point estimate

$$\int p(\mathbf{C}^* | \mathbf{Z}^*) q(\mathbf{Z}^*) d\mathbf{Z}^* \approx p(\mathbf{C}^* | \mathbb{E}_{q(\mathbf{Z}^*)}[\mathbf{Z}^*]),$$

where $\mathbb{E}_{q(\mathbf{Z}^*)}[\mathbf{Z}^*]$ is simply the matrix of ϕ_n^d's learned from $\mathcal{W}^*$. Next, we replace $p(\mathbf{u} | \mathcal{W}^*, \mathbf{Y})$ by the approximate posterior $q(\mathbf{u})$ learned during training

$$p(\mathbf{y}_r^* | \mathcal{W}^*, \mathcal{W}, \mathbf{Y})$$
$$\approx \int\int\int p(\mathbf{y}_r^* | \mathbf{f}_r^*) p(\mathbf{f}_r^* | \mathbf{u}_r, \mathbf{C}^*) q(\mathbf{u}_r) p(\mathbf{C}^* | \mathbb{E}_{q(\mathbf{Z}^*)}[\mathbf{Z}^*]) d\mathbf{C}^* d\mathbf{u}_r d\mathbf{f}_r^*.$$

The integral over $\mathbf{C}^*$ is also intractable, since this term appears in $p(\mathbf{f}_r^* | \mathbf{u}_r, \mathbf{C}^*)$ as input to the RBF kernel function. Hence, we approximate this integral also by the point estimate resulting in

$$p(\mathbf{y}_r^* | \mathcal{W}^*, \mathcal{W}, \mathbf{Y})$$
$$\approx \int\int\int p(\mathbf{y}_r^* | \mathbf{f}_r^*) p(\mathbf{f}_r^* | \mathbf{u}_r, \mathbb{E}_{p(\mathbf{C}^* | \mathbb{E}_{q(\mathbf{Z}^*)}[\mathbf{Z}^*])}[\mathbf{C}^*]) q(\mathbf{u}_r) d\mathbf{u}_r d\mathbf{f}_r^*.$$

All three factors remaining inside the integrals on the right hand side are Gaussians, hence their integration is analytically

available as

$$p(\mathbf{y}_r^*|\mathcal{W}^*, \mathcal{W}, \mathbf{Y}) \approx \mathcal{N}\Big(\mathbf{y}_r^*\Big|\mathbb{E}[\mathbf{K}_{\mathbf{C}^*\mathbf{G}}]\mathbf{K}_{\mathbf{GG}}^{-1}\mathbf{m}_r, \mathbb{E}[\mathbf{K}_{\mathbf{C}^*\mathbf{C}^*}] - \mathbb{E}[\mathbf{K}_{\mathbf{C}^*\mathbf{G}}](\mathbf{K}_{\mathbf{GG}}^{-1}\mathbf{S}_r - \mathbf{I})\mathbf{K}_{\mathbf{GG}}^{-1}\mathbb{E}[\mathbf{K}_{\mathbf{GC}^*}] + \kappa^{-1}\mathbf{I}\Big),$$

where $\mathbb{E}[\mathbf{K}_{\mathbf{C}^*\mathbf{G}}]$ is a $D^* \times P$ matrix containing the kernel response

$$k\Big(\mathbb{E}_{p(\mathbf{c}_d|\mathbb{E}_{q(\mathbf{z}_d^*)})}[\mathbf{c}_d], \mathbf{g}_p\Big)$$

in its dth row and pth column. The matrices $\mathbb{E}[\mathbf{K}_{\mathbf{C}^*\mathbf{C}^*}]$ and $\mathbb{E}[\mathbf{K}_{\mathbf{GC}^*}]$ are similarly defined.

4. Experiments

4.1. Baselines

We compare GPSTM with the following four baselines:

- **RBF-SVM:** This baseline trains a Support Vector Machine (SVM) with a RBF kernel directly on the high-dimensional bag-of-words representations of documents and the corresponding labels. This is a control baseline to quantify the need for topic modeling for the sake of dimensionality reduction. Prior work reports cases where this baseline has been observed to be extremely competitive. For instance see Fig. 5 in [16].
- **BSSML:** Bayesian Semi-Supervised Multilabel Learning [30] is a relatively recent method to perform joint linear (i.e. non-LDA) subspace learning and prediction. We choose this baseline to motivate LDA as a design choice for dimensionality reduction on document corpora.
- **sLDA:** Supervised LDA [7] serves as a baseline to quantify how much expressive power GPSTM gains from the non-linearity introduced into the predictor part of the model by the GP.
- **LDA+GP:** This baseline trains LDA first and then fits a GP on the inferred topic distributions as two disjoint steps. Comparing to this baseline, we measure how advantageous it is to learn LDA and GP jointly. While cascading disjointly trained LDA and GP appears as a control baseline for our setup, this approach has been recently shown to be greatly competitive in text categorization [34].

4.2. Setup

All four baselines and our GPSTM require an iterative training procedure. We train all models in comparison until a maximum of 50 iterations unless they converge earlier. For all sparse GPs in GPSTM and LDA+GP, we set the inducing points to cluster centroids found by k-means. We assign a balanced proportion of inducing points to different classes by computing k-means separately for each subset of the training set belonging to a different class. For all three data sets, we use five inducing points per class. We use the RBF kernel as the covariance function for all GPs. We either choose the hyperparameters of the models by cross validation on the training split or set them to values advised by their authors. For GPSTM and GP+LDA, we choose the kernel bandwidth from the set {0.15, 0.25, 0.5, 0.75, 1.0}, topic count from the set {10, 20, 30, 40, 50}, and for GPSTM we choose learning rate from the set $\{10^{-10}, 10^{-11}, 10^{-12}, 10^{-13}, 10^{-14}\}$. For RBF-SVM, we set the kernel bandwidth to the square-root of the number of vocabulary elements and choose the regularization parameter C from the set {0.1, 1, 10}. It is noteworthy that the kernel bandwidth used for GPSTM and GP+LDA does not apply directly to RBF-SVM, as the former two models take average topic distributions as input, while RBF-SVM operates on the higher-dimensional bag-of-words input. For BSSML, we set the Gamma priors to the projection matrix, bias and weight parameters and projection standard deviation to 0.1, following the setting promoted by the authors.

4.3. Data sets

We evaluate GPSTM on three data sets from two challenging real-world applications: text categorization and image tagging. The details of the data sets are given below:

4.3.1. Yahoo! answers

This is a text categorization data set[2] which has been frequently used in recent literature for benchmarking [36,37]. The collection contains more than four million questions and their corresponding answers in different categories and languages. We select the ten largest main categories in English. For each question category a randomly selected 1000 samples are used for training and another 1000 is used for testing, summing up to 10,000 training and 10,000 test documents. Each question is represented by its question subject, question content, and best answer content.

The recently introduced vector space representation method [38,39] captures both semantic and syntactic regularities between words drastically more effectively than explicit language modeling approaches. Although this representation brings about a reasonable degree of interpretability at the word level, neural network based text categorization models that build on word vectors do not provide the document-level interpretability of topic models. With this experiment, we intend to demonstrate a proof-of-concept study for bringing together the strengths of word vector representations and topic models for document classification. To this end, first we pass the Yahoo! Answers corpus through the pre-trained Glove6B[3] network and convert all the words into word vector form. Then we cluster the word-vector-represented corpus into 50 clusters ($\approx$0.01% of the vocabulary size) with k-means. Finally, we extract a histogram representation for each document by assigning each word within a document to the closest cluster. We use the resultant representation in place of performing bag-of-words on raw text data. Even though Google's pre-trained word vectors were trained over the larger Google News data set, we use Glove6B due it being trained on not just newswire text but also Wikipedia text, which can provide a better and more general domain match with the Yahoo! Answers data set.

4.3.2. Corel5K

This is a public image tagging data set [29] with a predefined training and test split of 4500 and 500 images, respectively. Images are tagged with 260 object categories. Some sample images from this data set are shown in Fig. 2 along with their ground-truth tags. Automatic discovery of these tags is difficult especially because the objects of interest are both not centered and can appear in extremely diverse forms. For instance, compare the trees in the image at the bottom left corner with ones in the image at the bottom right corner. While the former contains them at the far background, the latter presents them much more saliently. The promoted remedy is to exploit co-localizations of categories, which is only possible with multilabel supervised learning.

We adopt the protocol and features provided by Guillaumin et al. [31] and use a bag-of-visual-words representation to characterize each image. This widespread technique extracts keypoints and the corresponding SIFT descriptors from an image data set, clusters these keypoints with k-means (i.e. generating visual words), and represents each image by a histogram of the cluster

[2] L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0 (multi part) retrieved from https://webscope.sandbox.yahoo.com/catalog.php?datatype=l.

[3] https://nlp.stanford.edu/projects/glove/.

Fig. 2. Sample images and tags from Corel5K data set.

assignments of the keypoints within that image. For practical reasons, we restrict our analysis to the most frequent 20 object categories and filter out the remaining 240. The object categories used in our experiments are: water, sky, tree, people, grass, buildings, mountain, flowers, snow, clouds, rocks, stone, street, plane, bear, field, sand, birds, beach, and boats. We also reduce the vocabulary size from 1000 to the most frequent 500. In this reduced setting, all documents in the training and test splits still contain at least one word and at least one object. Hence, the data set is not downsampled with respect to instances.

4.3.3. NUS-WIDE

This is a public data set [42] widely used for benchmarking image tagging approaches. The data set consists of a training split of 161,789 images, and a test split of 107,859 images that contain 81 object categories. A close investigation reveals the remarkable fact that earlier work does not have consensus on a standard experiment setup for the NUS-WIDE data set. Various seminal work report experiments on different setups. While [45] use 150,000 images for train and 59,347 for test, [43] choose 110,000 training, 40,000 validation, and 40,253 test images. Lastly, [44] uses 132,575 training images and 35,665 test images.

In absence of a gold standard, we adopt a setup akin to [44] and use 35,665 test images, and 50,000 training images. We prefer this setup over others, as it assumes the smallest test set size. GPSTM is not a sort of model that can trivially be scaled up via a deep learning library implementation, due to exploding number of symbolic variables coming from the inducing point set. Thus, its fully scalable implementation deserves a dedicated study, which we leave out of the scope of this proof-of-concept work.

In order to benefit from the power of existing deep learning methods, we introduce a method to adapt pretrained deep neural nets to the topic modeling framework as feature extractors. We feed each NUS-WIDE image into the pretrained VGG-16 net and record the activation map of the final fully-connected layer as its feature vector. Then we choose 1000 random subsets of size 40 from this feature vector. We merge the feature vector subsets of 1000 random images into one set and train K-means clustering with $K = 50$. Eventually, we obtain a Bag-of-Words (BoW) vector for each image by extracting the histogram of the cluster assignments of its feature subsets. The outcome is a rich representation of the co-activation patterns of the neurons residing at the highest abstraction level. The eventual BoW features are in a shape that can be directly fed into any sort of topic model, including our GPSTM.

4.4. Performance metrics

The *Yahoo! Answers* data set contains documents which are evenly distributed into ten categories. This is a multiclass classification problem, as a document can belong to only one category. For such a setup, we satisfy with the natural performance measure and use only plain accuracy. In the remaining two image tagging data sets, there are twenty object categories, multiple of which can be active for an image. Hence, we assess the performance of models with the following set of metrics, some of which are standard measures and some are special to retrieval setups:

- **Accuracy:** Percentage of correctly classified documents.
- **Mean Average Precision (mAP):** Retrieval precision averaged with respect to all possible recalls, hence all possible decision thresholds.
- **Mean image-centric Average Precision (miAP):** Introduced by Li et al. [40], this metric is defined for image k as $\frac{1}{R}\sum_j^Q \frac{r_j}{j}\delta(k,t_j)$, where R is the number of active tags in k, r_j is the number of active tags among top j ranking tags, Q is the number of tag types, and $\delta(k, t_j)$ returns 1 if tag t_j is active in image k and 0 otherwise. The mean of iAP's for all categories is called the miAP.
- **Top-3 F1 Score (F1@3):** We adopt the top-3 precision and recall definitions given in [41]. For each image, we choose the highest ranking three tags. Then we calculate top-3 precision as $\frac{1}{c}\sum_{i=1}^{c}\frac{N_i^c}{N_i^g}$, where c is the number of tags, N_i^c is the true positive count for tag i and N_i^g is the number of ground-truth positive images. Similarly, top-3 recall is defined as $\frac{1}{c}\sum_{i=1}^{c}\frac{N_i^c}{N_i^p}$, where N_i^p is the number of positive predictions. As precision and recall are not meaningful metrics when used alone, we summarized them with the F1 Score. For precision p and recall r, F1 score is defined as $F_1 = 2pr/(p+r)$.
- **Top-5 F1 Score (F1@5):** This metric is identical to F1@3 except that highest ranking five tags are chosen for each image [41].

4.5. Results

4.5.1. The effect of the topic count on performance

To provide an insight on how the topic count effects prediction performance, we conduct a small experiment on a binary classification task chosen from the TechTC-100 data set [28]: categorization of web pages based on whether their content is *music* or *artist*. The data set contains 79 documents belonging to the *music* cate-

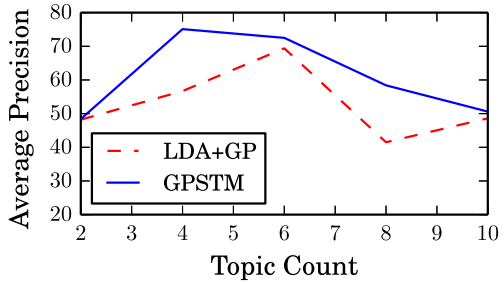

Fig. 3. Change of prediction performance as a function of topic count when LDA and GP are trained jointly (GPSTM) and disjointly (LDA+GP) on music versus artist classification task taken from the TechTC-100 data set.

Table 1
Performance scores of GPSTM and four baseline models for multioutput classification of ten object categories on images taken from the Yahoo! Answers data set.

	GPSTM (Ours)	LDA+GP	sLDA	BSSML	RBF-SVM
Accuracy (%)	**48.2**	47.7	44.8	47.4	15.2

gory and 84 to the *artist* category. We prefer this data set to assess the effect of topic count on performance, as it contains a relatively small but challenging set of documents. The challenge stems from the fact that the corpus is collected by a web crawler in an uncontrolled environment. Hence, the task demands from a model to achieve high expressiveness from limited number of observations, making design choices such as topic count maximally important. We reserve a randomly chosen 80% of the documents for training and use the remaining 20% for evaluation. We repeat this procedure five times and report mAP averaged across repetitions. Fig. 3 compares the evolution of the performance of GPSTM and LDA+GP as a function of the number of topics. GPSTM consistently outperforms LDA+GP, promoting joint training of the LDA and GP components. It is noteworthy that GPSTM reaches its best performance with a topic count of four. Fewer topics fall short in expressiveness and more topics cause overfitting. LDA+GP, on the other hand, has its peak at six topics.

4.5.2. Yahoo! answers data set

The immediate observation emerging from Table 1 is that GPSTM outperforms the runner-up LDA+GP, supporting our central hypothesis that training LDA and GP jointly improves on using LDA and GP as disjoint models. Notably, the performance of sLDA comes after both GPSTM and LDA+GP. This outcome illustrates how the non-linear RBF kernel used by the GP enhances the expressive power of the model, as opposed to the linear label predictor of sLDA. Lastly, RBF-SVM lags far behind all other baselines, demonstrating how vital the dimensionality reduction step is for bag-of-words data.

4.5.3. Corel5k data set

We train both GPSTM and the four baselines on the predefined training split and evaluate on the predefined test split. The performance scores of all models in comparison are listed in Table 2. GPSTM outperforms all baselines in four of the five performance metrics. Similarly to the Yahoo! Answers data set, LDA+GP proves to be the most competitive baseline. While sLDA gives second highest F1@5 score, it performs drastically worse than GPSTM in all other metrics. Lastly, BSSML, and RBF-SVM stay far behind all three topic model variants in all metrics.

Table 2
Performance scores of GPSTM and four baseline models for multi-output classification of 20 object categories on images taken from the Corel5k data set.

	GPSTM (Ours)	LDA+GP	sLDA	BSSML	RBF-SVM
Accuracy	**37.5**	32.5	16.0	27.3	20.8
mAP	**34.7**	29.0	26.7	17.4	19.9
miAP	**56.8**	53.6	47.0	43.6	36.9
F1@3	**50.3**	48.1	42.8	37.4	40.2
F1@5	44.8	**46.6**	45.9	38.5	40.4

Table 3
Performance scores of GPSTM and the baseline models for multioutput classification of 81 object categories on images taken from the NUS-WIDE data set. No results have been reported for sLDA since its public implementation does not complete 50 training iterations within 48 hours.

	GPSTM (Ours)	LDA+GP	Logistic	BSSML	sLDA
mAP	**29.3**	26.5	27.6	24.6	N/A
miAP	**62.9**	61.6	60.3	53.9	N/A
F1@3	**47.3**	46.5	45.5	41.5	N/A
F1@5	**42.4**	41.8	41.4	37.2	N/A

Table 4
Training durations of GPSTM and the baselines for 50 iterations in seconds.

	GPSTM (Ours)	LDA+GP	sLDA	BSSML	RBF-SVM
Yahoo! Answers	254.3	43.9	> 36,000	144.6	2.5
Corel5k	170.1	39.8	> 18,000	258.2	6.5
NUS-WIDE	2821.4	592.7	> 36,000	> 36,000	N/A

4.5.4. NUS-WIDE data set

Differently from Corel5k, we do not report plain accuracy in this experiment, as for almost all classes the positive cases are extremely rare, making accuracy an uninformative metric. We also replace RBF-SVM by logistic regression (denoted in Table 3 as Logistic), as the quadratic memory complexity of the kernel SVM with respect to data set size makes it intractable to be trained at the NUS-WIDE scale. Notably, applying logistic regression on VGG-16 activations is a standard baseline reported in various earlier work [43–45].

As seen in Table 3, GPSTM outperforms all baselines in all four performance metrics. It is should be noted that our experiment results are not comparable to previous work that introduce models dedicated to the image tagging problem [43–45], since the precise training and test splits are not released by any of them. Such a comparison would also be neither informative nor fair, as the main point of this paper is a novel method for supervising topic models independently from the application at hand.

4.6. Computational complexity

GPSTM improves on LDA+GP by end-to-end training a topic model and a GP. The bottleneck of GPSTM is calculation of the gradient $\nabla_{\mathbf{a}_d}\mathcal{L}$ that links the topic model to the GP. As given in Table 4, although this additional update naturally makes GPSTM slower than LDA+GP, its time complexity still remains within feasibility margins. Remarkably, the C++ implementation of sLDA provided by the authors[4] completes 50 iterations in approximately 5 h on Corel5K and NUS-WIDE, even longer for Yahoo! Answers. RBF-SVM runs much faster, as it does not learn a lower-dimensional representation for each document. Consequently, GPSTM brings

[4] https://github.com/blei-lab/class-slda.

about a profitable trade-off between performance and complexity. With a reasonable time overhead, it improves prediction accuracy consistently in all three data sets.

5. Conclusion

The main reason for the earlier work on supervised LDA to use a simple linear predictor has been to keep approximate posterior inference tractable. While replacing this linear predictor with the more complicated GP, we overcome the difficulties emerging on the inference side by importing ideas from the Bayesian GPLVM and the Deep GP literature. Consequently, we end up with a novel and powerful model, called GPSTM, to perform topic modeling and document classification jointly. In the experiments we conducted on three data sets from two challenging real-world applications, free-form text categorization and image tagging, we observed GPSTM to consistently outperform strong alternative approaches including sLDA, disjoint training of LDA and GP, and direct non-linear classification of document features.

In future work, we intend to extend the GPSTM in both inference and representation aspects. Firstly, we are interested in deriving a variant of GPSTM that can operate also on image patches, relaxing the common assumption and preserving the spatial structure inherent in image applications. Secondly, our current framework learns the inducing points by clustering the training set. For some applications, optimizing the inducing points jointly with the latent variables during posterior inference could be more beneficial. One direction forward in this regard can be to resort to information-theoretic heuristics for selection of the inducing points from the training set, coupled with the inference framework [32]. Last but not least, we have shown promising results on data sets far smaller than the practical needs of quite many areas. An interesting but challenging direction is to scale GPSTM up to big data regimes using stochastic variational inference [33] in a cost-effective manner.

Acknowledgments

Authors acknowledge funding by the Dutch Organization for Scientific Research (NWO; grant 612.001.301), thank David M.J. Tax and Julian Kooij for helpful discussions and Manuel Haußmann for proofreading the manuscript.

References

[1] D. Newman, C. Chemudugunta, P. Smyth, M. Steyvers, Analyzing entities and topics in news articles using statistical topic models, in: International Conference on Intelligence and Security Informatics, 2006, pp. 93–104.
[2] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[3] S. Gerrish, D.M. Blei, How they vote: Issue-adjusted models of legislative behavior, in: Advances in Neural Information Processing Systems, 2012, pp. 2753–2761.
[4] L. AlSumait, D. Barbará, C. Domeniconi, On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking, in: International Conference on Data Mining, 2008.
[5] X. Wang, E. Grimson, Spatial latent Dirichlet allocation, Adv. Neural Inf. Process. Syst. (2008) 1577–1584.
[6] B. Liu, L. Liu, A. Tsykin, G.J. Goodall, J.E. Green, M. Zhu, C.H. Kim, J. Li, Identifying functional miRNA–mRNA regulatory modules with correspondence latent dirichlet allocation, Bioinformatics 26 (24) (2010) 3105–3111.
[7] J.D. McAuliffe, D.M. Blei, Supervised topic models, in: Advances in Neural Information Processing Systems, 2008, pp. 121–128.
[8] J. Zhu, A. Ahmed, E.P. Xing, MedLDA: Maximum margin supervised topic models for regression and classification, in: International Conference on Machine Learning, 2009, pp. 1257–1264.
[9] J. Reisinger, A. Waters, B. Silverthorn, R.J. Mooney, Spherical topic models, in: International Conference on Machine Learning, 2010.
[10] S. Lacoste-Julien, F. Sha, M.I. Jordan, DiscLDA: Discriminative learning for dimensionality reduction and classification, Adv. Neural Inf. Process. Syst. (2009) 897–904.
[11] J.D. McAuliffe, D.M. Blei, Supervised topic models, Adv. Neural Inf. Process. Syst. (2008) 121–128.
[12] M.K. Titsias, N.D. Lawrence, Bayesian Gaussian process latent variable model, in: International Conference on Artificial Intelligence and Statistics, 2010, pp. 844–851.
[13] A.C. Damianou, N.D. Lawrence, Deep Gaussian processes, in: International Conference on Artificial Intelligence and Statistics, 2013, pp. 207–215.
[14] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: International Conference on Uncertainty in Artificial Intelligence, 2004, pp. 487–494.
[15] D.M. Mimno, A. McCallum, Topic models conditioned on arbitrary features with Dirichlet-multinomial regression., in: International Conference on Uncertainty in Artificial Intelligence, 2008, pp. 411–418.
[16] J. Zhu, A. Ahmed, E.P. Xing, Medlda: maximum margin supervised topic models, J. Mach. Learn. Res. 13 (2012) 2237–2278.
[17] C. Rasmussen, C. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.
[18] M. Rabinovich, D.M. Blei, The inverse regression topic model., in: International Conference on Machine Learning, 32, 2014, pp. 199–207.
[19] Z. Shi, T.M. Hospedales, T. Xiang, Bayesian joint topic modelling for weakly supervised object localisation., in: International Conference on Computer Vision, 2013, pp. 2984–2991.
[20] C. Zhang, C.H. Ek, X. Gratal, F.T. Pokorny, H. Kjellström, Supervised hierarchical Dirichlet processes with variational inference, in: ICCV Workshop on Inference for Probabilistic Graphical Models, 2013.
[21] Y. Zheng, Y.-J. Zhang, H. Larochelle, A deep and autoregressive approach for topic modeling of multimodal data, IEEE Trans. Pattern Anal. Mach. Intell. 38 (6) (2016) 1056–1069.
[22] Z. Niu, G. Hua, X. Gao, Q. Tian, Semi-supervised relational topic model for weakly annotated image recognition in social media., in: International Conference on Computer Vision and Pattern Recognition, 2014, pp. 4233–4240.
[23] C. Zhang, H. Kjellström, How to supervise topic models, in: ECCV 2014 Workshops, 2015, pp. 500–515.
[24] P. Hennig, D. Stern, R. Herbrich, T. Graepel, Kernel topic models, in: International Conference on Artificial Intelligence and Statistics, Vol. 22, pp. 511–519.
[25] N. Lawrence, Gaussian process latent variable models for visualisation of high dimensional data, Adv. Neural Inf. Process. Syst. 16 (3) (2004) 329–336.
[26] M.D. Hoffman, D.M. Blei, F. Bach, Online learning for latent Dirichlet allocation, Advances in Neural Information Processing Systems, 2010.
[27] J. Hensman, N. Fusi, N.D. Lawrence, Gaussian processes for big data, in: International Conference on Uncertainty in Artificial Intelligence, 2013.
[28] E. Gabrilovich, S. Markovitch, Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with c4.5, in: International Conference on Machine Learning, 2004.
[29] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation, in: International Conference on Computer Vision, 2009, pp. 309–316.
[30] M. Gönen, Coupled dimensionality reduction and classification for supervised and semi-supervised multilabel learning, Pattern Recognit. Lett. 38 (2014) 132–141.
[31] M. Guillaumin, J. Verbeek, C. Schmid, Multimodal semi-supervised learning for image classification, in: International Conference on Computer Vision and Pattern Recognition, 2010, pp. 902–909.
[32] M.K. Titsias, Variational Learning of Inducing Variables in Sparse Gaussian Processes, in: International Conference on Artificial Intelligence and Statistics, 5, 2009, pp. 567–574.
[33] M.D. Hoffman, D.M. Blei, C. Wang, J. Paisley, Stochastic variational inference, J. Mach. Learn. Res. 14 (2013) 1303–1347.
[34] S.-H. Chen, Y.-S. Lee, T.-C. Tai, J.-C. Wang, Gaussian process based text categorization for healthy information, in: International Conference on Orange Technologies, 2015, pp. 30–33.
[35] A. Agovic, A. Banerjee, Gaussian process topic models, in: International Conference on Uncertainty in Artificial Intelligence, 2010.
[36] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, Advances on Neural Information Processing Systems (NIPS), 2015.
[37] Z. Yang, D. Yang, C. Dyer, X. He, A.J. Smola, E.H. Hovy, Hierarchical attention networks for document classification, HLT-NAACL, 2016.
[38] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances on Neural Information Processing Systems (NIPS), 2013.
[39] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, Empirical Methods in Natural Language Processing (EMNLP), 2014. http://www.aclweb.org/anthology/D14-1162.
[40] X. Li, T. Uricchio, L. Ballan, M. Bertini, C.G.M. Snoek, A.D. Bimbo., Socializing the semantic gap: a comparative survey on image tag assignment, Refin. Retr. (2015). arXiv: 1503.08248.
[41] Y. Gong, Y. Jia, T. Leung, A. Toshev, S. Ioffe, Deep convolutional ranking for multilabel image annotation, 2014, arXiv:1312.4894.
[42] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, NUS-WIDE: a real-world web image database from National University of Singapore, in: Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR), 2009.
[43] J. Johnson, L. Ballan, L. Fei-Fei, Love thy neighbors: Image annotation by exploiting image metadata, in: Proceedings of the International Conference on Computer Vision (ICCV), 2015.
[44] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, G. Mori, Learning structured inference neural networks with label relations, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[45] F. Liu, T. Xiang, T. Hospedales, W. Yang, C. Sun, Semantic regularisation for recurrent image annotation", in: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[46] R. Fernandez-Beltran, F. Pla, Latent topics-based relevance feedback for video retrieval, Pattern Recognition 51 (2016) 72–84. (Supplement C).
[47] S. Hou, L. Chen, D. Tao, S. Zhou, W. Liu, Y. Zheng, Multi-layer multi-view topic model for classifying advertising video, Pattern Recognition 68 (2017) 66–81. (Supplement C) ISSN 0031–3203.
[48] P. Hu, W. Liu, W. Jiang, Z. Yang, Latent topic model for audio retrieval, Pattern Recognition 47 (3) (2014) 1138–1143.
[49] X. Li, J. Ouyang, X. Zhou, Centroid prior topic model for multi-label classification, Pattern Recognition Letters 62 (2015) 8–13. (Supplement C).
[50] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, J. Bu, Attribute-restricted latent topic model for person re-identification, Pattern Recognition 45 (12) (2012) 4204–4213.
[51] H. Wu, J. Bu, C. Chen, J. Zhu, L. Zhang, H. Liu, C. Wang, D. Cai, Locally discriminative topic modeling, Pattern Recognition 45 (1) (2012) 617–625. ISSN 0031–3203.
[52] Y. Yun, H.-C. Kim, S.-Y. Shin, J. Lee, A.D. Deshpande, C. Kim, Statistical method for prediction of gait kinematics with gaussian process regression, J. Biomech. 47 (1) (2014) 186–192.
[53] M. Kemmler, E. Rodner, E.-S. Wacker, J. Denzler, One-class classification with Gaussian processes, Pattern Recognition 46 (12) (2013) 3507–3518.
[54] M.F. Huber, Recursive gaussian process: on-line regression and learning, Pattern Recognition Letters 45 (2014) 85–91. (Supplement C).
[55] L. Liu, L. Shao, F. Zheng, X. Li, Realistic action recognition via sparsely-constructed gaussian processes, Pattern Recognition 47 (12) (2014) 3819–3827.

Melih Kandemir received his Ph.D. from Aalto University in 2013 under supervision of Samuel Kaski. He was a postdoctoral researcher at Heidelberg University, Heidelberg Collaboratory for Image Processing (HCI), until 2017. He is currently an assistant professor at Özyeğin University, Department of Computer Science, Istanbul, Turkey. Bayesian modeling and inference, weakly supervised learning, active learning, and application of these techniques to computer vision and medical image analysis problems are among his research interests.

Taygun Kekeç received his B.S. degree in Computer Engineering from Yildiz Technical University, Istanbul, Turkey in 2010. He received his M.S. degree in Mechatronics Engineering from Sabanci University, Istanbul, Turkey, in 2013. He is currently a Ph.D. student at Delft University of Technology, Delft, Netherlands. His research interests in machine learning are unsupervised text analysis, word vector embeddings and approximate inference techniques.

Reyyan Yeniterzi obtained her B.Sc. and M.Sc. from Sabanci University, Department of Computer Science and Engineering in 2007 and 2009, M.Sc. and Ph.D. from School of Computer Science at Carnegie Mellon University in 2012 and 2015. During her studies, in addition to internships at Microsoft Turkey and Google, she also worked as a research visitor at the International Computer Science Institute (ICSI), Vanderbilt and Qatar Carnegie Mellon Universities. She is currently an assistant professor at Özyeğin University, Department of Computer Science, Istanbul, Turkey. Her research interests are natural language processing, text mining and information retrieval.