Method of Moments for Topic Models with Mixed Discrete and Continuous Features

Joachim Giesen[1] , **Paul Kahlmeyer**[1] , **Sören Laue**[1,2] , **Matthias Mitterreiter**[1] , **Frank Nussbaum**[1,3*] , **Christoph Staudt**[1] and **Sina Zarrieß**[1,4]
[1]Friedrich-Schiller-Universität Jena
[2]Data Assessment Solutions GmbH, Hannover
[3]DLR Institute of Data Science, Jena
[4]Universität Bielefeld

Abstract

Topic models are characterized by a latent class variable that represents the different topics. Traditionally, their observable variables are modeled as discrete variables like, for instance, in the prototypical latent Dirichlet allocation (LDA) topic model. In LDA, words in text documents are encoded by discrete count vectors with respect to some dictionary. The classical approach for learning topic models optimizes a likelihood function that is non-concave due to the presence of the latent variable. Hence, this approach mostly boils down to using search heuristics like the EM algorithm for parameter estimation. Recently, it was shown that topic models can be learned with strong algorithmic and statistical guarantees through Pearson's method of moments. Here, we extend this line of work to topic models that feature discrete as well as continuous observable variables (features). Moving beyond discrete variables as in LDA allows for more sophisticated features and a natural extension of topic models to other modalities than text, like, for instance, images. We provide algorithmic and statistical guarantees for the method of moments applied to the extended topic model that we corroborate experimentally on synthetic data. We also demonstrate the applicability of our model on real-world document data with embedded images that we preprocess into continuous state-of-the-art feature vectors.

1 Introduction

Multimodal topic models have applications in many fields, including computer vision [Zheng *et al.*, 2014], natural language processing [Roller and im Walde, 2013], bioinformatics [Liu *et al.*, 2016], and the social sciences [Wang *et al.*, 2019]. Here, we focus on efficient parameter learning for a multimodal extension of the latent Dirichlet allocation (LDA) topic model [Blei *et al.*, 2001] by the method of moments. The method of moments was first used by [Pearson, 1894] for estimating the parameters of a mixture of two univariate Gaussians $x \sim w \cdot \mathcal{N}(\mu_1, \sigma_1^2) + (1-w) \cdot \mathcal{N}(\mu_2, \sigma_2^2)$. Pearson estimated the parameters of this model by matching expected moments of the variable x with empirical moments computed from data. Since the expected moments are polynomials in the five parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$, and w, the method of moments amounts to solving a system of polynomial equations. At his time, Pearson's approach was considered an impressive feat, but deemed too complex to be of further practical use, see [McLachlan and Peel, 2004, Page 3].

The skepticism towards the method of moments was not completely unjustified since solving systems of polynomial equations is, in general, a hard problem. The running times of state-of-the-art algorithms for solving systems of polynomial equations by Gröbner bases [Buchberger, 1976] grows exponentially in the number of variables in the zero-dimensional case, that is, when the system has only a finite number of solutions, see [Hashemi and Lazard, 2011]. Note that in the polynomial system that is derived from the method of moments, the variables are the parameters that need to be estimated.

However, [Hsu and Kakade, 2012] [2013] were able to show that the polynomial system resulting from the method of moments for multivariate mixtures of Gaussians can be solved efficiently, that is, in polynomial time, by spectral methods. Furthermore, they proved that parameter estimation by the method of moments is consistent, that is, true model parameters can be recovered with high probability with a growing number of data samples. Such a statistical guarantee is not known for the alternative maximum likelihood approach. The log-likelihood function for Gaussian mixture models can have arbitrarily many critical points, see [Cerón, 2017]. This poses not only statistical, but also algorithmic challenges. For instance, theoretical guarantees are lacking for solutions computed by search heuristics like the popular EM algorithm [Dempster *et al.*, 1977].

The spectral approach towards the method of moments does not only work for mixture models but also for several discrete mixed membership models, see [Anandkumar *et al.*, 2014a], among them the LDA topic model [Anandkumar *et al.*, 2015]. The difference between mixture models and mixed membership models is subtle. Both families of models have a finite latent class variable. The difference is that every observation in a mixture model is from a single component (topic) of the model, while it is a mixture itself (admixture) in mixed

*Contact Author

membership models. Here, we extend the LDA topic model, which is a mixed membership model, such that it can also accommodate continuous observed features. This enables a smooth transition from the text-only modality to multimodal documents, for instance, text documents with embedded images or images with captions.

LDA has been used directly for topic models on multimodal documents, where images are represented by bags of visual words, that is, by converting continuous visual features into discrete features [Feng and Lapata, 2010]. However, state-of-the art feature representations are continuous, see for instance [He *et al.*, 2016]. This motivates our extension of LDA into a topic model for mixed discrete and continuous features. We keep the discrete part as in standard LDA, especially the Dirichlet prior, and add a Gaussian mixed membership component for the continuous features. Our model mixes t topics by the following generative process for a multi-topic text document and embedded image:

1. Draw the topic proportions $\boldsymbol{h} \sim \text{Dir}(\alpha_1, \ldots, \alpha_t)$
2. Draw a document with l words:
   (a) Draw the number of words per topic $(l_1, \ldots, l_t) \sim \text{Mult}(l, \boldsymbol{h})$ (multinomial distrib.)
   (b) For $i = 1, \ldots, t$ do: Draw the word count vector for the i-th topic as $\boldsymbol{x_i} \sim \text{Mult}(l_i, \boldsymbol{p_i})$
3. Create a single count vector $\boldsymbol{x} = \sum_{i=1}^{t} \boldsymbol{x_i}$
4. For $i = 1, \ldots, t$ do: Draw an image feature vector for the i-th topic as $\boldsymbol{y_i} \sim \mathcal{N}(\boldsymbol{\mu_i}, \sigma_i^2 \boldsymbol{I})$
5. Create a single feature vector $\boldsymbol{y} = \sum_{i=1}^{t} h_i \boldsymbol{y_i}$
6. Output $(\boldsymbol{x}, \boldsymbol{y})$

A sample from this model consists of a discrete count vector $\boldsymbol{x}$ for words from a given dictionary and a continuous image feature vector $\boldsymbol{y}$. The word count vector is as in the classical LDA topic model, while the image feature vector is a convex combination of topic-specific normal distributions. The discrete and the continuous components are drawn independently from each other, but both are governed by the mixing proportions drawn from the Dirichlet prior. Our model is similar to the Corr-LDA model of [Blei and Jordan, 2003], where the continuous components are drawn from a mixture of Gaussians. The Gaussian mixed membership model that we employ here for the continuous part contains a mixture of Gaussians as special case when all the α_i of the Dirichlet prior are much smaller than 1. This allows us to empirically compare the mixture and the mixed membership approach on state-of-the-art image features in Section 5. But before we do so, we focus on our main contribution, namely, deriving the method-of-moments system of equations for our model and proving algorithmic and statistical guarantees for its solution.

2 Method of Moments

The method of moments is a general method for estimating the parameters of probabilistic models. Still, for each model, the expected moments have to be computed in terms of the model parameters, and the system of equations that equates expected moments and their corresponding empirical moments has to be solved. Both are typically non-trivial tasks. Deriving the system of equations for our model requires tedious but standard calculations, which we skip here. In this section, we only summarize the key characteristics of the resulting system of equations. In the following section, we show how to solve the system efficiently.

For the method of moments to succeed, there must be at least as many equations as there are model parameters. Assuming d discrete features (size of the dictionary) and m continuous features, our model has $t \cdot (d+m+2) - 1$ parameters, namely, $t(d+m)$ parameters from the t topic-specific mean vectors $\boldsymbol{\theta_i} = (\boldsymbol{p_i}, \boldsymbol{\mu_i}) \in R^{d+m}$ in addition to t parameters for the topic-specific variances σ_i^2 and $t-1$ parameters for the Dirichlet prior (given that the α_i add up to α_0, which we assume to be a hyperparameter). Hence, in total there are $t(m+d+2) - 1$ model parameters. The second-moment matrix

$$\boldsymbol{M}_2 = E[(\boldsymbol{x}, \boldsymbol{y}) \otimes (\boldsymbol{x}, \boldsymbol{y})],$$

where $\otimes$ denotes the outer product, yields $(d+m)(d+m+1)/2$ equations by symmetry, which exceeds the number of model parameters provided that the number of topics is not larger than $(d+m)/2$.

If $t \leq m$, we can reduce the system of equations from the second moment to $\boldsymbol{T}_2 = \sum_{i=1}^{t} \lambda_i \boldsymbol{\theta_i} \otimes \boldsymbol{\theta_i}$, where $\boldsymbol{T}_2$ is a transformed second-moment matrix that can be computed from moments up to the second order. In principle, the transformed equation system can be solved by a matrix decomposition. However, since the parameter vectors $\boldsymbol{\theta_i}$ are not orthogonal, the solution is not unique. Therefore, we also use the third-moment tensor

$$\boldsymbol{M}_3 = E[(\boldsymbol{x}, \boldsymbol{y}) \otimes (\boldsymbol{x}, \boldsymbol{y}) \otimes (\boldsymbol{x}, \boldsymbol{y})].$$

Here, we can also derive a transformed third-moment tensor that has a tensor decomposition with rank-1 components formed by the parameter vectors $\boldsymbol{\theta_i}$. The following proposition shows how the decompositions of the transformed moments relate to the model parameters.

Proposition 1. *If $t \leq m$, then there exist transformed moments $\boldsymbol{T}_2$ and $\boldsymbol{T}_3$ that can be computed from the observed first, second, and third moments and satisfy the following:*

$$\boldsymbol{T}_2 = \sum_{i=1}^{t} \lambda_i \boldsymbol{\theta_i} \otimes \boldsymbol{\theta_i} \quad \textit{and} \quad \boldsymbol{T}_3 = \sum_{i=1}^{t} c\lambda_i \boldsymbol{\theta_i} \otimes \boldsymbol{\theta_i} \otimes \boldsymbol{\theta_i},$$

where $\lambda_i = \alpha_i / [\alpha_0(\alpha_0+1)]$ and $c = 2/(\alpha_0+2)$. □

The proof of Proposition 1 requires technical algebraic manipulations that we skip here.

Proposition 1 suggests to retrieve the model parameters by performing matrix and tensor decompositions on the transformed moments, or rather their empirical versions. The CP decomposition (decomposition into rank-1 components as in Proposition 1) on tensors is unique under mild conditions [Kruskal, 1977]. Unfortunately, calculating the CP decomposition of a general tensor is NP hard [Hillar and Lim, 2013]. We can bypass this computational difficulty by using $\boldsymbol{T}_2$ to derive a whitening matrix that orthogonalizes the rank-1 components of $\boldsymbol{T}_3$. We describe the exact algorithm for learning the model parameters in the next section.

3 Decomposition Algorithm

Assuming linear independence of the parameter vectors $\boldsymbol{\theta_i}$, the matrix $\boldsymbol{T}_2$ from Proposition 1 is a real positive semidefinite symmetric matrix of rank t. Hence, there exists a rank-t decomposition $\boldsymbol{T}_2 = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\gamma})\boldsymbol{U}^\top$ with a matrix $\boldsymbol{U} \in R^{(d+m)\times t}$ that has orthonormal columns and a vector $\boldsymbol{\gamma} \in R^t$ of positive eigenvalues. The whitening matrix for $\boldsymbol{T}_2$ is given by

$$\boldsymbol{W} = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\gamma}^{-1/2})$$

since it holds that $\boldsymbol{W}^\top \boldsymbol{T}_2 \boldsymbol{W} = \boldsymbol{I}$. Next, set

$$\boldsymbol{\nu_i} = \boldsymbol{W}^\top \lambda_i^{1/2} \boldsymbol{\theta_i}, \quad i = 1, \ldots, t.$$

It can be shown that the vectors $\boldsymbol{\nu_i}$ are orthogonal. Consequently, the tensor

$$\begin{aligned}\boldsymbol{T}_3(\boldsymbol{W}, \boldsymbol{W}, \boldsymbol{W}) &= \sum_{i=1}^{t} c\lambda_i (\boldsymbol{W}^\top \boldsymbol{\theta_i}) \otimes (\boldsymbol{W}^\top \boldsymbol{\theta_i}) \otimes (\boldsymbol{W}^\top \boldsymbol{\theta_i}) \\ &= \sum_{i=1}^{t} \rho_i \, \boldsymbol{\nu_i} \otimes \boldsymbol{\nu_i} \otimes \boldsymbol{\nu_i}\end{aligned}$$

has orthonormal eigenvectors $\boldsymbol{\nu_i}$ along with scaled eigenvalues $\rho_i = c \cdot \lambda_i^{-1/2}$. Using the whitened transformed third moment $\boldsymbol{T}_3(\boldsymbol{W}, \boldsymbol{W}, \boldsymbol{W})$, the NP-hard problem of decomposing a general tensor reduces to the decomposition of an orthogonal tensor. Decompositions of orthogonal tensors can be efficiently computed [Ge *et al.*, 2015]. The original parameters can be retrieved by un-whitening using the pseudo-inverse $\boldsymbol{W}^\dagger = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\gamma}^{1/2})$ of $\boldsymbol{W}^\top$:

$$\lambda_i = c^2/\rho_i^2 \quad \text{and} \quad \boldsymbol{\theta_i} = \rho_i \boldsymbol{W}^\dagger \boldsymbol{\nu_i}/c.$$

The practical algorithm is outlined in Algorithm 1 and also displayed as a flowchart in Figure 1. Here, we have to rely on empirical quantities, which we denote with hats. First, the transformed empirical moments $\widehat{\boldsymbol{T}}_2$ and $\widehat{\boldsymbol{T}}_3$ are derived, for which the empirical raw moments are used. Second, the rank-1 components of $\widehat{\boldsymbol{T}}_3$ are orthogonalized using the empirical whitening matrix $\widehat{\boldsymbol{W}}$, which is obtained from $\widehat{\boldsymbol{T}}_2$. Third, we use the robust tensor power method from [Anandkumar *et al.*, 2014b] to compute the t largest rank-1 components of the empirical whitened tensor $\widehat{\boldsymbol{T}}_3(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{W}}, \widehat{\boldsymbol{W}})$. In the last step, we use the un-whitening operator to obtain the final estimated model parameters. Moreover, the topic-specific variances $\hat{\sigma}_i^2$ can be calculated by solving a linear system of equations.

Importantly, Algorithm 1 runs in polynomial time: The computation of the first two steps requires $\mathcal{O}((d+m)^3)$ time, the robust tensor power method is a polynomial time algorithm, and the linear system of equations can be solved and set up in polynomial time as well.

Note that the robust tensor power method finds the largest components of the whitened tensor only with high probability. Moreover, the empirical inputs of Algorithm 1 lead to errors in the estimation process. In the next section, we show that our method-of-moments estimator enjoys consistency properties nevertheless.

Algorithm 1 Method of moments for mixed topic models

Input: observed feature vectors in R^{d+m}
Output: model parameters $(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}, \hat{\alpha}_i, \hat{\sigma}_i^2)$, $i = 1, \ldots, t$

1: Compute the transformed empirical moments $\widehat{\boldsymbol{T}}_2$ and $\widehat{\boldsymbol{T}}_3$.
2: Compute rank-t SVD $\widehat{\boldsymbol{T}}_2 = \hat{\boldsymbol{U}}\,\mathrm{diag}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{U}}^\top$ and set $\widehat{\boldsymbol{W}} = \hat{\boldsymbol{U}}\,\mathrm{diag}(\hat{\boldsymbol{\gamma}}^{-1/2})$ and $\widehat{\boldsymbol{W}}^\dagger = \hat{\boldsymbol{U}}\,\mathrm{diag}(\hat{\boldsymbol{\gamma}}^{1/2})$.
3: Compute the t largest eigenvalues $\hat{\rho}_i$ and corresponding eigenvectors $\hat{\boldsymbol{\nu}}_{\boldsymbol{i}}$ of the whitened tensor $\widehat{\boldsymbol{T}}_3(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{W}}, \widehat{\boldsymbol{W}})$.
4: For $i = 1, \ldots, t$ set
$\hat{\boldsymbol{\theta}}_{\boldsymbol{i}} = \hat{\rho}_i \widehat{\boldsymbol{W}}^\dagger \hat{\boldsymbol{\nu}}_{\boldsymbol{i}}/c, \quad \hat{\lambda}_i = c^2/\hat{\rho}_i^2, \quad$ and
$\hat{\alpha}_i = \alpha_0(\alpha_0+1)\hat{\lambda}_i$.
Compute $\hat{\sigma}_i^2$ by solving a linear system of equations.

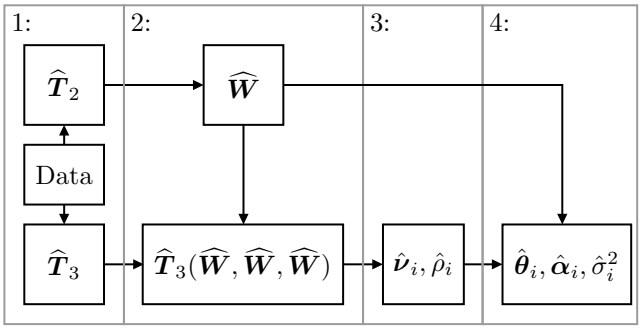

Figure 1: A flowchart visualizing the four steps of Algorithm 1.

4 Consistency

In this section, we investigate when an assumed true model can be recovered from data using the proposed method-of-moments estimator. There are two main sources of error in the estimation process. The first source are errors in the empirical moments. These errors propagate through the calculations of the transformed empirical moments as well as the whitening and un-whitening operators. The second source of error stems from performing the tensor decomposition on the empirical whitened tensor. To control the errors, we begin with a result that shows that small errors on the transformed empirical moments also imply a small error on the empirical whitened tensor.

Lemma 1. *Suppose that the true matrix $\boldsymbol{T}_2$ and tensor $\boldsymbol{T}_3$ are as in Proposition 1 with linearly independent parameter vectors $\boldsymbol{\theta_i}$. Moreover, assume that the errors of the noisy/perturbed versions of the transformed moments satisfy*

$$\|\boldsymbol{E}_2\| \leq \varepsilon_2 \quad \textit{and} \quad \|\boldsymbol{E}_3\| \leq \varepsilon_3,$$

where $\widehat{\boldsymbol{T}}_2 = \boldsymbol{T}_2 + \boldsymbol{E}_2$ and $\widehat{\boldsymbol{T}}_3 = \boldsymbol{T}_3 + \boldsymbol{E}_3$. Let $\sigma_{\min}(\boldsymbol{T}_2)$ be the smallest nonzero singular value of $\boldsymbol{T}_2$ and suppose that $\varepsilon_2 \leq \sigma_{\min}(\boldsymbol{T}_2)/4$. Then, it holds that

$$\|\boldsymbol{T}_3(\boldsymbol{W}, \boldsymbol{W}, \boldsymbol{W}) - \widehat{\boldsymbol{T}}_3(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{W}}, \widehat{\boldsymbol{W}})\| \leq \varepsilon,$$

where for a universal constant C_1 we define

$$\varepsilon = C_1 \left(\frac{\varepsilon_3}{\sigma_{\min}(\boldsymbol{T}_2)^{3/2}} + \rho_{\max} \frac{\varepsilon_2}{\sigma_{\min}(\boldsymbol{T}_2)} \right)$$

with $\rho_{\max} = \max_i \rho_i = \|\boldsymbol{T}_3(\boldsymbol{W}, \boldsymbol{W}, \boldsymbol{W})\|$. □

This lemma can be proven using results from matrix perturbation theory as in [Janzamin *et al.*, 2019]. It becomes apparent from the definition of ε that the error of the whitened tensor stems from both the estimation error of $\boldsymbol{T}_3$ (first term) and the error of the whitening procedure (second term). The next theorem, which is based on [Janzamin *et al.*, 2019, Theorem 3.5], states that the error of the parameters $(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}, \hat{\lambda}_i)$ as estimated by Algorithm 1 are small provided that the error on the whitened tensor is small.

Theorem 1. *In addition to the assumptions of Lemma 1, suppose that*

$$\varepsilon \leq \min\left\{\frac{C_2\rho_{\min}}{t}, \frac{\rho_{\min}}{10}\right\},$$

where $\rho_{\min} = \min_i \rho_i$ and C_2 is a universal constant (the same ε as in Lemma 1 is used here). Moreover, pick any $\eta \in (0,1)$ and assume that the robust tensor power method performs $\Omega(\log(t) + \log\log(\rho_{\max}/\varepsilon))$ power updates and assume that it is restarted poly$(t)\log(1/\eta)$ times for each eigenvalue/eigenvector pair. Under these assumptions, Algorithm 1 estimates pairs $(\hat{\boldsymbol{\theta}}_{\boldsymbol{1}}, \hat{\lambda}_1), \ldots, (\hat{\boldsymbol{\theta}}_{\boldsymbol{t}}, \hat{\lambda}_t)$ in polynomial time such that with probability at least $1-\eta$ there exists a permutation π on $[t] = \{1, \ldots, t\}$ for which it holds for all $i \in [t]$ that

$$\left|\lambda_{\pi(i)} - \hat{\lambda}_i\right| \leq \frac{4\lambda_{\pi(i)}^2}{c^2}\left(25\varepsilon + 10c\lambda_{\pi(i)}^{-1/2}\right)\varepsilon \qquad \textit{and}$$

$$\left\|\boldsymbol{\theta}_{\boldsymbol{\pi(i)}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\right\| \leq \frac{1}{c}\left(45\lambda_{\pi(i)}^{1/2}\varepsilon + 14\right)\|\boldsymbol{T}_2\|^{1/2}\varepsilon. \qquad \square$$

The error bounds of Theorem 1 can be understood as relative error bounds. For example, the error bound for the estimation error of $\hat{\lambda}_i$ is given relative to the size of the matching eigenvalue $\lambda_{\pi(i)}$. Similarly, the spectral norm $\|\boldsymbol{T}_2\|$ can be seen as an upper estimate of the scale of the parameters $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$. Moreover, we point out that the error bounds become arbitrarily close to zero as $\varepsilon_2 \to 0$ and $\varepsilon_3 \to 0$ in Lemma 1 since then also $\varepsilon \to 0$. Observe that Theorem 1 immediately implies that the errors on the Dirichlet parameters α_i are also small since by Proposition 1 they differ from λ_i only by the constant factor $\alpha_0(\alpha_0+1)$.

At this point, we briefly discuss some of the assumptions of Theorem 1 and why they are necessary. First, ε is required to be small in terms of $\rho_{\min}$ to ensure success of the orthogonal tensor decomposition of the whitened tensor. Here, intuitively, the perturbation may not exceed the smallest component with eigenvalue $\rho_{\min}$ of the whitened tensor because otherwise recovery becomes impossible. The required bound on ε becomes slightly stronger when more topic components need to be recovered.

The next assumptions concern the robust tensor power method. The power method iteratively generates pairs of eigenvalues and corresponding eigenvectors. The restarts guarantee that at each step the largest eigenvalue is found with high probability. The lower bound on the number of power iteration updates ensures convergence to an eigenvalue/eigenvector pair.

Based on the previous results, the next theorem shows that the method of moments is a consistent method for estimating the model parameters.

Theorem 2. *Assume that Algorithm 1 is provided with data drawn from a mixed-domain topic model with parameters $(\boldsymbol{\theta}_{\boldsymbol{i}}, \alpha_i, \sigma_i^2)$, $i \in [t]$ (for given α_0 and number of words l). Then, Algorithm 1 yields consistent parameter estimates $(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}, \hat{\alpha}_i, \hat{\sigma}_i^2)$, $i = 1, \ldots, t$. More precisely, the estimated parameters converge in probability to the true model parameters, that is, for any $\delta > 0$ the probabilities of the events*

$$\|\boldsymbol{\theta}_{\boldsymbol{\pi(i)}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\| > \delta, |\alpha_{\pi(i)} - \hat{\alpha}_i| > \delta, \textit{ and } |\sigma_{\pi(i)}^2 - \hat{\sigma}_i^2| > \delta$$

converge to zero as the sample size grows to infinity. $\square$

The proof of Theorem 2 proceeds by first showing that the transformed moments $\widehat{\boldsymbol{T}}_2$ and $\widehat{\boldsymbol{T}}_3$ converge in probability to $\boldsymbol{T}_2$ and $\boldsymbol{T}_3$, respectively. In conjunction with Lemma 1 and Theorem 1, this leads to convergence in probability of $\hat{\lambda}_i$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$. The convergence in probability of the remaining model parameters can then be easily shown. In the next section, we verify the consistency properties of our method-of-moments estimator experimentally.

5 Experiments

First, we conduct experiments on synthetic data to corroborate our theoretical findings. Then, we show on real-world data how adding images to word topic models changes the learned topics. All experiments were run on a machine with an Intel Core i9-10980XE processor with 18 cores, 128GB RAM using Python 3.8 and PyTorch 1.6.

5.1 Synthetic Data

By performing experiments on synthetic data, we want to show that the proposed method of moments can recover true model parameters effectively and efficiently. For this we define models and respectively try to recover the true known model parameters from generated samples.

Setup. In our experiments, a model configuration is defined by the number of topics t, the hyperparameter $\alpha_0 = \sum_{i=1}^{t}\alpha_i$ that controls the Dirichlet prior, the numbers of discrete and continuous features d and m, and the number l of words per document. Here, we can only sample the five-dimensional configuration space. For our experiments with synthetic data, we consider models with $t \in \{5, 10, 20\}$ topics. Apart from that, we choose a base configuration with values $\alpha_0 = 5$, $m = 128$, $d = 500$, and $l = 50$ for the remaining parameters. Starting from this base configuration, we respectively vary one of the parameters α_0, m, d, or l, while keeping the other parameters fixed.

For each configuration, we sample ten random models. Dirichlet parameters $\alpha_i \in R_+$ are drawn from the uniform distribution $\mathcal{U}(0,1)$ and scaled to sum up to α_0. Feature mean vectors $\boldsymbol{\mu}_{\boldsymbol{i}} \in R^m$ are drawn from the uniform distribution $\mathcal{U}(-10, 10)$ while it is made sure that these vectors are pairwise independent. Corresponding variances $\sigma_i^2 \in R_+$ are drawn from $\mathcal{U}(0,2)$. For natural language topics, the frequency of any word is typically inversely proportional to its rank in the frequency table [Cohen *et al.*, 1997], which is known as "Zipf's law". Hence, we sample word probability vectors $\boldsymbol{p}_{\boldsymbol{i}} \in R_+^d$ that adhere to Zipf's law.

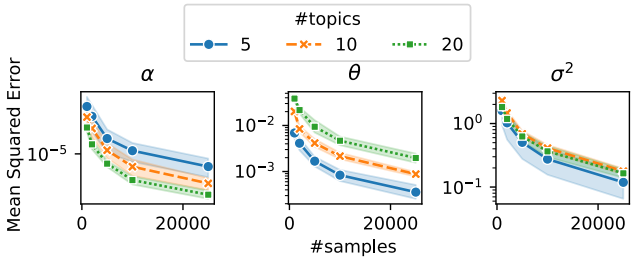

Figure 2: Synthetic Experiment for base configuration. The plots show the MSE on the different parameters that are estimated along with a 95% confidence interval over ten different models. From left to right: MSE_α, MSE_θ and MSE_{σ^2}.

Effectiveness. We consider Algorithm 1 effective if it successfully estimates the true model parameters, where as a consequence of Theorem 2 in Section 4 we expect better recovery as the number of samples increases. For quantifying the estimation quality in our experiments, we use mean squared errors (MSEs). For the variances and parameters of the Dirichlet distribution, we use the MSEs of the vectors $(\sigma_i^2)_{i\in[t]}$ and $(\alpha_i)_{i\in[t]}$, respectively comparing them against their estimated counterparts. For the parameter vectors $\boldsymbol{\theta_i}$, we use $1/t \cdot \sum_{i=1}^{t} \text{MSE}(\boldsymbol{\theta_i}, \hat{\boldsymbol{\theta}}_{\boldsymbol{i}})$, where $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$ is the estimate of $\boldsymbol{\theta_i}$. We denote the different MSEs with MSE_{σ^2}, MSE_α, and MSE_θ. As the estimated model parameters are permuted over the topics, we match estimated parameters and true parameters using the Kuhn-Munkres algorithm [Munkres, 1957] with matching cost MSE_θ.

To confirm the consistency claims of Theorem 2, we track the MSEs for an increasing number of samples. The experimental results for the base configuration with respectively $t = 5, 10, 20$ topics are shown in Figure 2. All plots in Figure 2 show the mean MSEs with 95% confidence intervals computed from the respective populations of ten sampled models for each configuration. The results show that the MSEs quickly decrease towards zero as more samples are used, confirming the theoretical consistency result from Theorem 2. The same can be observed for all experimental configurations, which indicates the robustness of these results.

Efficiency. A second theoretical claim from Theorem 1 is that the method-of-moments system of equations can be solved efficiently by Algorithm 1 from Section 3. To validate this claim, we recorded the training times for different model configurations. For each configuration, we tracked the median running time of Algorithm 1 over a population of ten models. It turns out that the computation time mainly depends on the number of topics t, that is, it is almost independent from all other parameters. For instance, the running times for the base configuration and $t = 10$ topics are $5.1s$ and $6.9s$ for 10 000 and 25 000 samples, respectively. For $t = 20$ topics, the running times are $19.6s$ and $22.4s$, respectively. Hence, training can be done efficiently.

5.2 Real-World Data

In order to show the real-world applicability of our model, we use a data set proposed by [Krause *et al.*, 2017], which contains 19 511 RGB images along with single paragraphs of text describing their content. In our experiments, first, we examine the quality of the learned topics. Second, we compare the learned topics of a purely discrete (LDA) model with the topics from our mixed-domain extension. Third, we shed light on the significance of the hyperparameters t and α_0. Finally, we briefly evaluate the computational performance. However, before we can learn models, some preprocessing of the data set is necessary.

Setup. For the discrete part, the text paragraphs were lemmatized and stop words as well as words that occur in less than ten documents or in more than 50% of the documents were removed. Each paragraph was then translated into a count vector over the remaining dictionary with $d = 2\,237$ words, resulting in on average 21.18 remaining words per paragraph. For the continuous part, the images were preprocessed using the first 17 layers of a pretrained ResNet18 torchvision model [He *et al.*, 2016], yielding continuous feature vectors with $m = 512$ dimensions.

We tested the proposed method of moments (Algorithm 1) on the data set by learning models with $t \in \{10, 15\}$ topics and hyperparameter values $\alpha_0 \in \{1, 50\}$. For each combination of α_0 and t, we trained a purely discrete (LDA) topic model and our extended mixed-domain topic model.

Quality of the topics. The quality of topics is still best evaluated by human inspection. Hence, we validate the discrete part by showing the words corresponding to the seven highest word probabilities for each topic. Examples are shown in Tables 1 and 2. Representing the continuous part is a little more involved. Samples from the continuous part of our model are mixtures of Gaussians: The i-th topic/mixture component has mean vector $\boldsymbol{\mu_i}$ and mixing proportion given by the i-th entry h_i of the vector $\boldsymbol{h}$. Hence, for representing the continuous part of the topics, we solve the inference problem $\arg\max_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{y})$ for each image feature vector $\boldsymbol{y}$ in the data set. Then, for each topic i, we show the five images for which the inference problem yields the highest mixing proportions h_i, see Figure 3.

Topic	Most likely words (discrete topic model)
1	black, woman, wear, dog, person, cat, stand
2	man, wear, shirt, black, stand, blue, hold
3	table, plate, sit, pizza, food, glass, bowl
4	train, track, yellow, red, platform, blue, tree
5	giraffe, tree, grass, stand, field, zebra, elephant
6	water, boat, wave, people, sky, blue, small
7	wall, toilet, sit, cat, black, bed, room
8	plane, sky, blue, airplane, cloud, fly, red
9	building, clock, tower, large, sign, street, window
10	street, bus, sign, building, car, red, road

Table 1: The seven most likely words for each topic of the learned purely discrete topic model with $t = 10$ topics and hyperparameter $\alpha_0 = 1$ (topics are permuted to ease comparison with Table 2).

The results shown in Table 1, Table 2, and Figure 3 have been obtained for $t = 10$ topics and $\alpha_0 = 1$. It can be seen that for each topic, by human judgement, the corresponding

Topic	Most likely words (mixed topic model)
1	snow, ski, black, wear, man, person, tree
2	man, wear, black, tennis, shirt, baseball, blue
3	plate, table, sit, pizza, food, brown, red
4	train, bus, track, street, red, building, car
5	elephant, tree, grass, stand, brown, giraffe, zebra
6	water, blue, man, sky, wave, plane, boat
7	wall, sit, black, toilet, room, table, bed
8	plane, black, blue, sky, red, airplane, motorcycle
9	building, sign, street, tree, clock, sky, red
10	man, wear, black, people, woman, stand, elephant

Table 2: The seven most likely words for each topic of the learned mixed topic model with $t = 10$ topics and hyperparameter $\alpha_0 = 1$.

most probable words are from the same domain–both for the purely discrete and our extended mixed model. Moreover, for the mixed model, the word topics match the most representative images well.

Effect of the multimodal extension. Next, we briefly discuss the differences between the purely discrete (LDA) model and our mixed-domain model extension. For our learned models, it can be observed that many topics remain similar, compare the words from Topics 2 to 8 in Tables 1 and 2. On the other hand, the visual information from the images can also lead to new topics, such as, Topic 1. The images for Topic 1 in Figure 3, which all show skiing people in snow environments, demonstrate that this new topic makes sense. Hence, adding the image modality to LDA models can lead to new insights.

Robustness. We also probed the influence of the choices for the hyperparameters t and α_0 on the learned topics. Here, we can only report high-level trends due to space constraints,

Increasing t for the mixed model does not seem to remove topics: For instance, increasing the number of topics from $t = 10$ to $t = 15$, while keeping $\alpha_0 = 50$ fixed, did not remove any topics. The five additional topics are either completely new or result from splitting existing topics. This indicates a certain robustness of our model under the choice of t for real-world data, where the true number of topics is unknown.

The choice of the hyperparameter α_0 impacts the nature of our model. If we set α_0 to a small value in comparison to t, then we force most of the values of α_i to be smaller than one. In this case, our model gets close to an ordinary mixture model, where each sample mostly belongs to one topic. For instance, setting $\alpha_0 = 1$ with $t = 15$ leads to unspecific topics. There are fewer unspecific topics for $\alpha_0 = 50$ with $t = 15$. In general, the mixed membership model performs better (in terms of human judgment) on our continuous deep learning features than the mixture model.

Computational aspects. Finally, we briefly touch on computational performance. As we have already observed on synthetic data, also here the running times hardly depend on α_0 as well. The dependency on the number of topics t can be seen in Table 3.

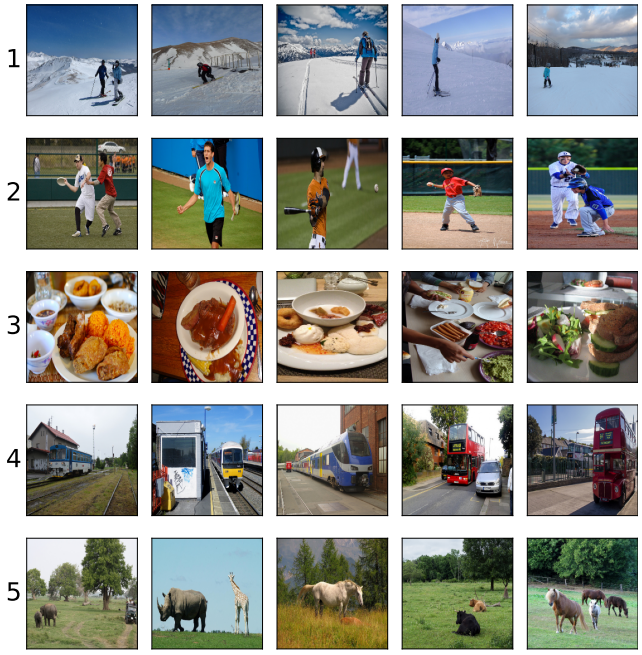

Figure 3: The five most representative images for the first five topics learned for the mixed topic model with $\alpha_0 = 1, t = 10$. Each row represents a topic. For each topic, the images with the highest mixing proportions for the topic are shown.

t	2	5	10	15	20	25
discrete	2.66	3.86	6.53	9.09	20.24	27.7
mixed	7.65	12.49	22.69	30.92	47.02	59.47

Table 3: Running times (in seconds) of Algorithm 1 on the real-world data with $\alpha_0 = 50$ and different numbers of topics t for estimating discrete (LDA) and mixed topic models.

6 Conclusions

We have extended the classical LDA topic model such that it also accommodates continuous features. The continuous features are modeled by mixed membership Gaussians. The parameters of the extended model can be learned in polynomial time and with statistical consistency guarantees. Such a combination of guarantees is not known for the competing maximum likelihood approach. We used synthetic data to experimentally corroborate the theoretical guarantees. Additionally, experiments on a mixed real-world data set with text and images, which we processed into continuous state-of-the-art image features, show that the mixed membership model gives qualitatively better results than a standard mixture model.

Acknowledgments

This work was supported by the German Science Foundation (DFG) grant (GI-711/5-1) within the priority program (SPP 1736) “Algorithms for Big Data”, and by the Carl Zeiss Foundation within the project “A Virtual Werkstatt for Digitization in the Sciences”.

References

[Anandkumar *et al.*, 2014a] Animashree Anandkumar, Rong Ge, Daniel J. Hsu, and Sham M. Kakade. A tensor approach to learning mixed membership community models. *Journal of Machine Learning Research*, 15(1):2239–2312, 2014.

[Anandkumar *et al.*, 2014b] Animashree Anandkumar, Rong Ge, Daniel J. Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[Anandkumar *et al.*, 2015] Anima Anandkumar, Dean P. Foster, Daniel J. Hsu, Sham M. Kakade, and Yi-Kai Liu. A Spectral Algorithm for Latent Dirichlet Allocation. *Algorithmica*, 72(1):193–214, 2015.

[Blei and Jordan, 2003] David M. Blei and Michael I. Jordan. Modeling annotated data. In *Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 127–134, 2003.

[Blei *et al.*, 2001] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 601–608, 2001.

[Buchberger, 1976] Bruno Buchberger. A theoretical basis for the reduction of polynomials to canonical forms. *SIGSAM Bulletin*, 10(3):19–29, 1976.

[Cerón, 2017] Carlos Enrique Améndola Cerón. *Algebraic Statistics of Gaussian Mixtures*. PhD thesis, Technische Universität Berlin, 2017.

[Cohen *et al.*, 1997] A. Cohen, R. Mantegna, and S. Havlin. Numerical Analysis of Word Frequencies in Artificial and Natural Language Texts. *Fractals*, 5:95–104, 1997.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[Feng and Lapata, 2010] Yansong Feng and Mirella Lapata. Topic Models for Image Annotation and Text Illustration. In *Proceedings of Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, pages 831–839, 2010.

[Ge *et al.*, 2015] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping From Saddle Points - Online Stochastic Gradient for Tensor Decomposition. In *Proceedings of The Conference on Learning Theory (COLT)*, pages 797–842, 2015.

[Hashemi and Lazard, 2011] Amir Hashemi and Daniel Lazard. Sharper Complexity Bounds for Zero-Dimensional Gröbner Bases and Polynomial System Solving. *Algebra and Computation*, 21(5):703–713, 2011.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[Hillar and Lim, 2013] Christopher J. Hillar and Lek-Heng Lim. Most Tensor Problems Are NP-Hard. *Journal of the ACM*, 60(6):45:1–45:39, 2013.

[Hsu and Kakade, 2012] Daniel J. Hsu and Sham M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. *CoRR*, abs/1206.5766, 2012.

[Hsu and Kakade, 2013] Daniel J. Hsu and Sham M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, pages 11–20, 2013.

[Janzamin *et al.*, 2019] Majid Janzamin, Rong Ge, Jean Kossaifi, and Anima Anandkumar. Spectral learning on matrices and tensors. *Foundations and Trends in Machine Learning*, 12(5-6):393–536, 2019.

[Krause *et al.*, 2017] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A Hierarchical Approach for Generating Descriptive Image Paragraphs. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 3337–3345, 2017.

[Kruskal, 1977] Joseph B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95–138, 1977.

[Liu *et al.*, 2016] Lin Liu, Lin Tang, Wen Dong, and Wei Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5:1608, 2016.

[McLachlan and Peel, 2004] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[Munkres, 1957] James Munkres. Algorithms for the assignment and transportation problems. *Journal of The Society for Industrial and Applied Mathematics*, 10:196–210, 1957.

[Pearson, 1894] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, 185:71–110, 1894.

[Roller and im Walde, 2013] Stephen Roller and Sabine Schulte im Walde. A Multimodal LDA Model integrating Textual, Cognitive and Visual Modalities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1157, 2013.

[Wang *et al.*, 2019] Kai Wang, Weiyi Meng, Shijun Li, and Sha Yang. Multi-Modal Mention Topic Model for mentionee recommendation. *Neurocomputing*, 325:190–199, 2019.

[Zheng *et al.*, 2014] Yin Zheng, Yu-Jin Zhang, and Hugo Larochelle. Topic Modeling of Multimodal Data: An Autoregressive Approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1370–1377, 2014.