Model Selection for Topic Models via Spectral Decomposition

Dehua Cheng* dehua.cheng@usc.edu University of Southern California

Xinran He* xinranhe@usc.edu University of Southern California

Yan Liu yanliu.cs@usc.edu University of Southern California

Abstract

Topic models have achieved significant successes in analyzing large-scale text corpus. In practical applications, we are always confronted with the challenge of model selection, i.e., how to appropriately set the number of topics. Following the recent advances in topic models via tensor decomposition, we make a first attempt to provide theoretical analysis on model selection in *latent Dirichlet allocation*. With mild conditions, we derive the upper bound and lower bound on the number of topics given a text collection of finite size. Experimental results demonstrate that our bounds are correct and tight. Furthermore, using *Gaussian mixture model* as an example, we show that our methodology can be easily generalized to model selection analysis in other latent models.

1 Introduction

Recently topic models, such as latent Dirichlet allocation (LDA) [BNJ03] and its variants [TJBB06], have been proven extremely successful in modeling large, complex text corpus. These models assume that the words in a document are generated from a mixture of latent topics represented by multinomial distributions. Therefore, the major inference problem becomes recovering latent topics from text corpus. Popular inference algorithms for LDA include variational inference [BNJ03, TKW07, WPB11, HBWP13], sampling methods [GS04, PNI$^+$08], and the recent development via tensor decomposition [AGM12, AFH$^+$12, AGH$^+$12]. However, all of them require that the number of topics K is given as input.

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

It is known that model selection, i.e., choosing the appropriate number of topics K plays a vital role in successfully applying LDA models [TMN$^+$14, KRS14]. For example, [TMN$^+$14] has shown that choosing K too large leads to severe deterioration in the learning rate; [KRS14] points out that incorrect number of mixture components can result in unbounded error when estimating parameters of mixture model with spectral method. Moreover, as K increases, the computational cost of inference for the LDA model grows significantly.

Unfortunately it is extremely challenging to choose the right number of topics for the LDA model. The traditional method for Bayesian models selection, marginal model likelihood, is generally intractable for LDA model. In practice, [Tad12] approximates the marginal likelihood via Laplace's method, while [AEF$^+$10, GS04] computes the likelihood via MCMC. Moreover, [Tad12] provides another model selection method by analysis of residuals. However, it only provides rough measures for evidence in favor of a larger K. Other model selection criteria, such as AIC [Aka74], BIC [S$^+$78] and cross validation can be applied. Though achieving practical success [AEF$^+$10], they only have asymptotic consistency. Moreover, they require multiple runs of the learning algorithm with different K, which limits the practicality of the approaches to large-scale datasets. On the other hand, Bayesian nonparametrics, such as *Hierarchical Dirichlet Processes* (HDP)[TJBB06], provide alternatives to select K in a principled way. However, it has been shown in a recent paper [MH13] that HDP is inconsistent for estimating the number of topics of LDA model even with infinite amount of data.

In this paper, we provide theoretical analysis on the number of topics for latent topic models. By the results from Anandkumar et al. [AGH$^+$12], the second-order moment of LDA follows a special structure as the summation over the outer product of the topic vectors. We show that a spectral decomposition on the

*Dehua Cheng and Xinran He contributed equally to this article.

second-order empirical moment with proper thresholding on the singular values can lead to the correct number of topics. With mild assumptions, we show that our thresholding provides both the lower bound and the upper bound on number of topics K in the LDA model. To the best of our knowledge, this is the first work to utilize such connection explicitly to analyze the number of topics with provable guarantee. Moreover, we show that our methodology can be generalized naturally to analyzing the number of mixture components in other mixture models, for example the *Gaussian Mixture Model* (GMM).

Our main contributions are:

(1) We analyze the empirical second-order moment of LDA model and derive an upper bound on its variance in terms of the corpus statistics, i.e., the number of documents, the length of each document and the number of unique words. Essentially, our results provide an informative and computable guideline to the convergence of second-order moment, which can be of its own practical value, e.g., determining the correct downsampling rate on a large-scale dataset.

(2) We analyze the spectral structure of the expected second-order moment of LDA model. That is, we provide the spectral information on the covariance of *Dirichlet* design matrix.

(3) Based on the results on empirical and expected second-order moment of LDA model, we derived three inequalities involving the number of topics K, which in turn provide both upper and lower bounds on K without unknown parameters or constants. We also present the simulation study for our theoretical results.

(4) We show that our results and techniques can be generalized to other mixture models, where the results on *Gaussian mixture models* is presented as an example.

The rest of the paper is organized as follows: In section 2, we present our main result on how to analyze the number of topics in the LDA model. We carry out experiments on the synthetic datasets with different settings to demonstrate the validity and tightness of our bounds in section 3. We conclude the paper and show how our methodology generalizes to other mixture models in section 4.

2 Analyze the Number of Topics in LDA

Latent Dirichlet Allocation [BNJ03] (LDA) is a powerful generative model for topic modeling. It has been applied to a variety of applications and also serves as building blocks in other powerful models. Most existing methods follow the Bayesian inference principle to estimate the parameters of the model [BNJ03, TKW07, GS04, PNI$^+$08]. Recently, method of moments has been explored, leading to a series of interesting work and insight into the LDA model from a traditional yet brand new perspective. It has been shown in [AFH$^+$12, AGH$^+$12] that the latent topics can be directly derived from the properly constructed third-order moment (which can be directly estimated from the data) by orthogonal tensor decomposition. Following this line of work, we observe that the low-order moments are also useful for discovering the number of topics in the LDA model, due to their close connection. In this section, we will investigate the structure of both empirical and expected second-order moment, and show that they lead to both upper and lower bound on the number of topics.

2.1 Notation and Problem Formulation

We first introduce the notation for our later discussion. As introduced in [BNJ03], the full generative process for the d-th document in the LDA model is described as follows:

1. Generate the topic mixing $\mathbf{h}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$.
2. For each word $l = 1, \ldots, L$ in document d:
   (a) Generate a topic $z_{d\ell} \sim \mathrm{Multi}(\mathbf{h}_d)$, where $\mathrm{Multi}(\mathbf{h}_d)$ denotes the multinomial distribution.
   (b) Generate a word $\mathbf{x}_{d\ell} \sim \mathrm{Multi}(\boldsymbol{\mu}_{z_{d\ell}})$, where $\boldsymbol{\mu}_{z_{d\ell}}$ is the multinomial parameter associated with topic $z_{d\ell}$.

The notation is summarized in Table 1. $\mathbf{x}_{d\ell}$ is represented by natural basis $\mathbf{e}_v$, meaning that the ℓ-th word in d-th document is the v-th word in the dictionary.

In [AGH$^+$12], the authors proposed the method of moment for the LDA model, where the empirical first-order moment $\hat{\mathbf{M}}_1$ is defined as

$$\hat{\mathbf{M}}_1 = \frac{\sum_d \sum_\ell \mathbf{x}_{d\ell}}{DL},$$

and the empirical second-order moment $\hat{\mathbf{M}}_2$ as

$$\hat{\mathbf{M}}_2 = \frac{\sum_d \sum_{\ell \neq \ell'} \mathbf{x}_{d\ell} \otimes \mathbf{x}_{d\ell'}}{DL(L-1)} - \frac{\alpha_0}{\alpha_0 + 1} \hat{\mathbf{M}}_1 \otimes \hat{\mathbf{M}}_1,$$

Table 1: Notation for LDA

Notation	Definition
D (d)	Number(index) of documents
L (ℓ)	Number(index) of words in a document
V (v)	Number(index) of unique words
K (k)	Number(index) of latent topics
$\boldsymbol{\mu}_k$	Multinomial parameters for the k-th topic
$\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$	Collection of all topics
$\boldsymbol{w_d} = \{\mathbf{x}_{d\ell}\}_{\ell=1}^{L}$	Collection of all words in d-th document
$\mathbf{x}_{d\ell}$	ℓ-th word in d-th document
$\mathbf{h}_d$	Topic mixing for d-th document
$z_{d\ell}$	Topic assignment for word $\mathbf{x}_{d\ell}$
$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top$	Hyperparameter for document topic distribution
$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_V)^\top$	Hyperparameter for generating topics

where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ and the outer product $\mathbf{x} \otimes \mathbf{x} := \mathbf{x}\mathbf{x}^\top$ for any column vector $\mathbf{x}$. Then we define the first-order and second-order moments as the expectation of the empirical moments, i.e., $\mathbf{M}_1 = \mathbb{E}[\hat{\mathbf{M}}_1]$ and $\mathbf{M}_2 = \mathbb{E}[\hat{\mathbf{M}}_2]$ respectively. Furthermore, it has been shown in [AGH+12] that $\mathbf{M}_2$ equals the weighted sum of the outer products of the topic parameter $\boldsymbol{\mu}$, i.e.,

$$\mathbf{M}_2 = \sum_{k=1}^{K} \frac{\alpha_k}{(\alpha_0 + 1)\alpha_0} \boldsymbol{\mu}_k \otimes \boldsymbol{\mu}_k.$$

It implies that the rank of $\mathbf{M}_2$ is exactly the number of topics K. Another interesting observation from this derivation is that since $\mathbf{M}_2$ is the summation of K rank-1 matrices and all the topics $\boldsymbol{\mu}_k$ are linearly independent almost surely under our full generative model, we have the K-th largest singular value $\sigma_K(\mathbf{M}_2) > 0$ and K+1-th largest singular value $\sigma_{K+1}(\mathbf{M}_2) = 0$. Therefore, the number of non-zero singular values of $\mathbf{M}_2$ is exactly the number of topics, which provides a way to estimate K under the noiseless scenario. However, in practice, we only have access to the estimated $\hat{\mathbf{M}}_2$ as an approximation to the true second-order moment $\mathbf{M}_2$. As a result, the rank of $\hat{\mathbf{M}}_2$ may not be K and $\sigma_{K+1}(\hat{\mathbf{M}}_2)$ may be larger than zero. To overcome this obstacle, we need to study (1) the spectral structure of $\mathbf{M}_2$, and (2) the relationship between $\mathbf{M}_2$ and its estimator $\hat{\mathbf{M}}_2$.

2.2 Solution Outline

The second-order moment $\mathbf{M}_2$ can be estimated directly from the observations, without inferring the topic mixing and estimating parameters. Our idea follows that when the sample size becomes large enough, $\hat{\mathbf{M}}_2$ can approximate $\mathbf{M}_2$ well enough, i.e., $\sigma_{K+1}(\hat{\mathbf{M}}_2)$ is very close to zero while $\sigma_K(\hat{\mathbf{M}}_2)$ is bounded away from zero. Then, by picking a proper threshold θ satisfying $\sigma_{K+1}(\hat{\mathbf{M}}_2) < \theta < \sigma_K(\hat{\mathbf{M}}_2)$, we can obtain the value of K by simply counting the number of singular values of $\hat{\mathbf{M}}_2$ greater than θ. In order to justify our idea, we need to achieve two subtasks: (1) examine the convergence rate of the singular values of $\hat{\mathbf{M}}_2$ when increasing the number of observations; (2) investigate how the spectral structure of $\mathbf{M}_2$ is related to the model parameters, thus providing a lower bound for $\sigma_K(\mathbf{M}_2)$.

We will provide details of our theoretical results on these two subtasks in the following sections.

2.3 Convergence of $\hat{\mathbf{M}}_2$

Without loss of generality, we assume that both $\boldsymbol{h}_k$ and $\boldsymbol{\mu}_k$ are generated from symmetrical Dirichlet distribution, namely $\alpha_k = \alpha$ for $k = 1, \ldots, K$ and $\beta_v = \beta$ for $v = 1, \ldots, V$. We also assume that all documents have the same length L for simplicity. Since $\hat{\mathbf{M}}_2$ is an unbiased estimator of $\mathbf{M}_2$ by definition, we can bound the difference between the singular values of $\hat{\mathbf{M}}_2$ and $\mathbf{M}_2$ by bounding their variance as follows:

Theorem 2.1. *For the LDA model, with probability at least $1 - \delta$, we have*

$$|\sigma_i(\hat{\mathbf{M}}_2) - \sigma_i(\mathbf{M}_2)| \leq \delta_{\mathbf{R}}, 1 \leq i \leq V$$

where $\delta_{\mathbf{R}} = \frac{1}{\sqrt{D\delta}} \sqrt{\frac{2}{L^2} + \frac{2}{V^2} + \mathcal{O}(\epsilon)}$, ϵ represents higher-order terms.

Especially, when $i \geq K + 1$, we have

$$\sigma_i(\hat{\mathbf{M}}_2) \leq \delta_{\mathbf{R}}. \tag{1}$$

Proof. Let $\mathbf{R} = \mathbf{M}_2 - \hat{\mathbf{M}}_2$ and $||\mathbf{R}||_2, ||\mathbf{R}||_{\mathrm{F}}$ be the spectral and Frobenius norm of $\mathbf{R}$, respectively. We denote $\lambda_i(\mathbf{M})$ as the i-th largest eigenvalue of matrix $\mathbf{M}$. We establish the result through the following chain

of inequalities:

$$\begin{aligned}\max_i |\sigma_i(\hat{\mathbf{M}}_2) - \sigma_i(\mathbf{M}_2)| &\overset{(1)}{\leq} \max_i |\lambda_i(\hat{\mathbf{M}}_2) - \lambda_i(\mathbf{M}_2)| \\ &\overset{(2)}{\leq} ||\mathbf{R}||_2 \\ &\overset{(3)}{\leq} ||\mathbf{R}||_{\mathrm{F}}.\end{aligned}$$

Inequality (1) follows the semi-definiteness of matrix $\mathbf{M}_2$ and the symmetry of matrix $\hat{\mathbf{M}}_2$. The detailed proof is deferred to Lemma A.1 in Appendix. (2) and (3) are well-known results on matrix norm and matrix perturbation theory [HJ]. And in Lemma 2.2, we provide essential upper bound on the Frobenius norm of matrix $\mathbf{R}$. Because $\mathrm{Rank}(\mathbf{M}_2) \leq K$, i.e., $\sigma_i(\mathbf{M}_2) = 0$ for $i \geq K+1$, the second statement holds true. □

Lemma 2.2. *For the LDA model, with probability at least* $1-\delta$*, we have* $||\mathbf{R}||_{\mathrm{F}} \leq \delta_{\mathbf{R}}$.

Proof. We first compute the expectation $\mathbb{E}[||\mathbf{R}||_{\mathrm{F}}^2]$ and then use Markov inequality to complete the proof. The square of Frobenius norm is $||\mathbf{R}||_{\mathrm{F}}^2 = \sum_{i,j} \mathbf{R}_{ij}^2$. Since we have $\mathbb{E}[\mathbf{R}_{ij}|\boldsymbol{\mu}] = 0$, so $Var[\mathbf{R}_{ij}|\boldsymbol{\mu}] = \mathbb{E}[\mathbf{R}_{ij}^2|\boldsymbol{\mu}] - \mathbb{E}^2[\mathbf{R}_{ij}|\boldsymbol{\mu}] = \mathbb{E}[\mathbf{R}_{ij}^2|\boldsymbol{\mu}]$. The expectation of $||\mathbf{R}||_{\mathrm{F}}^2$ can be calculated as

$$\begin{aligned}\mathbb{E}[||\mathbf{R}||_{\mathrm{F}}^2] =& \mathbb{E}[\mathbb{E}[||\mathbf{R}||_{\mathrm{F}}^2|\boldsymbol{\mu}]] \\ =& \mathbb{E}[\sum_{i\neq j} Var[\mathbf{R}_{ij}|\boldsymbol{\mu}] + \sum_i Var[\mathbf{R}_{ii}|\boldsymbol{\mu}]].\end{aligned}$$

The remaining task is to calculate the conditional variance of $\mathbf{R}_{ij}$ and $\mathbf{R}_{ii}$, where we provide the result in Lemma 2.3.

Then by Markov inequality, for any $t > 0$, we have

$$\Pr(||\mathbf{R}||_{\mathrm{F}}^2 \geq t \times \mathbb{E}[||\mathbf{R}||_{\mathrm{F}}^2]) \leq 1/t$$

By setting $t = 1/\delta$, with probability at least $1-\delta$, we have that

$$||\mathbf{R}||_{\mathrm{F}} \leq \frac{1}{\sqrt{D\delta}} \sqrt{\frac{2}{L^2} + \frac{2}{V^2} + \mathcal{O}(\epsilon)} = \delta_{\mathbf{R}}.$$

□

Lemma 2.3. *For the LDA model, the following holds*

$$\mathbb{E}[Var[\mathbf{R}_{ij}|\boldsymbol{\mu}]] \leq \frac{1}{DL^2V^2} + \frac{2}{DV^4} + \mathcal{O}(\epsilon), \quad \forall i \neq j,$$

and

$$\mathbb{E}[Var[\mathbf{R}_{ii}|\boldsymbol{\mu}]] \leq \frac{1}{DL^2V} + \frac{2}{DV^4} + \mathcal{O}(\epsilon), \quad \forall i,$$

for $i, j = 1, 2, \ldots, V$ *and* ϵ *represents higher-order terms.*

Because we only need an upper bound on the variance, we make a few relaxations and introduce $\mathcal{O}(\cdot)$ notation to simplify the expression, i.e., we only keep the dominant terms and absorb the rest into $\mathcal{O}(\epsilon)$. To be rigorous, we have the following assumptions on the scale of each statistics or parameters: $L = \mathcal{O}(D)$, $V = \mathcal{O}(D)$, $L = \mathcal{O}(V)$, $K = \mathcal{O}(L)$, $K = \Omega(1)$, $\alpha = \Theta(1)$, and $\beta = \Theta(1)$. The calculation of the variance is provided in Appendix D.

It is interesting to examine the role of D, L, and V in $\delta_{\mathbf{R}}$. $\delta_{\mathbf{R}}$ decreases to 0 as $D \to +\infty$. Even if there are only two words in each document, $\hat{\mathbf{M}}_2$ would still converge to $\mathbf{M}_2$. This observation agrees with the discussion in [AGH+12]. L and V have similar influence over $\delta_{\mathbf{R}}$, which is limited by each other.

To apply the results above, we simply ignore the higher-order terms. However, because ϵ will increase as α, β, or K decreases, one should pay extra attention when the statistics D, L, V are far from the asymptotic region. As shown in our simulated studies, our bound yields convincing results when D, L, V are on the scale of hundreds or above, which is more than common in real-world applications.

2.4 Spectral Structure of $\mathbf{M}_2$

The spectral structure of $\mathbf{M}_2$ depends on K, V and $\boldsymbol{\mu}_k, \alpha_k, k = 1, 2, \ldots, K$. We use the following theorem to characterize the spectral structure of $\mathbf{M}_2$.

Theorem 2.4. *Assume that* $\alpha_{\min} = \min_k\{\alpha_k\}$, $\alpha_{\max} = \max_k\{\alpha_k\}$, *and* $\beta_v = \beta, \forall v = 1, \ldots, V$ *and*

$$\delta' = \left(\frac{\log(K/\delta_3)K\left(\beta + 2\log\left(K/\delta_2\right)\right)^2}{V\beta}\right)^{\frac{1}{2}},$$

(1) With probability at least $1-\delta_1-\delta_2-\delta_3$*, we have*

$$\begin{aligned}&\sigma_1(\mathbf{M}_2) \leq \overline{\sigma_1} \qquad (2)\\ =&\frac{\alpha_{\max}}{\alpha_0(\alpha_0+1)} \frac{(1+\delta')V(\beta + K\beta^2)}{\max\left\{0_+, V\beta - \sqrt{2V\beta\log(K/\delta_1)}\right\}^2}.\end{aligned}$$

(2) With probability at least $1-\delta_1-\delta_2-\delta_3$*, we have*

$$\begin{aligned}&\sigma_K(\mathbf{M}_2) \geq \underline{\sigma_K} \qquad (3)\\ =&\frac{\alpha_{\min}}{\alpha_0(\alpha_0+1)} \frac{(1-\delta')V\beta}{(V\beta + 2\sqrt{V\beta}\log(K/\delta_1))^2}.\end{aligned}$$

Proof. We have $\mathbf{M}_2 = \frac{1}{\alpha_0(\alpha_0+1)} \sum_{k=1}^K \alpha_k \boldsymbol{\mu}_k \otimes \boldsymbol{\mu}_k = \frac{1}{\alpha_0(\alpha_0+1)} \mathbf{O}\mathbf{A}\mathbf{O}^\top$, where $\mathbf{O} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ is a $V \times K$ matrix and $\mathbf{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_K)$ is a diagonal matrix. The first K singular values of $\mathbf{M}_2$ are also the first K

singular values of $\frac{1}{\alpha_0(\alpha_0+1)}\mathbf{A}^{\frac{1}{2}}\mathbf{O}^\top\mathbf{O}\mathbf{A}^{\frac{1}{2}}$. And we have

$$\sigma_1(\mathbf{A}^{\frac{1}{2}}\mathbf{O}^\top\mathbf{O}\mathbf{A}^{\frac{1}{2}}) \leq \sigma_1(\mathbf{A})\sigma_1(\mathbf{O}^\top\mathbf{O}),$$

and

$$\sigma_K(\mathbf{A}^{\frac{1}{2}}\mathbf{O}^\top\mathbf{O}\mathbf{A}^{\frac{1}{2}}) \geq \sigma_K(\mathbf{A})\sigma_K(\mathbf{O}^\top\mathbf{O}).$$

To estimate the singular value of $\mathbf{O}^\top\mathbf{O}$, we need to utilize that fact that $\boldsymbol{\mu}_k \sim \mathrm{Dir}(\beta)$. The random variables in the same column of $\mathbf{O}$ are dependent with each other, which keeps us from applying powerful results from random matrix theory. To decouple the dependency, we design a diagonal matrix $\boldsymbol{\Lambda}$, whose diagonal elements are drawn from Gamma$(V\beta, 1)$ independently. Therefore, $\hat{\mathbf{O}} = \mathbf{O}\boldsymbol{\Lambda}$ is a matrix with independent elements, i.e., each element is an i.i.d. random variable following Gamma$(\beta, 1)$.

We denote each row of $\hat{\mathbf{O}}$ as $\mathbf{r}_v, v = 1, \ldots, V$, then $\hat{\mathbf{O}}^\top\hat{\mathbf{O}} = \sum_{v=1}^V \mathbf{r}_v^\top\mathbf{r}_v$. In order to apply matrix Chernoff bound [Tro12], we need to bound the spectral norm of $\mathbf{r}_v^\top\mathbf{r}_v$, i.e., $\max_v\{\sigma_1(\mathbf{r}_v^\top\mathbf{r}_v)\}$. Because $\mathbf{r}_v^\top\mathbf{r}_v$ is a rank 1 matrix, we have $\sigma_1(\mathbf{r}_v^\top\mathbf{r}_v) = \mathbf{r}_v\mathbf{r}_v^\top$. By Lemma C.3 (see Appendix) and the union bound, with probability greater than $1 - KVe^{-\frac{c_1}{2}\min\{\frac{c_1}{2},\sqrt{\beta}\}}$, we have

$$R = \max_{v=1,\ldots,V}\{\sigma_1(\mathbf{r}_v^\top\mathbf{r}_v)\} \leq K(\beta + c_1\beta^{1/2})^2.$$

We also have $\sigma_1(\mathbb{E}[\hat{\mathbf{O}}^\top\hat{\mathbf{O}}]) = V\beta(1 + K\beta)$ and $\sigma_K(\mathbb{E}[\hat{\mathbf{O}}^\top\hat{\mathbf{O}}]) = V\beta$. Applying the matrix Chernoff bound to $\hat{\mathbf{O}}^\top\hat{\mathbf{O}}$, with probability greater than

$$1 - KVe^{-\frac{c_1}{2}\min\{\frac{c_1}{2},\sqrt{\beta}\}} - K\left[\frac{e^{-\delta'}}{(1-\delta')^{1-\delta'}}\right]^{\frac{V\beta}{K(\beta+c_1\beta^{1/2})^2}},$$

we have

$$\sigma_K(\hat{\mathbf{O}}^\top\hat{\mathbf{O}}) \geq (1-\delta')V\beta.$$

And with probability greater than

$$1 - KVe^{-\frac{c_1}{2}\min\{\frac{c_1}{2},\sqrt{\beta}\}} - K\left[\frac{e^{\delta'}}{(1+\delta')^{1+\delta'}}\right]^{\frac{V\beta}{K(\beta+c_1\beta^{1/2})^2}},$$

we have

$$\sigma_1(\hat{\mathbf{O}}^\top\hat{\mathbf{O}}) \leq (1+\delta')V\beta(1 + K\beta).$$

By definition, for $i = 1, \ldots, K$, it follows

$$\sigma_i(\mathbf{M}_2) = \frac{1}{\alpha_0(\alpha_0+1)}\sigma_i(\mathbf{A}^{\frac{1}{2}}\boldsymbol{\Lambda}^{-1}\hat{\mathbf{O}}^\top\hat{\mathbf{O}}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\frac{1}{2}}).$$

Therefore, we have

$$\sigma_1(\mathbf{M}_2) \leq \frac{\alpha_{\max}}{\alpha_0(\alpha_0+1)}\frac{\sigma_1(\hat{\mathbf{O}}^\top\hat{\mathbf{O}})}{\sigma_K^2(\boldsymbol{\Lambda})},$$

and

$$\sigma_K(\mathbf{M}_2) \geq \frac{\alpha_{\min}}{\alpha_0(\alpha_0+1)}\frac{\sigma_1(\hat{\mathbf{O}}^\top\hat{\mathbf{O}})}{\sigma_1^2(\boldsymbol{\Lambda})}.$$

Since $\sigma_1(\boldsymbol{\Lambda})$ and $\sigma_K(\boldsymbol{\Lambda})$ are the maximum and minimum of a set of random variables following Gamma$(V\beta, 1)$, we can bound them by Lemma C.4 with coefficient c_2. By setting the coefficients c_1, c_2, δ' carefully, we conclude the Theorem 2.4. We provide the details on the coefficients setting in Appendix A.1. □

With certain assumptions on $\alpha_{\max}$ and $\alpha_{\min}$, we can fully utilize the bounds above. If we assume that $\alpha_k = \Theta(\frac{1}{K}\sum_i \alpha_i) = \Theta(1)$, $\forall k$, then $\frac{\alpha_{\min}}{\alpha_0} = \Theta(\frac{1}{K})$ and $\alpha_0 = \Theta(K)$. Therefore, $\underline{\sigma_K}$ decreases rapidly as K increases, where $\sigma_K(\mathbf{M}_2) \propto \frac{1}{K^2}$ approximately. This fact leads to increasing difficulty in distinguishing the topics with small singular values from noise. Note that $\overline{\sigma_1}$ also decreases with a slower rate as K increases.

2.5 Implications on the Number of Topics

The convergence of $\hat{\mathbf{M}}_2$ and the spectral structure of $\mathbf{M}_2$ provide us the upper bounds and the lower bounds on the singular values of the empirical second-order moments $\hat{\mathbf{M}}_2$. We can discover the number of topics by the following steps to select the appropriate threshold θ:

First, by setting $\theta > \delta_{\mathbf{R}}$, thresholding provides a lower bound on K, since with high probability, every spurious topic has singular value smaller than $\delta_{\mathbf{R}}$.[1]

Secondly, if we set $\theta < \underline{\sigma_K} - \delta_{\mathbf{R}}$, thresholding provides a upper bound on K, since with high probability, every true topic has singular value greater than the threshold. However, the above threshold is not computable, because that $\underline{\sigma_K}$ depends on the true number of topics K.

Instead, we can directly utilize the upper bound $\overline{\sigma_1}$ on $\sigma_1(\hat{\mathbf{M}}_2)$ to provide an upper bound for the number of topics. We have $\sigma_1(\hat{\mathbf{M}}_2) \leq \overline{\sigma_1} + \delta_{\mathbf{R}}$ as shown in Theorem 2.4. The left hand side, $\sigma_1(\hat{\mathbf{M}}_2)$, is determined by the observed corpus, and the right hand side $\overline{\sigma_1} + \delta_{\mathbf{R}}$ is a function on the the number of topics. When $\overline{\sigma_1} + \delta_{\mathbf{R}}$ decreases as the the number of topics K increases (see

[1] Strictly speaking, there is no one-to-one correspondence between topics and the singular values of the second-order moments. Here we refer to the correspondence in terms of the total number of topics.

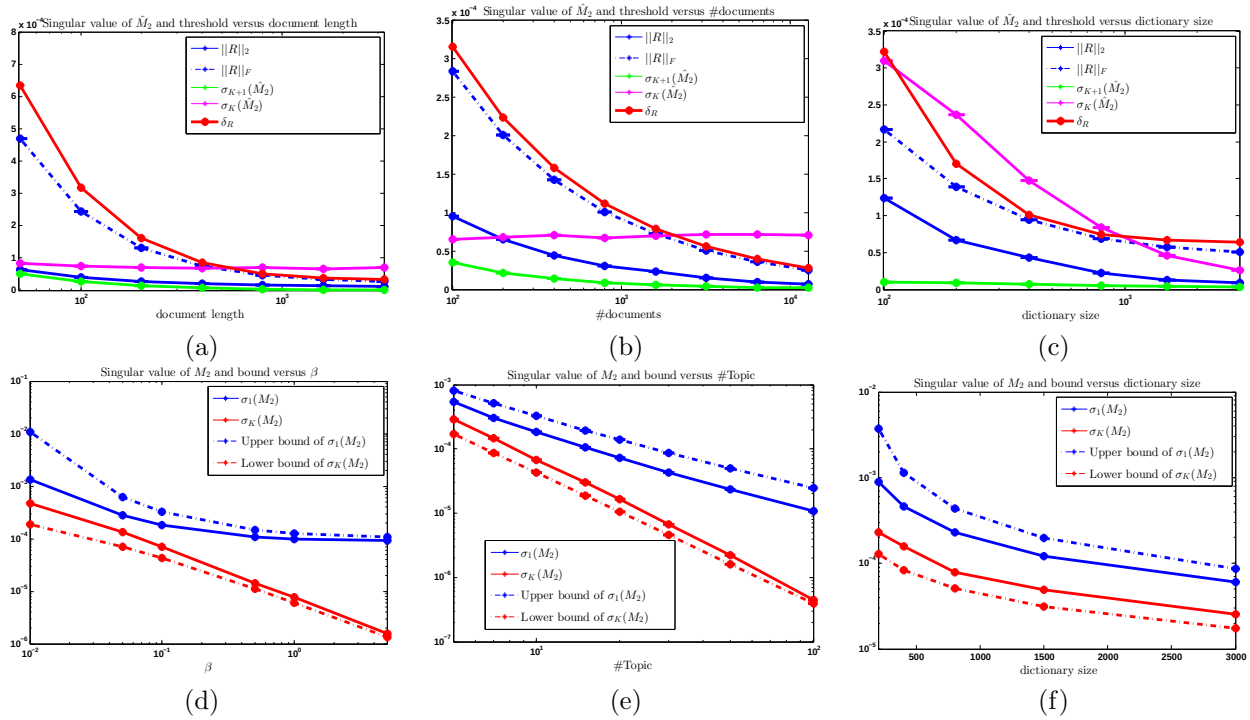


Figure 1: Experimental results on synthetic data under LDA model. Results on $\delta_{\mathbf{R}}$ are illustrated in Figure (a-c). $\underline{\sigma_K}$ and $\overline{\sigma_1}$ are illustrated in Figure (d-f).

discussion in Section 2.4), solving the inequality for K provides an upper bound on K.

3 Experimental Results

We validate our theoretical results by conducting experiments on the synthetic datasets generated according to the LDA model. For each experiment setting, we report the results by averaging over five random runs.

In the first set of experiments, we test the convergence of the second-order moment $\hat{\mathbf{M}}_2$ in terms of $\delta_{\mathbf{R}}$. The parameter setting is as follows: $K = 10$, $\forall k, \alpha_k = 1$ and $\forall v, \beta_v = 0.1$. We vary the dictionary size V, document length L, or document number D while keeping the other two fixed. The detailed settings are summarized as belows:

(a) Fix $D = 2000$ and $V = 1000$, vary length of document L from 50 to 3200.

(b) Fix $L = 500$ and $V = 1000$, vary number of documents D from 100 to 12800.

(c) Fix $L = 500$ and $D = 2000$, vary size of dictionary V from 100 to 3000.

Figure 1 (a-c) shows the matrix norms on $\mathbf{R} = \hat{\mathbf{M}}_2 - \mathbf{M}_2$ and the K-th and $(K+1)$-th largest singular values of $\hat{\mathbf{M}}_2$. The results completely agree with our theoretical analysis as expected, where $\delta_{\mathbf{R}}$ serves as an accurate upper bound on the Frobenius norm of $\hat{\mathbf{M}}_2 - \mathbf{M}_2$. When the amount of data is large enough, the red line goes below the purple line, which indicates that with enough data, thresholding with $\delta_{\mathbf{R}}$ provides a tight lower bound on the number of topics.

In the second experiment, we evaluate our bounds on the spectral structure of $\mathbf{M}_2$ in Theorem 2.4. Similarly, we vary K, β, or V while keeping the other two parameters fixed. The detailed settings are as follows:

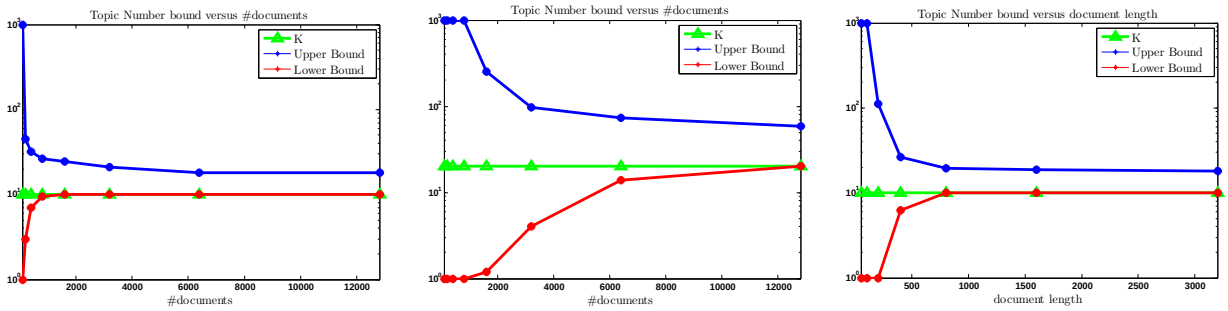
(d) Fix $\alpha_k = 1$, $V = 1000$, and $K = 10$, vary $\beta_v = \beta$ from 0.01 to 5.

(e) Fix $\alpha_k = 1$, $V = 1000$, and $\beta_v = 0.1$, vary number of topics K from 5 to 100.

(f) Fix $\alpha_k = 1$, $K = 10$, and $\beta_v = 0.1$, vary the size of dictionary V from 200 to 3000.

The results in Figure 1 (d-f) match well with our theoretical analysis.

In the last experiment, we calculate the upper and lower bound on K when varying the number of documents or the length of documents. The results are presented in Figure 2. As we can see, the lower bound



(a)$K = 10, L = 500$, varying D (b)$K = 20, L = 500$, varying D (c)$K = 10, D = 2000$, varying L

Figure 2: The upper and lower bound on number of topics for LDA based on discussion in section 2.5.

indeed converges to the true number of topics. However, the upper bound converges to a value other than the ground truth. This is due to the fact that the upper bound involves both $\overline{\sigma_1}$ and $\delta_{\mathbf{R}}$, whereas $\overline{\sigma_1}$ does not change as the size of dataset increases. The experiment results demonstrate that our theoretical upper and lower bound on K can effectively narrow down the range of possible K.

4 Discussion and Conclusions

So far we have shown that for the LDA model, by investigating the convergence of the empirical moments $\hat{\mathbf{M}}_2$ and the spectral structure of the expected moment $\mathbf{M}_2$, the singular values of the empirical moment provide useful information on the number of topics. The convergence rate $\delta_{\mathbf{R}}$ provides upper bounds for the singular value of spurious topics which leads to the lower bound on K by thresholding. Moreover, solving inequality on the first singular value $\sigma_1(\hat{\mathbf{M}}_2)$ provides an upper bound on the number of topics K. This line of research provides an interesting direction for analyzing other types of mixture models as explored in [HK13]. Here we formalize our methodology and present an example on Gaussian Mixture Models (GMM). As our purpose is methodology demonstration, we omit the comparison with the excellent existing works on GMM, such as [SR09].

4.1 Generalization

The methodology can be easily generalized to other mixture models whose empirical low-order moments have the same structures as the weighted sum of the outer products of mixture components. In order to derive the convergence bound $\delta_{\mathbf{R}}$, the variance of $\mathbf{R}_{ij}$ need to be computed for the model at hand. Moreover, we need to explore the spectral structure of the true moment to provide upper and lower bound on the first and the K-th singular values respectively. Then by combining the new convergence results and the knowledge on spectral structure, similar upper and lower bound on the number of mixture components can be derived. As an example, we next show how to bound the number of mixture components for the *Gaussian Mixture Model* [Bis06] with spherical mixture components.

GMM is one of the most popular mixture models due to its simplicity and effectiveness. It models the data points as the mixture of several multivariate Gaussian components. The generative process of GMM is summarized as follows: for a dataset $\{\mathbf{x}_i\}_{i=1}^N$ generated from *spherical Gaussian mixtures* with K components, we assume that

$$\begin{aligned} h_i \sim & \text{Multi}(w_1, w_2, \ldots, w_K), \\ \mathbf{x}_i \sim & \mathcal{N}(\boldsymbol{\mu}_{h_i}, \sigma^2 \mathbf{I}), \\ i = & 1, 2, \ldots, N \end{aligned}$$

where $(w_1, w_2, \ldots, w_K)$ is the pmf for each mixture component, h_i is the component assignment for the i-th data point, and $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ is a m-dimensional spherical Gaussian distribution with $M \geq K$. We also assume $\boldsymbol{\mu}_i \sim \mathcal{N}(\mathbf{0}, \sigma_\mu^2 \mathbf{I})$ and $(w_1, w_2, \ldots, w_K) \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_K)$ to complete the generative process. Note that we assume the following parameters are known: $\sigma, \sigma_\mu, \alpha_k, k = 1, 2, \ldots, K$.

The problem on how to correctly choosing the number of mixture components has been extensively studied. Besides traditional methods such as cross validation, AIC and BIC [LV10], other methods such as penalized likelihood method [THK13] and variational approach [CB01] are also proposed to solve the problem. Similar to the LDA model, we show that analyzing the empirical moments provides an alternative approach to bound the number of mixture components.

We define the empirical second-order moment as $\hat{\mathbf{M}}_2 = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i \otimes \mathbf{x}_i - \sigma^2 \mathbf{I}$ and the second-order moment $\mathbf{M}_2$ as the expectation of the empirical moment, namely

$\mathbf{M}_2 = \mathbb{E}[\hat{\mathbf{M}}_2]$. Then by similar analysis, we have the following theorem for bounding the number of mixture components in spherical GMM:

Theorem 4.1. *Let $\alpha_k = \alpha, \forall k$, then*

(1) Let K_l be the number of singular values of $\hat{\mathbf{M}}_2$ such that $\sigma(\hat{\mathbf{M}}_2) > \delta_{\mathbf{R}}$, where

$$\delta_{\mathbf{R}} = \frac{\sigma m}{\sqrt{N\delta}} \sqrt{2\sigma_\mu^2 + \frac{m+1}{m}\sigma^2},$$

then with probability at least $1-\delta$, we have

$$K \geq K_l.$$

(2) Let K_u be the maximal integer such that

$$\sigma_1(\hat{\mathbf{M}}_2) \leq \frac{\sigma_\mu^2}{K_u} \frac{(\alpha + 2\log(K_u/\delta_1))\left((\sqrt{m}+\sqrt{K_u}+t)^2\right)}{\max\{0_+, \alpha - \sqrt{2\alpha\log(1/\delta_2)/K_u}\}} + \delta_{\mathbf{R}}.$$

Then with probability at least $1-\delta_1-\delta_2$, we have

$$K \leq K_u.$$

The proof for Theorem 4.1 is similar to that for LDA model by providing convergence rate $\delta_{\mathbf{R}}$ on the singular values of $\hat{\mathbf{M}}_2$ and bounds on the singular values of $\mathbf{M}_2$. The detailed proof is in Appendix B due to space limit.

4.2 Conclusion

In this paper, we provide theoretical analysis on model selection for LDA model. Specifically, we present both upper and lower bound on the number of topics K based on the connection between second-order moments and latent topics. The upper bound is obtained by bounding the difference between the estimated second-order moment $\hat{\mathbf{M}}_2$ and the true moment $\mathbf{M}_2$. The lower bound is obtained via analyzing the largest singular value of $\hat{\mathbf{M}}_2$. Furthermore, our analysis can be easily generalized to other latent models, such as Gaussian mixture models.

The major limitation of our approach is that all our analysis assumes that the data are generated exactly according to the model. As a result, the current methodology may fail when applying to real world dataset due to model misspecification or lacking of low rank structure in the moments.

For future work, we would like to explore the practicality of our analysis to real applications. In addition, we will examine several ways to improve the theoretical results. For example, if we can bound the higher-order moments of $\hat{\mathbf{M}}_2 - \mathbf{M}_2$, we can improve the results by replacing Markov inequality with tighter inequalities. Moreover, we could bound the spectral norm of $\hat{\mathbf{M}}_2 - \mathbf{M}_2$ directly instead of bounding its Frobenius norm, which will yield tighter bounds.

5 Acknowledgment

We thank Fei Sha and David Kale for helpful discussion and suggestion. The research was sponsored by the NSF research grants IIS-1254206, and U.S. Defense Advanced Research Projects Agency (DARPA) under Social Media in Strategic Communication (SMISC) program, Agreement Number W911NF-12-1-0034. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agency, or the U.S. Government.

References

[AEF+10] Edoardo M. Airoldi, Elena A. Erosheva, Stephen E. Fienberg, Cyrille Joutard, Tanzy Love, and Suyash Shringarpure. Reconceptualizing the classification of pnas articles. *Proceedings of the National Academy of Sciences*, 107(49):20899–20904, 2010.

[AFH+12] Anima Anandkumar, Dean P Foster, Daniel Hsu, Sham Kakade, and Yi-Kai Liu. A spectral algorithm for latent dirichlet allocation. In *NIPS*, pages 926–934, 2012.

[AGH+12] Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *arXiv preprint arXiv:1210.7559*, 2012.

[AGM12] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *FOCS*, pages 1–10, 2012.

[Aka74] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control*, 19(6):716–723, 1974.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[CB01] Adrian Corduneanu and Christopher M Bishop. Variational Bayesian model selection for mixture distributions. In *AISTATS*, pages 27–34, 2001.

[GS04] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, April 2004.

[HBWP13] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *JMLR*, 14(1):1303–1347, 2013.

[HJ] Roger A. Horn and Charles R. Johnson. Matrix analysis, 1985.

[HK13] Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *ITCS*, 2013.

[KRS14] Alex Kulesza, N Raj Rao, and Satinder Singh. Low-rank spectral learning. In *ICML*, 2014.

[LM00] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The annals of statistics*, 28(5):1302–1338, 2000.

[LV10] Olga Lukociene and Jeroen K. Vermunt. Determining the number of components in mixture models for hierarchical data. In *Advances in Data Analysis, Data Handling and Business Intelligence*, pages 241–249. 2010.

[MH13] Jeffrey W Miller and Matthew T Harrison. A simple example of dirichlet process mixture inconsistency for the number of components. In *NIPS*, pages 199–206. 2013.

[PNI+08] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *KDD*, pages 569–577, 2008.

[S+78] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[SR09] Russell J Steele and Adrian E Raftery. Performance of bayesian model selection criteria for gaussian mixture models. *Dept. Stat., Univ. Washington, Washington, DC, Tech. Rep*, 559, 2009.

[Tad12] Matt Taddy. On estimation and selection for topic models. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, pages 1184–1193, 2012.

[THK13] Huang Tao, Peng Heng, and Zhang Kun. Model selection for gaussian mixture models. *arXiv preprint arXiv:1301.3558*, 2013.

[TJBB06] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

[TKW07] Yee Whye Teh, Kenichi Kurihara, and Max Welling. Collapsed variational inference for hdp. In *NIPS*, 2007.

[TMN+14] Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *ICML*, 2014.

[Tro12] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[WPB11] Chong Wang, John W Paisley, and David M Blei. Online variational inference for the hierarchical dirichlet process. In *AISTATS*, pages 752–760, 2011.