Analyzing the Number of Latent Topics via Spectral Decomposition

Dehua Cheng* USC
Xinran He* USC
Yan Liu USC

October 24, 2014

Abstract

Correctly choosing the number of topics plays an important role in successfully applying topic models to real world applications. Following the latest tensor decomposition framework by Anandkumar et al., we make the first attempt to provide theoretical analysis on the number of topics under *Latent Dirichlet Allocation* model. With mild conditions, our method provides accessible information on the number of topics, which includes both upper and lower bounds. Experimental results on synthetic datasets demonstrate that our proposed bounds are correct and tight. Furthermore, using *Gaussian Mixture Model* as an example, we show that our methodology can be easily generalized for analyzing the number of mixture components in other mixture models.

1 Introduction

Modeling large, complex real-world domains demands powerful models that can handle rich relational structures. One application that has attracted much interest from machine learning researchers is modeling text corpus or collections of documents via latent topic models. One particular model, *Latent Dirichlet Allocation* (LDA) [BNJ03], has gained extensive popularity and achieved significant success in both academia and industry.

As the LDA model assumes that documents are generated from mixture of topics, the main inference problem becomes recovering the latent topics from the observed corpus. Popular algorithms include variational inference [BNJ03, TKW07, WPB11, HBWP13], sampling methods [GS04, PNI$^+$08], and the recent decomposition based methods [AGM12, AFH$^+$12, AGH$^+$12]. However, all of them treat the number of topics K as a fixed and given parameter. It is known that choosing the correct number of topics K plays a vital role in successfully applying LDA models: setting K too small or too large will lead to inaccurate inference results. For example, [TMN$^+$14] has shown that choosing K too large leads to severe deterioration in the learning rate; [KRRS14] 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.

It is highly nontrivial to choose the number of topics K in the LDA model. Existing solutions are mostly based on statistical model selection techniques, such as AIC [Aka74], BIC [S$^+$78], or cross-validation. However, all of them require multiple runs of the learning algorithm with different K, which limits the practicality of this strategy 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. But HDP and similar approaches also suffer from the huge computational bottleneck, though recent work has accelerated HDP-related methods remarkably by exploring, for example, approaches based on stochastic variational inference [HBWP13, TKW07, WPB11] or parallel sampling [WDX13, ASW08, CL14]. Moreover, 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.

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

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 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. 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 down-sampling 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 bound on K without unknown parameters or constants. We also present the simulated study of 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$^+$08]. Recently, method of moments have been explored, leading to a series of interesting work and insight into the LDA model from a traditional yet brand new perspective. It has been shown in [AFH$^+$12, AGH$^+$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, since their close connection with each other. 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:

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

(a) Generate a topic $z_{d\ell} \sim \text{Multi}(\mathbf{h}_d)$, where $\text{Multi}(\mathbf{h}_d)$ denotes the multinomial distribution.

(b) Generate a word $\mathbf{x}_{d\ell} \sim \text{Multi}(\boldsymbol{\mu}_{z_{d\ell}})$, 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

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

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. One 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 any 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 $k = 1, \dots, K$ and $\beta_v = \beta$ for $v = 1, \dots, V$. 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*

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

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 $i \ge K + 1$, *we have*

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

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:

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

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 $\mathrm{Rank}(\mathbf{M}_2) \le K$, i.e., $\sigma_i(\mathbf{M}_2) = 0$ for $i \ge K + 1$, the second statement holds true. □

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

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

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

The last task, but not the most important task is to calculating 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*

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

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

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

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 $\mathbf{A} = \operatorname{diag}(\alpha_1, \ldots, \alpha_K)$ 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 $\boldsymbol{\mu}_k \sim \operatorname{Dir}(\beta)$. 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 $\operatorname{Gamma}(V\beta, 1)$ 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 $\operatorname{Gamma}(\beta, 1)$.

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

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

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

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

and

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

Since $\sigma_1(\mathbf{\Lambda})$ and $\sigma_K(\mathbf{\Lambda})$ 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. $\square$

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

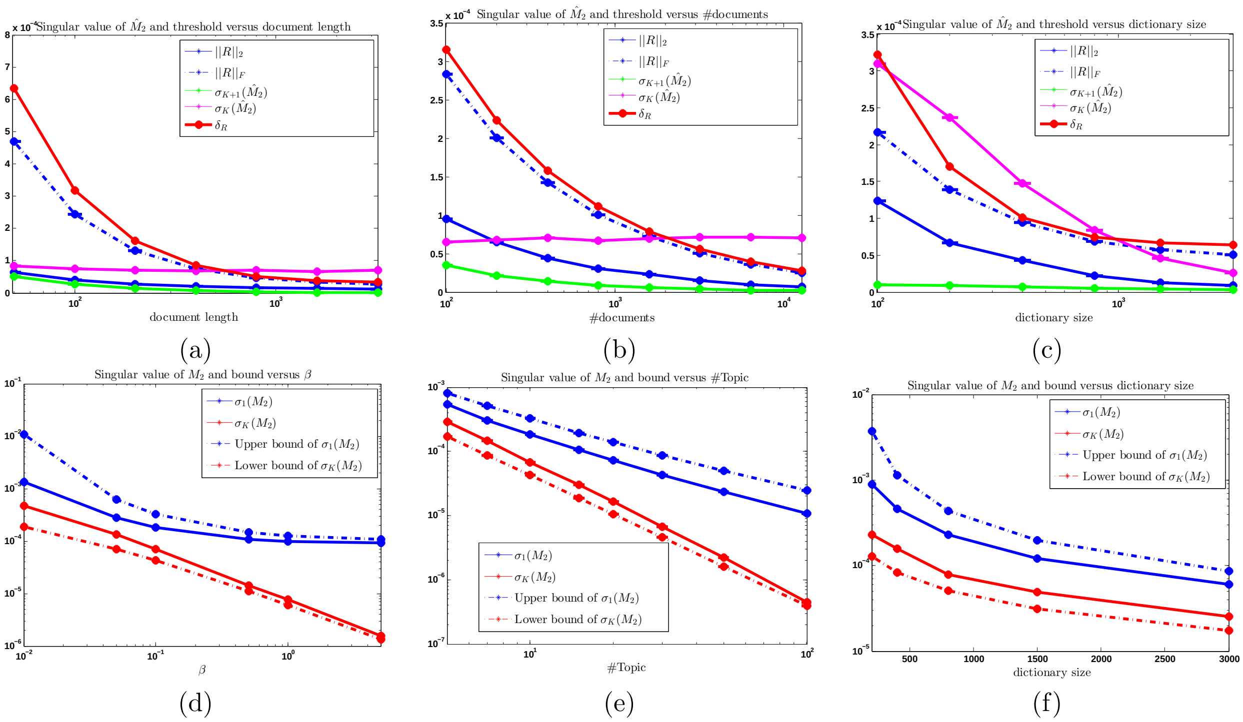


Figure 1: Experimental results on synthetic data under LDA model. Results on $\delta_{\mathbf{R}}$ are illustrated in Figure (a-c). $\underline{\sigma_K}$ and $\overline{\sigma_1}$ are illustrated in Figure (d-f).

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 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.

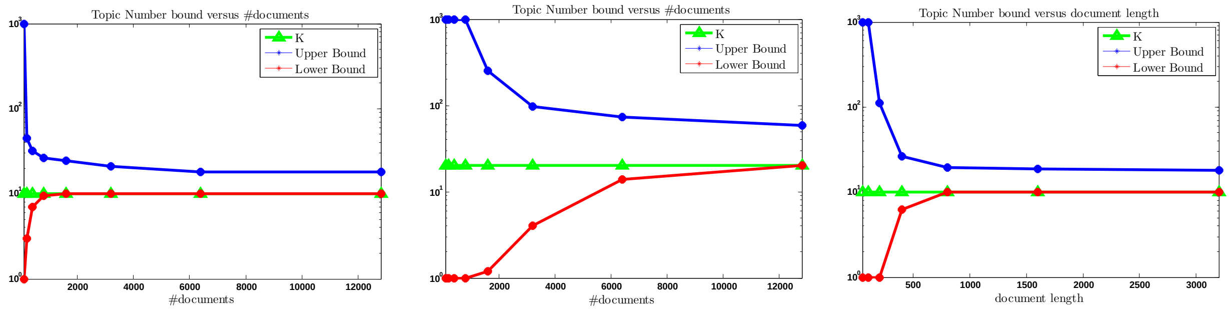
(e) Fix $\alpha_k = 1$, $V = 1000$, and $\beta_v = 0.1$, vary number of topics K from 5 to 100.

(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 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.

[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.



(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.

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.

In this section, we will discuss its generalization to other mixture models as well as its limitations.

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* (GMM) [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

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

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 Limitation and Future Work

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. Though it seems impossible to derive theoretic result under this situation, as a future work, we would like to explore heuristics for discovering the number of topics for real world dataset utilizing the connection between topics and the spectral structure of the moments.

From the technical point of view, there are several ways to improve the proposed 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. In addition, better understanding on the spectral structure of $\mathbf{M}_2$ will also lead to tighter bounds.

5 Acknowledgment

We thank Fei Sha, Mohammad Taha Bahadori, and David Kale for discussions and helpful suggestions.

References

[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.

[ASW08] Arthur Asuncion, Padhraic Smyth, and Max Welling. Asynchronous distributed learning of topic models. In *NIPS*, pages 81–88, 2008.

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

[CL14] Dehua Cheng and Yan Liu. Parallel gibbs sampling for hierarchical dirichlet processes via gamma processes equivalence. In *KDD*, 2014.

[GS04] T. L. Griffiths and M. 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.

[KRRS14] A. Kulesza, N. Raj Rao, and S. 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 JeroenK. 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.

[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.

[WDX13] Sinead Williamson, Avinava Dubey, and Eric P Xing. Parallel markov chain monte carlo for nonparametric mixture models. In *ICML*, pages 98–106, 2013.

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

A Theoretical results for LDA

A.1 Coefficient Setting for Theorem 2.4

Bound of $\sigma_1(M_2)$

We have that with probability greater than

$$
\begin{aligned}
1 - &K e^{-\frac{c_2^2}{2}} \\
&- KVe^{-\frac{c_1}{2}\min\{\frac{c_1}{2},\sqrt{\beta}\}} \\
&- K[\frac{e^{\delta'}}{(1+\delta')^{1+\delta'}}]^{\frac{V\beta}{K(\beta+c_1\beta^{1/2})^2}},
\end{aligned}
$$

we have

$$\sigma_1(M_2) \le \frac{1}{K(K\alpha+1)} \frac{(1+\delta')V(\beta+K\beta^2)}{(V\beta - c_2\sqrt{V\beta})^2}.$$

We can choose c_1, c_2 and δ' as follows to simplify the formula of the bound

- Choose $c_2 = \sqrt{2\log(K/\delta_1)}$, first probability term is less than δ_1.
- Choose $c_1 = \frac{2}{\sqrt{\beta}}\log(KV/\delta_2)$, third probability term is less than δ_2.
- Choose δ' as

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

  second probability term is less than δ_3.

As a result, with probability greater than $1-\delta_1-\delta_2-\delta_3$, we have

$$\sigma_1(M_2) \le \frac{1}{K(K\alpha+1)} \frac{(1+\delta')V(\beta+K\beta^2)}{(V\beta - \sqrt{2V\beta\log(K/\delta_1)})^2}.$$

As an alternative, we can choose c_1, c_2 and δ_1 as follows to simplify the formula of the bound

- Choose $c_2 = \sqrt{2\log(K/\delta)}$, first probability term is less than δ.
- Choose $c_1 = \frac{4}{\sqrt{\beta}}\log(KV)$, third probability term is less than $\frac{1}{KV}$.
- Choose $\delta' = 0.1$, second probability term is less than $K(0.995)^{\frac{V(\beta+K\beta^2)}{K(\beta+c_1\beta^{1/2})^2}}$.

As a result, with probability greater than

$$1 - \delta - \frac{1}{KV} - K(0.995)^{\frac{V\beta}{K(\beta+2\log(KV))^2}},$$

we have

$$\sigma_1(M_2) \le \frac{1.1}{K(K\alpha+1)} \frac{V(\beta+K\beta^2)}{(V\beta - \sqrt{2V\beta\log(K/\delta)})^2}.$$

Bound of $\sigma_K(M_2)$

We have that with probability greater than

$$\begin{aligned}1 - &Ke^{-\frac{c_2}{2}\min\{\frac{c_2}{2}, V\beta\}} \\ &-KVe^{-\frac{c_1}{2}\min\{\frac{c_1}{2}, \sqrt{\beta}\}} \\ &-K[\frac{e^{-\delta'}}{(1-\delta')^{1-\delta'}}]^{\frac{V\beta}{K(\beta+c_1\beta^{1/2})^2}},\end{aligned}$$

we have

$$\sigma_K(M_2) \geq \frac{1}{K(K\alpha+1)}\frac{(1-\delta')V\beta}{(V\beta+c_2\sqrt{V\beta})^2}$$

We can choose c_1, c_2 and δ' as follows to simplify the formula of the bound

- Choose $c_2 = 2\sqrt{\log(K/\delta_1)}$, first probability term is less than δ_1.
- Choose $c_1 = \frac{2}{\sqrt{\beta}}\log(KV/\delta_2)$, third probability term is less than δ_2.
- Choose δ' as

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

  second probability term is less than δ_3.

As a result, with probability greater than $1-\delta_1-\delta_2-\delta_3$, we have

$$\sigma_K(M_2) \geq \frac{1}{K(K\alpha+1)}\frac{(1-\delta')V\beta}{(V\beta+2\sqrt{V\beta}\log(K/\delta_1))^2}$$

As an alternative, we can choose c_1, c_2 and δ_1 as follows to simplify the formula of the bound

- Choose $c_1 = \frac{4}{\sqrt{\beta}}\log(KV)$, third probability term is less than $\frac{1}{KV}$.
- Choose $c_2 = 2\sqrt{\log(K/\delta)}$, first probability term is less than δ.
- Choose $\delta' = 0.1$, second probability term is less than $K(0.995)^{\frac{V(\beta+K\beta^2)}{K(\beta+c_1\beta^{1/2})^2}}$.

As a result, with probability greater than

$$1-\delta-\frac{1}{KV}-K(0.995)^{\frac{V\beta}{K(\beta+2\log(KV))^2}},$$

we have

$$\sigma_K(M_2) \geq \frac{0.9}{K(K\alpha+1)}\frac{V\beta}{(V\beta+2\sqrt{V\beta\log K/\delta})^2}.$$

A.2 Lemma for Theorem 2.1

Lemma A.1. *With $\hat{\mathbf{M}}_2$ and $\mathbf{M}_2$ previously defined, we have that*

$$\max_i |\sigma_i(\hat{\mathbf{M}}_2) - \sigma_i(\mathbf{M}_2)| \leq \max_i |\lambda_i(\hat{\mathbf{M}}_2) - \lambda_i(\mathbf{M}_2)|$$

Proof. Because $\mathbf{M}_2$ is a symmetric semidefinite matrix, so we have

$$\sigma_i(\mathbf{M}_2) = \lambda_i(\mathbf{M}_2), \quad \forall i,$$

And because $\hat{\mathbf{M}}_2$ is a symmetric matrix, we have

$$\sigma_i(\hat{\mathbf{M}}_2) = |\lambda_{s(i)}(\hat{\mathbf{M}}_2)|, \quad \forall i,$$

for some permutation s.

Because we have $\lambda_i(\hat{\mathbf{M}}_2) \leq |\lambda_i(\hat{\mathbf{M}}_2)| = \sigma_j(\hat{\mathbf{M}}_2)$, so we have $\lambda_i(\hat{\mathbf{M}}_2) \leq \sigma_i(\hat{\mathbf{M}}_2)$.

Let j be the smallest index that $|\lambda_j(\hat{\mathbf{M}}_2)| \neq \sigma_j(\hat{\mathbf{M}}_2)$, for $i < j$, we have

$$\begin{aligned}|\sigma_i(\hat{\mathbf{M}}_2) - &\sigma_i(\mathbf{M}_2)| \\ &= |\lambda_i(\hat{\mathbf{M}}_2) - \lambda_i(\mathbf{M}_2)| \\ &\leq \max_i |\lambda_i(\hat{\mathbf{M}}_2) - \lambda_i(\mathbf{M}_2)|\end{aligned}$$

By the fact that $\lambda_i(\mathbf{M}_2) \geq 0$, we have that for $\forall i \geq j$,

$$\sigma_i(\hat{\mathbf{M}}_2) \leq \max_k |\lambda_k(\hat{\mathbf{M}}_2) - \lambda_k(\mathbf{M}_2)|$$

We also have

$$\sigma_i(\hat{\mathbf{M}}_2) \geq \lambda_i(\hat{\mathbf{M}}_2)$$

Because

$$|\lambda_i(\hat{\mathbf{M}}_2) - \sigma_i(\mathbf{M}_2)| \leq \max_k |\lambda_k(\hat{\mathbf{M}}_2) - \lambda_k(\mathbf{M}_2)|$$

We can prove that

$$|\sigma_i(\hat{\mathbf{M}}_2) - \sigma_i(\mathbf{M}_2)| \leq \max_k |\lambda_k(\hat{\mathbf{M}}_2) - \lambda_k(\mathbf{M}_2)|$$

Therefore,

$$\max_i |\sigma_i(\hat{\mathbf{M}}_2) - \sigma_i(\mathbf{M}_2)| \leq \max_i |\lambda_i(\hat{\mathbf{M}}_2) - \lambda_i(\mathbf{M}_2)|$$

□

B Theoretical results for GMM

The proof of Theorem 4.1 is achieved by analyzing the concentration result $\delta_{\mathbf{R}}$ of empirical second order moments and also upper bound for the first singular value of the true moment $\mathbf{M}_2$. Thresholding with $\delta_{\mathbf{R}}$ leads to the first claim, while solving the inequality on the $\sigma_1(\hat{\mathbf{M}}_2)$ provides the second claim.

B.1 Relation Between $\mathbf{M}_2$ and $\hat{\mathbf{M}}_2$

We bound the different between singular values of $\mathbf{M}_2$ through the following Theorem.

Theorem B.1. *For* spherical Gaussian mixtures *with probability at least* $1 - \delta$, $\forall i \in \{1, 2, \dots, m\}$,*we have*

$$|\sigma_i(\hat{\mathbf{M}}_2) - \sigma_i(\mathbf{M}_2)| \leq \frac{\sigma m}{\sqrt{N\delta}} \sqrt{2\sigma_\mu^2 + \frac{m+1}{m}\sigma^2} = \delta_{\mathbf{R}}$$

Especially, when $i \leq K + 1$, *we have*

$$\sigma_i(\hat{\mathbf{M}}_2) \leq \frac{\sigma m}{\sqrt{N\delta}} \sqrt{2\sigma_\mu^2 + \frac{m+1}{m}\sigma^2}. \tag{4}$$

Proof. We establish the result by bounding the Frobenius of matrix $\mathbf{R}$ as we do for LDA model. The square of Frobenius norm is $||\mathbf{R}||_{\mathrm{F}}^2 = \sum_{i,j} \mathbf{R}_{ij}^2$. Since we have $\mathbb{E}[\mathbf{R}_{ij}|\mu] = 0$, thus

$$Var[\mathbf{R}_{ij}|\mu] = \mathbb{E}[\mathbf{R}_{ij}^2|\mu] - \mathbb{E}^2[\mathbf{R}_{ij}|\mu] = \mathbb{E}[\mathbf{R}_{ij}^2|\mu],$$

and

$$\begin{aligned}
\mathbb{E}[||\mathbf{R}||_{\mathrm{F}}^2] =& \mathbb{E}[\mathbb{E}[||\mathbf{R}||_{\mathrm{F}}^2|\mu]] \\
=& \mathbb{E}[\sum_{i,j} Var[\mathbf{R}_{ij}|\mu]|\mu] \\
=& \mathbb{E}[\sum_{i \neq j} Var[\mathbf{R}_{ij}|\mu] + \sum_i Var[\mathbf{R}_{ii}|\mu]|\mu] \\
=& \frac{m(m-1)}{N}\sigma^2(2\sigma_\mu^2 + \sigma^2) + \frac{m}{N}\sigma^2(2\sigma_\mu^2 + 2\sigma^2) \\
=& \frac{m^2\sigma^2}{N}(2\sigma_\mu^2 + \frac{m+1}{m}\sigma^2).
\end{aligned}$$

Then by Markov inequality, we have

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

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

$$\|\mathbf{R}\|_{\mathrm{F}} \leq \frac{\sigma m}{\sqrt{N\delta}}\sqrt{2\sigma_\mu^2 + \frac{m+1}{m}\sigma^2}$$

□

B.2 Spectral Structure of $\mathbf{M}_2$

We use following theorem to characterize the spectral structure of $\mathbf{M}_2$.

Theorem B.2. *Assume that $\alpha_i = \alpha$ in the* spherical Gaussian mixtures*, we have*
(1) With probability at least $1 - \delta_1 - \delta_2 - 2\exp(-t^2/2)$, we have

$$\sigma_1(\mathbf{M}_2) \leq \frac{\sigma_\mu^2}{K}\frac{\alpha + 2\log(K/\delta_1)}{\alpha - \sqrt{2\alpha\log(1/\delta_2)/K}}(\sqrt{m} + \sqrt{K} + t)^2 \tag{5}$$

(2) Further assume that and $w_i \geq w_{\min}, \forall i$, then with probability at least $1 - 2\exp(-t^2/2)$, we have

$$\sigma_K(\mathbf{M}_2) \geq w_{\min}\sigma_\mu^2(\sqrt{m} - \sqrt{K} - t)^2 \tag{6}$$

Proof. We have $\mathbf{M}_2 = \sum_{k=1}^K w_k\mu_k \otimes \mu_k = \mathbf{OAO}^\top$, where $\mathbf{O} = (\mu_1, \mu_2, \ldots, \mu_K)$ is a $m \times K$ matrix and $\mathbf{A} = diag(w_1, w_2, \ldots, w_K)$ is a diagonal matrix. Because $\mathbf{M}_2 = \mathbf{OAO}^\top = \mathbf{OA}^{1/2}\mathbf{A}^{1/2}\mathbf{O}^\top$, we have that $\sigma_i(\mathbf{M}_2) = \sigma_i(\mathbf{A}^{1/2}\mathbf{O}^\top\mathbf{OA}^{1/2}), \forall i = 1, 2, \ldots, K$. Therefore, we have the following inequalities [HJ]:

$$\sigma_1(\mathbf{M}_2) \leq \sigma_1(\mathbf{O}^\top\mathbf{O})\sigma_1(\mathbf{A}), \tag{7}$$

$$\sigma_K(\mathbf{M}_2) \geq \sigma_K(\mathbf{O}^\top\mathbf{O})\sigma_K(\mathbf{A}). \tag{8}$$

Note that the elements of $\mathbf{O}$ are i.i.d. Gaussian random variables, i.e., $\mathbf{O}_{ij} \sim \mathcal{N}(0, \sigma_\mu^2)$. The distribution of $\sigma_i(\mathbf{O}^\top\mathbf{O})$ has been well-studied in random matrix theory [Ver10]. With probability at least $1 - 2\exp(-t^2/2)$, we have

$$\sigma_1(\mathbf{O}^\top\mathbf{O}) \leq \sigma_\mu^2(\sqrt{m} + \sqrt{K} + t)^2,$$
$$\sigma_K(\mathbf{O}^\top\mathbf{O}) \geq \sigma_\mu^2(\sqrt{m} - \sqrt{K} - t)^2.$$

And since $\sigma_1(\mathbf{A}) = \max_i\{w_i\}$, we can prove that with probability at least $1-\delta_1-\delta_2$, we have (see appendix C.3 for proof)

$$\max_i\{w_i\} \leq \frac{1}{K}\frac{\alpha + 2\log(K/\delta_1)}{\alpha - \sqrt{2\alpha\log(1/\delta_2)/K}}$$

We also have $\sigma_K(\mathbf{A}) = \min_i\{w_i\} \geq w_{\min}$. We complete the proof by substituting the above formulas into inequalities (7). $\square$

C Tail bound for Gamma distribution

In this section, we proof some tail bound related to the Gamma distribution. Our main tool is the following Lemma.

Lemma C.1. ***[Massart and Laurent] Tail Bound for Chi-square distribution*** *Let U be a χ^2_D random variable with D degree of freedom, then for any positive x, the following holds*

$$Pr(U \geq D + 2\sqrt{Dx} + 2x) \leq e^{-x},$$
$$Pr(U \leq D - 2\sqrt{Dx}) \leq e^{-x}.$$

Proof. See [LM00] for proof. $\square$

C.1 Tail Bound for a Single Gamma Distribution

In this section, we provide tail bound for a single Gamma random variable (R. V.).

Lemma C.2. ***Tail Bound for Gamma R.V.*** *Let $X \sim Gamma(\alpha, 1)$ be a Gamma R.V. with shape parameter α, and scale parameter 1, then for any positive c, the following holds*

$$Pr(X \geq \alpha + c\sqrt{\alpha}) \leq e^{-\frac{c}{2}\min\{\frac{c}{2},\sqrt{\alpha}\}},$$
$$Pr(X \leq \alpha - c\sqrt{\alpha}) \leq e^{-\frac{c^2}{2}}.$$

Proof. By relationship between Gamma R.V. and chi-square R.V., we have that $2X \sim \chi^2_{2\alpha}$. Apply Lemma C.1 directly, we have

$$\Pr(X \geq \alpha + c\sqrt{\alpha}) \leq e^{-c\sqrt{\alpha}+\alpha(\sqrt{1+2c\alpha^{-1/2}}-1)},$$
$$\Pr(X \leq \alpha - c\sqrt{\alpha}) \leq e^{-\frac{c^2}{2}}.$$

To get the same formula as in the lemma, we can easily prove that $c\sqrt{\alpha} - \alpha(\sqrt{1+2c\alpha^{-1/2}} - 1) > \frac{c}{2}\min\{\frac{c}{2},\sqrt{\alpha}\}, \quad \forall c, \alpha > 0$.

$\square$

Corollary C.3. ***Tail Bound for Sum of Square of Gamma R.V.*** *If we have n i.i.d Gamma R.V. $X_i \sim Gamma(\alpha, 1), i = 1, \ldots, n$, then for any positive c, the following holds*

$$Pr(\sum_i X_i^2 \geq n(\alpha + c\sqrt{\alpha})^2) \leq ne^{-\frac{c}{2}\min\{\frac{c}{2},\sqrt{\alpha}\}}.$$

C.2 Tail Bound for Maximum/Minimum of Gamma Random Variables

Lemma C.4. *If we have n i.i.d Gamma R.V. $X_i \sim Gamma(\alpha, 1), i = 1, \ldots, n$, we have that*

$$Pr(\max_i\{X_i\} \geq \alpha + c\sqrt{\alpha}) \leq ne^{-\frac{c}{2}\min\{\frac{c}{2},\sqrt{\alpha}\}},$$
$$Pr(\min_i\{X_i\} \leq \alpha - c\sqrt{\alpha}) \leq ne^{-\frac{c^2}{2}}.$$

Proof. It can be proved by applying union bound directly. $\square$

C.3 Tail Bound for Maximum/Minimum Element of Dirichlet Distribution

It is well known that a random vector $(x_1, x_2, \ldots, x_n) \sim \mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_n)$ is equivalent to a random vector $(y_1, y_2, \ldots, y_n)/\sum_i y_i$, where $y_i \sim \mathrm{Gamma}(\alpha_i, 1)$ independently. And we have $\max_i\{x_i\} = \max_i\{y_i\}/\sum_i y_i$.

Assume $\alpha_i = \alpha$, so we have

$$\Pr(\max_i\{y_i\} \geq \alpha + c_1\sqrt{\alpha}) \leq ne^{-\frac{c_1}{2}\min\{\frac{c_1}{2}, \sqrt{\alpha}\}}.$$

And since $\sum_i y_i \sim \mathrm{Gamma}(n\alpha, 1)$, we have

$$\Pr(\sum_i y_i \leq n\alpha - c_2\sqrt{n\alpha}) \leq e^{-\frac{c_2^2}{2}}$$

By setting $c_1 = 2\log(n/\delta_1)/\sqrt{\alpha}$ (when $n > \delta_1 e^{\alpha}$) and $c_2 = \sqrt{2\log(1/\delta_2)}$, we have that with probability at least $1 - \delta_1 - \delta_2$,

$$\max_i\{x_i\} \leq \frac{1}{n}\frac{\alpha + \log(n/\delta_1)}{\alpha - \sqrt{2\alpha\log(1/\delta_2)/n}}$$

Similarity, $\min_i\{x_i\} = \min_i\{y_i\}/\sum_i y_i$. And

$$\begin{aligned}\Pr(\min_i\{x_i\} \leq \alpha - c_1\sqrt{\alpha}) &\leq ne^{-\frac{c_1^2}{2}},\\ \Pr(\sum_i y_i \geq n\alpha + c_2\sqrt{n\alpha}) &\leq e^{-\frac{c_2}{2}\min\{\frac{c_2}{2}, \sqrt{n\alpha}\}}.\end{aligned}$$

By setting $c_1 = \sqrt{2\log(n/\delta_1)}$ and $c_2 = \sqrt{2\log(1/\delta_2)}$ (when $\delta_2 > e^(-2\alpha)$), we have that with probability at least $1 - \delta_1 - \delta_2$,

$$\min_i\{x_i\} \geq \frac{1}{n}\frac{\alpha - \sqrt{2\log(n\alpha/\delta_1)}}{\alpha + \sqrt{2\alpha\log(1/\delta_2)/n}}$$

which is nontrivial only when α is large enough.

D Variance Calculation for LDA

In this section, we presents the overall procedure and some important intermediate results of the variance calculation for LDA. Note that we have the following assumptions on the scale of each statistics or parameters: $L = \mathcal{O}(D)$, $V = \mathcal{O}(D)$, $L = \mathcal{O}(V)$, $K = \mathcal{O}(L)$, $1/K = \mathcal{O}(1)$, $\alpha = \Theta(1)$, and $\beta = \Theta(1)$.

First, we have

$$\begin{aligned}R =& \frac{1}{D}\sum_d \frac{1}{L(L-1)}\sum_{l \neq s} x_{d,l}x_{d,s}^\top\\ &- \frac{\alpha_0}{\alpha_0 + 1}[\frac{1}{D}\sum_d \frac{1}{L}\sum_l x_{d,l}][\frac{1}{D}\sum_d \frac{1}{L}\sum_l x_{d,l}]^\top\\ &- M_2.\end{aligned}$$

We represent each term by

$$\begin{aligned}R^{(1)} =& \frac{1}{D}\sum_d \frac{1}{L(L-1)}\sum_{l \neq s} x_{d,l}x_{d,s}^\top,\\ R^{(2)} =& \frac{\alpha_0}{\alpha_0 + 1}[\frac{1}{D}\sum_d \frac{1}{L}\sum_l x_{d,l}][\frac{1}{D}\sum_d \frac{1}{L}\sum_l x_{d,l}]^\top,\\ R^{(3)} =& \frac{1}{D}\sum_d \frac{1}{L}\sum_l x_{d,l}.\end{aligned}$$

And we have the following identity:

$$\begin{aligned}E_\mu Var_X[R_{ij}] =&E_\mu Var_X[R_{ij}^{(1)}] + E_\mu Var_X[R_{ij}^{(2)}]\\&-2E_\mu Cov_X[R_{ij}^{(1)}, R_{ij}^{(2)}],\end{aligned}$$

with $H = \{\mu, h\}$, $X = \{h, x\}$.

$$R_{ij}^{(2)} = \frac{\alpha_0}{\alpha_0 + 1} R_i^{(3)} R_j^{(3)}.$$

For simplicity of representation, we assume the following,

$$\begin{aligned}f_d^{(ij)} =&\frac{1}{L(L-1)} \sum_{l \neq s}^{L} x_{d,l}^{(i)} x_{d,s}^{(j)},\\g_d^{(i)} =&\frac{1}{L} \sum_{l=1}^{L} x_{d,l}^{(i)}.\end{aligned}$$

and the superscript (ij) or (i) will be omitted if there is no ambiguity. By this representation, we have

$$\begin{aligned}R^{(1)} =&\frac{1}{D} \sum_d f_d,\\R^{(3)} =&\frac{1}{D} \sum_d g_d.\end{aligned}$$

We also assume the representation $z_d^{(i)} = \sum_k \mu_k^{(i)} h_d^{(k)}$, which is the probability of e_i in the d-th documents conditioned on $H = \{\mu, h\}$. And $\delta_{ij} = 1$ if and only if $i = j$.

The intermediate results for diagonal and off-diagonal variance are different, so we provide them separately in the following sections.

D.1 Calculate Off-diagonal Variance

In this section, we assume that $i \neq j$. And we have the following results:

$$\begin{aligned}E_\mu Var_X[R_{ij}^{(1)}] \leq&\frac{1}{DL^2V^2} + \frac{2}{DLV^3} + \frac{1}{DV^4} + O(\epsilon)\\E_\mu Var_X[R_{ij}^{(2)}] \leq&\frac{2}{DLV^3} + \frac{1}{DV^4} + O(\epsilon)\\E_\mu Cov_X(R^{(1)}, R^{(2)}) \geq&\frac{2}{DLV^3} + O(\epsilon)\end{aligned}$$

Therefore, we have that

$$\begin{aligned}E_\mu Var_X[R_{ij}] \leq&\frac{1}{DL^2V^2} + \frac{2}{DLV^3} + \frac{1}{DV^4} + \frac{2}{DLV^3}\\&+\frac{1}{DV^4} - \frac{4}{DLV^3} + O(\epsilon)\\=&\frac{1}{DL^2V^2} + \frac{2}{DV^4} + O(\epsilon).\end{aligned}$$

D.2 Calculate Diagonal Variance

In this section, we assume that $i \neq j$. And we have the following results:

$$
\begin{aligned}
E_\mu Var_X[R_{ij}^{(1)}] \leq & \frac{1}{DL^2V} + \frac{4}{DLV^3} + \frac{1}{DV^4} + O(\epsilon), \\
E_\mu Var_X[R_{ij}^{(2)}] \leq & \frac{2}{DLV^3} + \frac{1}{DV^4} + O(\epsilon), \\
E_\mu Cov_X(R^{(1)}, R^{(2)}) \geq & \frac{3}{DLV^3} + O(\epsilon).
\end{aligned}
$$

Therefore, we have that

$$
\begin{aligned}
E_\mu Var_X[R_{ij}] \leq & \frac{1}{DL^2V} + \frac{4}{DLV^3} + \frac{1}{DV^4} + \frac{2}{DLV^3} \\
& + \frac{1}{DV^4} - \frac{6}{DLV^3} + O(\epsilon) \\
= & \frac{1}{DL^2V} + \frac{2}{DV^4} + O(\epsilon).
\end{aligned}
$$