Title
Spectral Methods for Correlated Topic Models

Authors
Arabshahi, Forough
Anandkumar, Animashree

Publication Date
2016-05-29

Peer reviewed

Beyond LDA: A Unified Framework for Learning Latent Normalized Infinitely Divisible Topic Models through Spectral Methods

Forough Arabshahi
University of California Irvine
farabsha@uci.edu

Animashree Anandkumar
University of California Irvine
a.anandkumar@uci.edu

Abstract

In this paper we propose guaranteed spectral methods for learning a broad range of topic models, which generalize the popular Latent Dirichlet Allocation (LDA). We overcome the limitation of LDA to incorporate arbitrary topic correlations, by assuming that the hidden topic proportions are drawn from a flexible class of Normalized Infinitely Divisible (NID) distributions. NID distributions are generated through the process of normalizing a family of independent Infinitely Divisible (ID) random variables. The Dirichlet distribution is a special case obtained by normalizing a set of Gamma random variables. We prove that this flexible topic model class can be learnt via spectral methods using only moments up to the third order, with (low order) polynomial sample and computational complexity. The proof is based on a key new technique derived here that allows us to diagonalize the moments of the NID distribution through an efficient procedure that requires evaluating only univariate integrals, despite the fact that we are handling high dimensional multivariate moments.

Keywords: Latent variable models, spectral methods, tensor decomposition, moment matching, infinitely divisible, Lévy processes.

1 Introduction

Topic models are a popular class of exchangeable latent variable models for document categorization. The goal is to uncover hidden topics based on the distribution of word occurrences in a document corpus. Topic models are *admixture* models, which go beyond the usual mixture model which allows for only one hidden topic to be present in each document. In contrast, topic models incorporate multiple topics in each document. It is assumed that each document has a latent proportions of different topics, and the observed words are drawn in a conditionally independent manner, given the set of topics.

Latent Dirichlet Allocation (LDA) is the most popular topic model [6], where the topic proportions are drawn from the Dirichlet distribution. While LDA has widespread applications, it is limited by the choice of the Dirichlet distribution. Notably, Dirichlet distribution can only model negative correlations [4], and thus, is unable to incorporate arbitrary correlations among the topics that may be present in different document corpora. Another drawback is that the elements with similar means need to have similar variances. While there have been previous attempt to go beyond the Dirichlet distribution, e.g. [5, 15], their correlation structures are still limited (e.g. only positive correlations can be modeled in [5]), learning these models is usually difficult and no guaranteed algorithms exist.

In this work, we consider a flexible class of topic models, and propose guaranteed and efficient algorithms for learning them. We employ the class of Normalized Infinitely Divisible (NID) distributions to model the topic proportions [8, 13]. These are a class of distributions on the simplex, formed by normalizing a set of independent draws from a family of positive Infinitely Divisible (ID) distributions. The draws from an ID distribution can be represented as a sum of an arbitrary number of i.i.d. random variables. The concept of infinite divisibility was introduced in 1929 by Bruno de Finetti, and the most fundamental results were developed by Kolmogorov, Lévy and Khintchine in the 1930s.

The Gamma distribution is an example of an ID distribution, and the Dirichlet distribution is obtained by normalizing a set of independent draws from Gamma distributions. We show that the class of NID topic models significantly generalize the LDA model: they can incorporate both positive and negative correlations among the topics, they involve additional parameters to vary the variance and higher order moments, while fixing the mean, and so on.

There are mainly three categories of algorithms for learning topic models, viz., variational inference [5, 6], Gibbs sampling [7, 9, 14], and spectral methods [2, 16]. Among them, spectral methods have gained increasing prominence over the last few years, due to their efficiency and guaranteed learnability. In this paper, we develop novel spectral methods for learning latent NID topic models.

Spectral methods have previously been proposed for learning LDA [2], and in addition, other latent variable models such as Independent Component Analysis (ICA), Hidden Markov Models (HMM), mixtures of ranking distributions, and so on [3]. The idea is to learn the parameters based on spectral decomposition of low order moment tensors (third or fourth order). Efficient algorithms for tensor decomposition have been proposed before [3], and implies consistent learning with (low order) polynomial computational and sample complexity.

The main difficulty in extending spectral methods to the more general class of NID topic models is the presence of arbitrary correlations among the hidden topics which need to be "untangled". For instance, take the case of a single topic model (i.e. each document has only one topic); here, the third order moment, which is the co-occurrence tensor of word triplets, has a CP decomposition, and computing the decomposition yields an estimate of the topic-word matrix. In contrast, for the LDA model, such a tensor decomposition is obtained by a combination of moments up to the third order. In other words, the moments of the LDA model need to be appropriately "centered" in order to have the tensor decomposition form.

Finding such a moment combination has so far been an "art form", since it is based on explicit manipulation of the moments of the hidden topic distribution. So far, there is no principled mechanism to automatically find the moment combination with the CP decomposition form. For arbitrary topic models, however, finding such a combination may not even be possible. In general, one requires all the higher order moments for learning.

In this work, we show that surprisingly, for the flexible class of NID topic models, moments up to third order suffice for learning, and we provide an efficient algorithm for computing the coefficients to combine the moments. The algorithm is based on computation of a univariate integral, that involves the Levy measure of the underlying ID distribution. The integral can be computed efficiently through numerical integration since it is only univariate, and has no dependence on the topic or word dimensions. Intriguingly, this can be accomplished, even when there exist no closed form probability density functions (pdf) for the NID variables.

The paper is organized as follows. In Section 2, we propose our "Latent Normalized Infinitely Divisible Topic Models" and present its generative process. We dedicate Section 3 to the properties of NID distributions and indicate how they overcome the drawbacks of the Dirichlet distribution and other distributions on the simplex. In Section 4 we present our efficient learning algorithm with guaranteed convergence for the proposed topic model based on spectral decomposition. Finally, we conclude the paper in Section 5.

2 Latent Normalized Infinitely Divisible Topic Models

Topic models incorporate relationships between words $\mathbf{x}_1, \mathbf{x}_2 \ldots \in \mathbb{R}^d$ and a set of k hidden topics. We represent the words $\mathbf{x}_i$ using one-hot encoding, i.e. $\mathbf{x}_i = \mathbf{e}_j$ if j^{th} word in the vocabulary occurs, and $\mathbf{e}_j$ is the standard basis vector. The proportions of topics in a document is represented by vector $\mathbf{h} \in \mathbb{R}^k$. We assume that $\mathbf{h}$ is drawn from an NID distribution.

The detailed generative process of a latent NID topic model for each document is as follows

1. Draw k independent variables, $z_1, z_2, \ldots, z_k$ from a family of ID distributions.
2. Set $\mathbf{h}$ to $(\frac{z_1}{Z}, \ldots, \frac{z_k}{Z})$ where $Z = \sum_{i \in [k]} z_i$.
3. For each word $\mathbf{x}_i$,
   (a) Choose a topic $\zeta_i \sim \text{Multi}(\mathbf{h})$ and represent it with one-hot encoding.
   (b) Choose a word $\mathbf{x}_i$ vector as a standard basis vector with probability

$$\mathbb{E}(\mathbf{x}_i|\zeta_i) = \mathbf{A}\zeta_i, \tag{1}$$

   conditioned on the drawn topic ζ_i, and $\mathbf{A} \in \mathbb{R}^{d \times k}$ is the topic-word matrix.

From (1), we also have

$$\mathbb{E}(\mathbf{x}_i|\mathbf{h}) = \mathbb{E}[\mathbb{E}(\mathbf{x}_i|\mathbf{h}, \zeta_i)] = \mathbb{E}(\mathbf{x}_i|\zeta_i)\mathbb{E}(\zeta_i|\mathbf{h}) = \mathbf{A}\mathbf{h}. \tag{2}$$

When the z_i is drawn from the Gamma$(\alpha_i, 1)$ distribution, we obtain the Dir$(\boldsymbol{\alpha})$ distribution for the hidden vector $\mathbf{h} = (h_1, \ldots, h_k)$, and the LDA model through the above generative process.

Our goal is to recover the topic-word matrix $\mathbf{A}$ given the document collection. In the following section we introduce the class of NID distribution and discuss its properties.

3 Properties of NID distributions

NID distributions are a flexible class of distributions on the simplex and have been applied in a range of domains. This includes hierarchical mixture modeling with Normalized Inverse-Gaussian distribution [12], and modeling overdispersion with the normalized tempered stable distribution [11], both of which are examples of NID distributions. For more applications, see [8].

Let us first define the concept of infinite divisibility and present the properties of an ID distribution, and then consider the NID distributions.

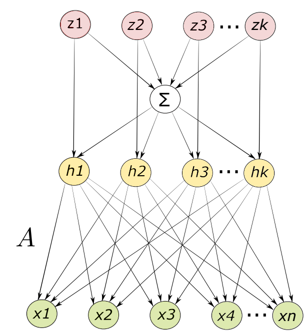


Figure 1: Graphical Model Representation of the Latent NID Topic Model. $z_1, z_2, \ldots, z_k$ are a collection of independent Infinitely Divisible positive variables that are characterized by the collection of their corresponding Lévy measures $\alpha_1\nu, \alpha_2\nu, \ldots, \alpha_k\nu$ And $h_1, h_2 \ldots, h_k$ are the resulting NID variables representing topic proportions in a document of length N with words $x_1, \ldots, x_N$

3.1 Infinitely Divisible Distributions

If random variable z has an Infinitely Divisible (ID) distribution, then for any $n \in \mathbb{N}$ there exists a collection of i.i.d random variables $y_1, \ldots, y_n$ such that $z \overset{\text{d}}{=} y_1 + \cdots + y_n$. In other words, an Infinitely Divisible distribution can be expressed as the sum of an arbitrary number of independent identically distributed random variables.

The Poisson distribution, compound Poisson, the negative binomial distribution, Gamma distribution, and the trivially degenerate distribution are examples of Infinitely Divisible distributions; as are the normal distribution, Cauchy distribution, and all other members of the stable distribution family. The Student's t-distribution is also another example of Infinitely Divisible distributions. The uniform distribution and the binomial distribution are not infinitely divisible, as are all distributions with bounded (finite) support.

The special decomposition form of ID distributions makes them natural choices for certain models or applications. E.g. a compound Poisson distribution is a Poisson sum of IID random variables. The discrete compound Poisson distribution, also known as the stuttering Poisson distribution, can model batch arrivals (such as in a bulk queue [1]) and can incorporate Poisson mixtures.

In the sequel, we limit the discussion to ID distributions on $\mathbb{R}^+$ in order to ensure that the Normalized ID variables are on the simplex. Let us now present how ID distributions can be characterized.

Lévy measure: A σ-finite Borel measure ν on $\mathbb{R}^+$ is called a Lévy measure if $\int_0^\infty \min(1,x)\nu(\mathrm{d}x) < \infty$. According to the Lévy-Khintchine representation given below, the Lévy measure uniquely characterizes an ID distribution along with a constant scale τ. This implies that every Infinitely Divisible distribution corresponds to a Lévy process, which is a stochastic process with independent increments.

Lévy-Khintchine representation [Theorem 16.14 [10]] Let $\mathcal{M}_1(\Lambda)$ and $\mathcal{M}_\sigma(\Lambda)$ indicate the subset of probability measures and the set of σ-finite measures on a non-empty set Λ, respectively. Let $\mu \in \mathcal{M}_1([0,\infty))$ and let $\Psi(u) = -\log \int\limits_0^\infty e^{-uz}\mathrm{d}(\mu)$ be the log-Laplace transform of μ. Then μ is Infinitely Divisible, if and only if there exists a $\tau \geq 0$ and a σ-finite measure $\nu \in \mathcal{M}_\sigma((0,\infty))$ with

$$\int\limits_0^\infty \min(1,z)\nu(\mathrm{d}z) < \infty, \tag{3}$$

such that

$$\Psi(u) = \tau u + \int\limits_0^\infty (1 - e^{-uz})\nu(\mathrm{d}z) \qquad \text{for} \quad u \geq 0, \tag{4}$$

In this case the pair (τ,ν) is unique, ν is called the Lévy measure of μ and τ is called the deterministic part. It can be shown that $\tau = \sup\{z \geq 0 : \mu([0,z)) = 0\}$.

In particular, let $\Phi_{z_i}(u) = \mathbb{E}[e^{\iota u z_i}] = \int\limits_0^\infty e^{\iota u z_i} f(z_i)\mathrm{d}z_i$ indicate the characteristic function of an Infinitely Divisible random variable z_i with pdf $f(z_i)$ and corresponding pair (τ_i,ν_i), where ι is the imaginary unit. Based on the Lévy-Khintchine representation it holds that $\Phi_{z_i}(\iota u) = \mathbb{E}[e^{-uz_i}] = e^{-\Psi_i(u)}$ where $\Psi_i(u) = \tau_i u + \int\limits_0^\infty (1-e^{-uz})\nu_i(\mathrm{d}z)$ is typically referred to as the Laplace exponent of z_i. This implies that the Laplace exponent of an ID variable is also completely characterized by pair (τ_i,ν_i). It holds for ID variables that if ν_i is a well-defined Lévy measure, so is $\alpha_i\nu_i$ for any $\alpha_i > 0$, which indicates that $\alpha_i\Psi_i(u)$ is also a well-defined Laplace exponent of an ID variable.

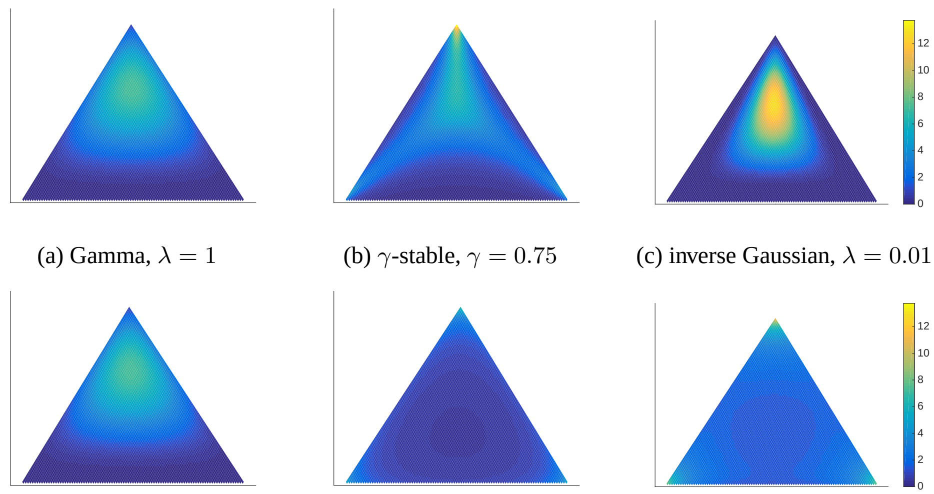
3.2 Normalized Infinitely Divisible Distributions

As defined in [8], a Normalized Infinitely Divisible (NID) random variable is a random variable that is formed by normalizing independent draws of strictly positive (not necessarily coinciding) Infinitely Divisible distributions. More specifically, let $z_1,\ldots,z_k$ be a set of independent strictly positive Infinitely Divisible random variables and $Z = z_1 + \cdots + z_k$. An NID distribution is defined as the distribution of the random vector $\mathbf{h} = (h_1,\ldots,h_k) := (\frac{z_1}{Z},\ldots,\frac{z_k}{Z})$ on the $(k-1)$-dimensional simplex, denoted as Δ^{k-1}. The strict positivity assumption implies that $\mathbf{h}$ is on the simplex [8, 13].

Let $[k]$ denote Natural numbers $1,\ldots,k$. As stated by the Lévy-Khintchine theorem, a collection of ID positive variables z_i for $i \in [k]$ is completely characterized by the collection of the corresponding Lévy measures $\nu_1,\ldots,\nu_k$. It was shown in [13] that this also holds for the normalized variables h_i for $i \in [k]$.

In this paper, we assume that the ID variables $z_1,\ldots,z_k$ are drawn independently from ID distributions that are characterized with the corresponding collection of Lévy measures $\alpha_i\nu,\ldots,\alpha_k\nu$, respectively. Which in turn translates respectively to variables with Laplace exponents $\alpha_1\Psi(u),\ldots,\alpha_k\Psi(u)$. Variables α_i will allow the distribution to vary in the interior of the simplex, providing the asymmetry needed to model latent models. The homogeneity assumption on the Lévy measure or the Laplace exponent provides the structure needed for guaranteed learning (Theorem 1). The overall graphical model representation is shown in Figure 1

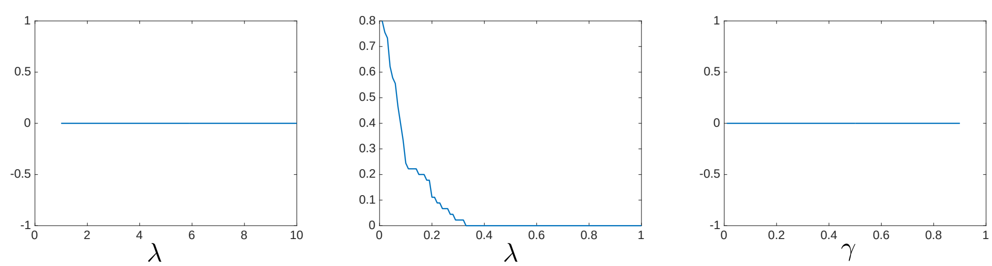
If the original ID variables z_i have probability densities f_i for all $i \in [k]$, then the distribution of vector $\mathbf{h}$, where $h_k = 1 - \sum_{i\in[k-1]} h_i$ is, $f(\mathbf{h}) = \int\limits_0^\infty \prod\limits_{i\in[k]} f_i(h_i Z) Z^{k-1}\mathrm{d}Z$. There are only three members of the NID class that have closed form densities namely, the Gamma distribution, Gamma(α_i,λ), the Inverse Gaussian distribution, $IG(\alpha_i,\lambda)$, and the $1/2$-stable distribution



(a) Gamma, $\lambda = 1$ (b) γ-stable, $\gamma = 0.75$ (c) inverse Gaussian, $\lambda = 0.01$

(d) Gamma, $\lambda = 10$ (e) γ-stable, $\gamma = 0.4$ (f) inverse Gaussian, $\lambda = 4$

Figure 2: Heat map of the pdf of three examples of the NID class that have closed form with respect to their parameters. All the figures have $\alpha = (2, 2, 4)$. For the Inverse Gaussian the distribution moves from the center to the vertices of the simplex as λ goes from 0 to ∞ with fixed $\boldsymbol{\alpha}$ and for the γ-stable we have the same behavior when γ changes from 1 to 0 with fixed $\boldsymbol{\alpha}$.



(a) Gamma NID. (b) Inverse Gaussian NID. (c) γ-stable NID.

Figure 3: Proportion of positively correlated elements of special cases of an NID distribution with 10 elements with respect to the parameter of the Laplace exponent for a fixed randomly drawn vector $\boldsymbol{\alpha} = [0.77, 0.70, 0.97, 0.46, 0.02, 0.44, 0.90, 0.33, 0.97, 0.45]$.

$St(\gamma, \beta, \alpha_i, \mu)$ with $\gamma = 1/2$. $\mu = 1$ and $\beta = 1$ to ensure positive support for the Stable distribution. As noted earlier, Gamma$(\alpha_i, 1)$ reduces to the Dirichlet distribution. An interested reader is referred to [8, 13] for the closed form of each distribution.

Figure 2 depicts the heatmap of the density of these distributions on the probability simplex for different value of their parameters. Note that all the distributions have the same α parameter and hence, the same mean values. However, their concentration properties are widely varying, showing that the NID class can incorporate variations in higher order moments through additional parameters.

Gamma ID distribution: When the ID distribution is Gamma with parameters $(\alpha_i, 1)$, we have the Dirichlet distribution as the resulting NID distribution. The Laplace exponent for this distribution will, therefore, be $\Psi_i(u) = \alpha_i \ln(1 + u)$.

γ-stable ID distribution: The variables are drawn from the positive stable distribution $St(\gamma, \beta, \alpha_i, \mu)$ with $\mu = 0$, $\beta = 1$ and $\gamma < 1$ which ensures that the distribution is on $\mathbb{R}^+$. The Laplace exponent of this distribution is $\Psi_i(u) = \alpha_i \frac{\Gamma(1-\gamma)}{\sqrt{2\pi\gamma}} u^\gamma$. Note that the γ-stable distribution can be represented in closed form for $\gamma = \frac{1}{2}$.

Inverse Gaussian ID distribution: The random variables are drawn from the Inverse-Gaussian (IG) distribution $IG(\alpha_i, \lambda)$. The Laplace exponent of this distribution is $\Psi_i(u) = \alpha_i\big(\sqrt{2u+\lambda^2} - \lambda\big)$.

Note: The Dirichlet distribution, the $1/2$-Stable distribution and the Inverse Gaussian distribution are all special cases of the generalized Inverse Gaussian distribution [8].

As mentioned earlier, the class of NID distributions is capable of modeling positive and negative correlations among the topics. This property is depicted in Figure 3. These figures show the proportion of positively correlated topics for the three presented distributions. As we can see the Inverse Gaussian NID distribution can capture both positive and negative correlations. It should be noted that the Logistic-Normal distribution that has been proposed before [5] can *only* capture positive correlations. But with NID topic models we can handle both positively *and* negatively correlated topics simultaneously.

4 Learning NID Topic Models through Spectral Methods

In this section we will show how the form of the moments of NID distributions enable efficient learning of this flexible class.

In order to be able to guarantee efficient learning using higher order moments, the moments need to have a very specific structure. Namely, the moment of the underlying distribution of $\mathbf{h}$ needs to form a diagonal tensor. If the components of $\mathbf{h}$ where indeed independent, this is obtained through the cumulant tensor. On the other hand, for LDA, it has been shown by Anandkumar et. al. [2] that a linear combination of moments of up to third order of $\mathbf{h}$ forms a diagonal tensor for the Dirichlet distribution. Below, we extend the result to the more general class of NID distributions.

4.1 Consistency of Learning through Moment Matching

Assumption 1 *ID random variables z_i for $i \in [k]$ are said to be* partially homogeneous *if they share the same Lévy measure. This implies that the corresponding Laplace exponent of variable z_i is given $\alpha_i \Psi(u)$ for some $\alpha_i \in \mathbb{R}^+$, and $\Psi(u)$ is the Laplace exponent of the common Lévy measure.*

Under the above assumption, we prove guaranteed learning of NID models through spectral methods. This is based on the following moment forms for NID models, which admit a CP tensor decomposition. The components of the decomposition will be the columns of the topic-word matrix: $\mathbf{A} := [\mathbf{a}_1|\mathbf{a}_2|\dots|\mathbf{a}_k]$.

Define

$$\Omega(m,n,p) = \int_0^\infty u^m \frac{\mathrm{d}^n}{\mathrm{d}u^n}\Psi(u)\Big(\frac{\mathrm{d}}{\mathrm{d}u}\Psi(u)\Big)^p e^{-\alpha_0\Psi(u)}\mathrm{d}u, \tag{5}$$

where $\Psi(u)$ is the Laplace exponent of the NID distribution and $\alpha_0 = \sum_{i\in[k]} \alpha_i$.

Theorem 1 ***(Moment Forms for NID models)*** *Let $\mathbf{M}_2$ and $\mathbf{M}_3$ be respectively the following matrix and tensor constructed from the following moments of the data,*

$$\mathbf{M}_2 = \mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2] + v\cdot \mathbb{E}[\mathbf{x}_1]\otimes\mathbb{E}[\mathbf{x}_2], \tag{6}$$

$$\mathbf{M}_3 = \mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \mathbf{x}_3] + v_2\cdot \mathbb{E}[\mathbf{x}_1]\otimes\mathbb{E}[\mathbf{x}_2]\otimes\mathbb{E}[\mathbf{x}_3]$$

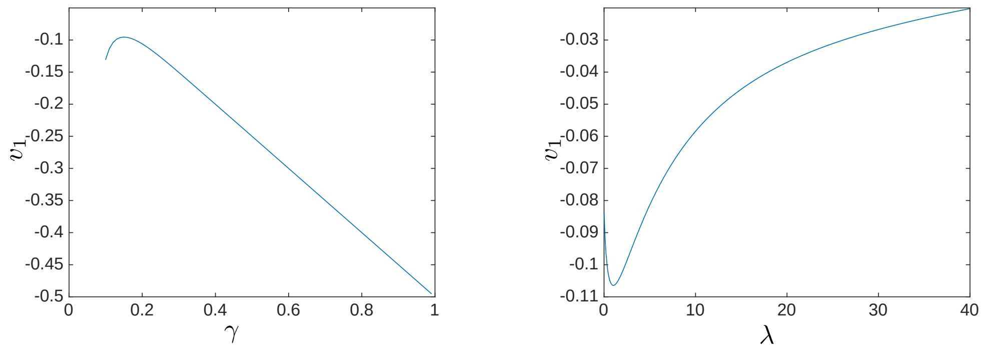
$$+ v_1\cdot\big[\mathbb{E}[\mathbf{x}_1\otimes\mathbf{x}_2]\otimes\mathbb{E}[\mathbf{x}_3] + \mathbb{E}[\mathbf{x}_1]\otimes\mathbb{E}[\mathbf{x}_2\otimes\mathbf{x}_3] + \mathbb{E}[\mathbf{x}_1\otimes\mathbb{E}[\mathbf{x}_2]\otimes\mathbf{x}_3]\big] \tag{7}$$

$$\tag{8}$$

where,

$$v = \frac{\Omega(1,1,1)}{\Big(\Omega(0,1,0)\Big)^2}, \quad v_1 = -\frac{\Omega(2,2,1)}{2\Omega(1,2,0)\Omega(0,1,0)}, \tag{9}$$

$$v_2 = \frac{-0.5\Omega(2,1,2) + 3v_1\Omega(1,1,1)\Omega(0,1,0)}{\Big(\Omega(0,1,0)\Big)^3}, \tag{10}$$



(a) Weight v_1 of theorem 1 for a Stable ID distribution $St(\gamma, \beta, \alpha_i, \mu)$ with $\mu = 0$, $\beta = 1$, $\gamma < 1$ and $\alpha_i > 0$ vs. γ for $\alpha_0 = 1$

(b) Weight v_1 of theorem 1 for an Inverse Gaussian distribution $IG(\alpha_i, \lambda)$ vs. $\lambda > 0$ and $\alpha_i \geq 0$ for $\alpha_0 = 1$

Figure 4: Weight v_1 for two different examples of the NID distribution. Weights v and v_2 in the theorem have similar behavior w.r.p the parameters.

Then given Assumption 1,

$$\mathbf{M}_2 = \sum_{j \in [k]} \kappa_j (\mathbf{a}_j \otimes \mathbf{a}_j), \qquad \mathbf{M}_3 = \sum_{j \in [k]} \lambda_j (\mathbf{a}_j \otimes \mathbf{a}_j \otimes \mathbf{a}_j). \tag{11}$$

for a set of κ_j's and λ_j's which are a function of the parameters of the distribution.

Remark 1: efficient computation of v, v_1 and v_2: What makes Theorem 1 specially intriguing is the fact that weights v, v_1 and v_2 can be computed through univariate integration, which can be computed efficiently, regardless of the dimensionality of the problem.

Remark 2: investigation of special cases When the ID distribution is Gamma with parameters $(\alpha_i, 1)$, we have the Dirichlet distribution as the resulting NID distribution. Weights v_1 and v_2 reduce to the results of Anandkumar et. al. [2] for the Gamma$(\alpha_i, 1)$ distribution, which are $v_1 = -\frac{\alpha_0}{\alpha_0+2}$ and $v_2 = \frac{2\alpha_0^2}{(\alpha_0+2)(\alpha_0+1)}$. When the variables are drawn from the positive stable distribution $St(1/2, \beta, \alpha_i, \mu)$ weights v_1 and v_2 in Theorem 1 can be represented in closed form as $v_1 = -\frac{1}{4}$ and $v_2 = -\frac{5}{8}$.

It is hard to find closed form representation of the weights for other stable distributions and the Inverse Gaussian distribution. Therefore, we give the form of the weights with respect to the parameters of each distribution in Figure 4. As it can be seen in Figures 2e and 2b, as γ increases, the distribution gets more centralized on the simplex. Therefore, as depicted in Figure 4a the weight becomes more negative to compensate for it. The same holds in Figure 4b.

The above result immediately implies guaranteed learning for non-degenerate topic-word matrix $\mathbf{A}$.

Assumption 2 *Topic-word matrix $\mathbf{A} \in \mathbb{R}^{d \times k}$ has linearly independent columns and the parameters $\alpha_i > 0$.*

Corollary 1 (Guaranteed Learning of NID Topic Models using Spectral Methods) *Given empirical versions of moments $\mathbf{M}_2$ and $\mathbf{M}_3$ in* (6) *and* (7)*, using tensor decomposition algorithm from [3], under the above assumption, we can consistently estimate topic-word matrix $\mathbf{A}$ and parameters $\boldsymbol{\alpha}$ with polynomial computational and sample complexity.*

The overall procedure is given in Algorithm 1.

Remark 3: third order moments suffice For the flexible class of latent NID topic models, only moments up to the third order suffice for efficient learning.

Remark 4: hyperparameter tuning In practice, we can tune for hyperparameters to compute the best fitting v, v_1 and v_2. This is based on their expressions in (9). If we do not want to limit

Algorithm 1 Parameter Learning

Require: Chosen NID distribution and hidden dimension k
Ensure: Parameters of NID distribution $\boldsymbol{\alpha}$ and topic-word matrix $\mathbf{A}$

1: Estimate empirical moments $\hat{\mathbb{E}}(\mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \mathbf{x}_3)$,$\hat{\mathbb{E}}(\mathbf{x}_1 \otimes \mathbf{x}_2)$ and $\hat{\mathbb{E}}(\mathbf{x}_1)$.
2: Compute weights v, v_1 and v_2 in (9) and (10) for the given NID distribution by numerical integration.
3: Estimate tensors $\mathbf{M}_2$ and $\mathbf{M}_3$ in (6) and (7) .
4: Decompose tensor $\mathbf{M}_3$ into its rank-1 components using the algorithm in [3] that requires $\mathbf{M}_2$.
5: Return columns of $\mathbf{A}$ as the components of the decomposition and $\boldsymbol{\alpha}$ as the set of weights.

ourselves to a single parametric NID family, we can employ a non-parametric estimation of the Lévy-Khintchine representation. Since this is one-dimensional, a small number of parameters will suffice for good performance.

Overview of the proof of Theorem 1 We begin the proof by forming the following second order and third order tensors using the moments of the NID distribution given in Lemma 1.

$$\mathbf{M}_2^{(\mathbf{h})} = \mathbb{E}(\mathbf{h} \otimes \mathbf{h}) + v\mathbb{E}(\mathbf{h}) \otimes \mathbb{E}(\mathbf{h}), \tag{12}$$

$$\begin{aligned}\mathbf{M}_3^{(\mathbf{h})} &= \mathbb{E}(\mathbf{h} \otimes \mathbf{h} \otimes \mathbf{h}) + v_2\mathbb{E}(\mathbf{h}) \otimes \mathbb{E}(\mathbf{h}) \otimes \mathbb{E}(\mathbf{h}) \\ &+ v_1\mathbb{E}(\mathbf{h} \otimes \mathbf{h}) \otimes \mathbb{E}(\mathbf{h}) + v_1\mathbb{E}(\mathbf{h} \otimes \mathbb{E}(\mathbf{h}) \otimes \mathbf{h}) + v_1\mathbb{E}(\mathbf{h}) \otimes \mathbb{E}(\mathbf{h} \otimes \mathbf{h})\end{aligned} \tag{13}$$

Weights v, v_1 and v_2 are as in Equations (9) and (10). They are computed by setting the off-diagonal entries of matrix $\mathbf{M}_2^{(\mathbf{h})}$ in Equation 12 and $\mathbf{M}_3^{(\mathbf{h})}$ in Equation 13 to 0. Due to the homogeneity assumption, all the off-diagonal entries can be simultaneously made to vanish with these choices of coefficients for v, v_1 and v_2. We obtain $\mathbf{M}_2^{(\mathbf{h})} = \sum_{i\in[k]} \kappa_i' \mathbf{e}_i^{\otimes 2}$ and $\mathbf{M}_3^{(\mathbf{h})} = \sum_{i\in[k]} \lambda_i' \mathbf{e}_i^{\otimes 3}$ where $\mathbf{e}_i$'s are the standard basis vectors, and this implies they are diagonal tensors. Due to this fact and the exchangeability of the words given topics according to (2), Equations 11 follow.

The exact forms of v, v_1 and v_2 are obtained by the following moment forms for NID distributions.

Lemma 1 ([13]) *The moments of NID variables $h_1, \dots h_k$ satisfy*

$$\mathbb{E}(h_1^{r_1} h_2^{r_2} \dots h_k^{r_k}) = \frac{1}{\Gamma(r)} \int_0^\infty u^{r-1} e^{-\alpha_0 \Psi(u)} \prod_{j\in[k]} B_{r_j}^j \mathrm{d}u, \tag{14}$$

where $r = \sum_{i\in[k]} r_i$ and $B_{r_j}^j$ can be written in terms of the partial Bell polynomial as

$$B_r^i = B_r(-\alpha_i \Psi^{(1)}(u), \dots, -\alpha_i \Psi^{(r)}(u)), \tag{15}$$

in which $\Psi^{(l)}(u)$ is the l-th derivative of $\Psi(u)$ with respect to u.

5 Conclusion

In this paper we introduce the new class of Latent Normalized Infinitely Divisible (NID) topic models that generalizes previously proposed topic models such as LDA. We provide guaranteed efficient learning for this class of distributions using spectral methods through untangling the dependence of the hidden topics. We provide evidence that our proposed NID topic model overcomes the shortcomings of the Dirichlet distribution by allowing for both positive and negative correlations among the topics. We provide a guaranteed learning result, despite not having closed form densities for NID variables.

Future extensions Our goal is to go beyond the partial homogeneity assumption and extend to more general distributions. Moreover, we plan to investigate other constraints than the simplex. We plan to allow for general functions that are composed of independent variables, and to design mechanisms to learn them efficiently.

References

[1] RM Adelson. Compound poisson distributions. *OR*, 17(1):73–75, 1966.

[2] Anima Anandkumar, Yi-kai Liu, Daniel J Hsu, Dean P Foster, and Sham M Kakade. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925, 2012.

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

[4] Ali Shojaee Bakhtiari and Nizar Bouguila. Online learning for two novel latent topic models. In *Information and Communication Technology*, pages 286–295. Springer, 2014.

[5] David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[7] Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng, and Bo Zhang. Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems*, pages 2445–2453, 2013.

[8] Stefano Favaro, Georgia Hadjicharalambous, and Igor Prünster. On a class of distributions on the simplex. *Journal of Statistical Planning and Inference*, 141(9):2987–3004, 2011.

[9] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[10] Achim Klenke. Infinitely divisible distributions. In *Probability Theory*, pages 331–349. Springer, 2014.

[11] Michalis Kolossiatis, Jim E Griffin, and Mark FJ Steel. Modeling overdispersion with the normalized tempered stable distribution. *Computational Statistics & Data Analysis*, 55(7):2288–2301, 2011.

[12] Antonio Lijoi, Ramsés H Mena, and Igor Prünster. Hierarchical mixture modeling with normalized inverse-gaussian priors. *Journal of the American Statistical Association*, 100(472):1278–1291, 2005.

[13] Francesca Mangili and Alessio Benavoli. New prior near-ignorance models on the simplex. *International Journal of Approximate Reasoning*, 56:278–306, 2015.

[14] David Mimno, Hanna M Wallach, and Andrew McCallum. Gibbs sampling for logistic normal topic models with graph-based priors. 2008.

[15] Issei Sato and Hiroshi Nakagawa. Topic models with power-law using pitman-yor process. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–682. ACM, 2010.

[16] Hsiao-Yu Tung and Alex J Smola. Spectral methods for indian buffet process inference. In *Advances in Neural Information Processing Systems*, pages 1484–1492, 2014.

Appendix

Proof of Theorem 1 *Proof:* The moment form of Lemma 1 can be represented as [13],

$$\mathbb{E}(h_1^{r_1} h_2^{r_2} \dots h_n^{r_n}) = \frac{1}{\Gamma(r)} \int_0^\infty u^{r-1} e^{-\sum_{i=n+1}^{k} \Psi_i(u)} \prod_{j\in[n]} (-1)^{r_j} \frac{\mathrm{d}^{r_j}}{\mathrm{d}u^{r_j}} e^{-\Psi_j(u)} \mathrm{d}u. \tag{16}$$

We use the above general form of the moments to compute and diagonalize the following moment tensors,

$$\mathbf{M}_2^{(\mathbf{h})} = \mathbb{E}(\mathbf{h}\otimes\mathbf{h}) + \eta\mathbb{E}(\mathbf{h})\otimes\mathbb{E}(\mathbf{h}), \tag{17}$$

$$\begin{aligned}\mathbf{M}_3^{(\mathbf{h})} &= \mathbb{E}(\mathbf{h}\otimes\mathbf{h}\otimes\mathbf{h}) \\ &+ \eta_1\mathbb{E}(\mathbf{h}\otimes\mathbf{h})\otimes\mathbb{E}(\mathbf{h}) + \eta_2\mathbb{E}(\mathbf{h}\otimes\mathbb{E}(\mathbf{h})\otimes\mathbf{h}) + \eta_3\mathbb{E}(\mathbf{h})\otimes\mathbb{E}(\mathbf{h}\otimes\mathbf{h}) \\ &+ \eta_4\mathbb{E}(\mathbf{h})\otimes\mathbb{E}(\mathbf{h})\otimes\mathbb{E}(\mathbf{h}).\end{aligned} \tag{18}$$

Setting the off-diagonal entries of Equations (17) and (18) to 0 and get the following set of equations

$$\mathbb{E}(h_ih_j) + \eta\mathbb{E}(h_i)\mathbb{E}(h_j) = 0 \qquad \text{for} \quad i\neq j, \tag{19}$$

$$\mathbb{E}(h_ih_jh_l) + \eta_1\mathbb{E}(h_ih_j)\mathbb{E}(h_l) + \eta_2\mathbb{E}(h_ih_l)\mathbb{E}(h_j) + \eta_3\mathbb{E}(h_jh_l)\mathbb{E}(h_i) + \eta_4\mathbb{E}(h_i)\mathbb{E}(h_j)\mathbb{E}(h_l) = 0 \quad \text{for} \quad i\neq j\neq l = 0, \tag{20}$$

$$\mathbb{E}(h_i^2h_l) + \eta_1\mathbb{E}(h_i^2)\mathbb{E}(h_l) + \eta_2\mathbb{E}(h_ih_l)\mathbb{E}(h_i) + \eta_3\mathbb{E}(h_ih_l)\mathbb{E}(h_i) + \eta_4\mathbb{E}(h_i)\mathbb{E}(h_i)\mathbb{E}(h_l) = 0 \quad \text{for} \quad i\neq l. \tag{21}$$

Writing the moments using Equation (16), assuming $\Phi_i(u) = \alpha_i\Psi(u)$, we get the following weights by some simple algebraic manipulations,

$$\eta = \frac{\int_0^\infty ue^{-\alpha_0\Psi(u)}\left(\frac{\mathrm{d}}{\mathrm{d}u}\Psi(u)\right)^2\mathrm{d}u}{\left(\int_0^\infty e^{-\alpha_0\Psi(u)}\frac{\mathrm{d}}{\mathrm{d}u}\Psi(u)\mathrm{d}u\right)^2} \tag{22}$$

$$\eta_1 = \eta_2 = \eta_3 = -\frac{\frac{1}{2}\int_0^\infty u^2e^{-\alpha_0\Psi(u)}\frac{\mathrm{d}^2}{\mathrm{d}u^2}\Psi(u)\frac{\mathrm{d}}{\mathrm{d}u}\Psi(u)\mathrm{d}u}{\int_0^\infty ue^{-\alpha_0\Psi(u)}\frac{\mathrm{d}^2}{\mathrm{d}u^2}\Psi(u)\mathrm{d}u\int_0^\infty e^{-\alpha_0\Psi(u)}\frac{\mathrm{d}}{\mathrm{d}u}\Psi(u)\mathrm{d}u} \tag{23}$$

$$\eta_4 = \frac{-\frac{1}{2}\int_0^\infty u^2e^{-\alpha_0\Psi(u)}\left(\frac{\mathrm{d}}{\mathrm{d}u}\Psi(u)\right)^3\mathrm{d}u + (\eta_1+\eta_2+\eta_3)\int_0^\infty ue^{-\alpha_0\Psi(u)}\left(\frac{\mathrm{d}}{\mathrm{d}u}\Psi(u)\right)^2\mathrm{d}u\int_0^\infty e^{-\alpha_0\Psi(u)}\frac{\mathrm{d}}{\mathrm{d}u}\Psi(u)\mathrm{d}u}{\left(\int_0^\infty e^{-\alpha_0\Psi(u)}\frac{\mathrm{d}}{\mathrm{d}u}\Psi(u)\mathrm{d}u\right)^3} \tag{24}$$

Setting $v = \eta$, $v_1 = \eta_1 = \eta_2 = \eta_3$ and $v_2 = \eta_4$ and defining

$$\Omega(m,n,p) := \int_0^\infty u^m \frac{\mathrm{d}^n}{\mathrm{d}u^n}\Psi(u)\Big(\frac{\mathrm{d}}{\mathrm{d}u}\Psi(u)\Big)^p e^{-\alpha_0\Psi(u)}\mathrm{d}u, \tag{25}$$

the set of weights v, v_1 and v_2 have the following form,

$$v = \frac{\Omega(1,1,1)}{\Big(\Omega(0,1,0)\Big)^2}, \tag{26}$$

$$v_1 = -\frac{\Omega(2,2,1)}{2\Omega(1,2,0)\Omega(0,1,0)}, \tag{27}$$

$$v_2 = \frac{-0.5\Omega(2,1,2) + 3v_1\Omega(1,1,1)\Omega(0,1,0)}{\Big(\Omega(0,1,0)\Big)^3}. \tag{28}$$

(29)

Weights v, v_1 and v_2 ensure that moment tensors $\mathbf{M}_2^{(\mathbf{h})}$ and $\mathbf{M}_3^{(\mathbf{h})}$ form diagonal tensors. Therefore they can be represented as,

$$\mathbf{M}_2^{(\mathbf{h})} = \sum_{i \in [k]} \kappa'_i \mathbf{e}_i^{\otimes 2}, \tag{30}$$

$$\mathbf{M}_3^{(\mathbf{h})} = \sum_{i \in [k]} \lambda'_i \mathbf{e}_i^{\otimes 3}. \tag{31}$$

The exchangeability assumption on the word space gives,

$$\mathbb{E}[\mathbf{x}_1] = \mathbb{E}\big(\mathbb{E}[\mathbf{x}_1|\mathbf{h}]\big) = \mathbf{A}\mathbb{E}(\mathbf{h}), \tag{32}$$

$$\mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2] = \mathbb{E}\big(\mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2|\mathbf{h}]\big) = \mathbf{A}\mathbb{E}(\mathbf{h} \otimes \mathbf{h})\mathbf{A}^\top, \tag{33}$$

$$\mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \mathbf{x}_3] = \mathbb{E}\big(\mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \mathbf{x}_3|\mathbf{h}]\big) = \mathbb{E}[\mathbf{h} \otimes \mathbf{h} \otimes \mathbf{h}](\mathbf{A}, \mathbf{A}, \mathbf{A}). \tag{34}$$

Therefore,

$$\mathbf{M}_2 = \mathbf{A}\mathbf{M}_2^{(\mathbf{h})}\mathbf{A}^\top = \sum_{j \in [k]} \kappa_j (\mathbf{a}_j \otimes \mathbf{a}_j), \tag{35}$$

$$\mathbf{M}_3 = \mathbf{M}_3^{(\mathbf{h})}(\mathbf{A}, \mathbf{A}, \mathbf{A}) = \sum_{j \in [k]} \lambda_j (\mathbf{a}_j \otimes \mathbf{a}_j \otimes \mathbf{a}_j) \tag{36}$$

□