Dirichlet Simplex Nest and Geometric Inference

Mikhail Yurochkin [1 2 *] **Aritra Guha** [3 *] **Yuekai Sun** [3] **XuanLong Nguyen** [3]

Abstract

We propose Dirichlet Simplex Nest, a class of probabilistic models suitable for a variety of data types, and develop fast and provably accurate inference algorithms by accounting for the model's convex geometry and low dimensional simplicial structure. By exploiting the connection to Voronoi tessellation and properties of Dirichlet distribution, the proposed inference algorithm is shown to achieve consistency and strong error bound guarantees on a range of model settings and data distributions. The effectiveness of our model and the learning algorithm is demonstrated by simulations and by analyses of text and financial data.[1]

1. Introduction

For many complex probabilistic models, especially those with latent variables, the probability distribution of interest can be represented as an element of a convex polytope in a suitable ambient space, for which model fitting may be cast as the problem of finding the extreme points of the polytope. For instance, a mixture density can be identified as a point in a convex set of distributions whose extreme points are the mixture components. In the well-known topic model (Blei et al., 2003) for text analysis, a document corresponds to a point drawn from the topic polytope, its extreme points are the topics to be inferred. This convex geometric viewpoint provides the basis for posterior contraction behavior analysis of topic models, as well as developing fast geometric inference algorithms (Nguyen, 2015; Tang et al., 2014; Yurochkin & Nguyen, 2016; Yurochkin et al., 2017).

The basic topic model – the Latent Dirichlet Allocation (LDA) of Blei et al. (2003), as well as the comparable finite admixtures developed in population genetics (Pritchard et al., 2000) were originally designed for categorical data. However, there are many real world applications in which the convex geometric probabilistic modeling continues to be a sensible approach, even if observed measurements are no longer discrete-valued, but endowed with a variety of distributions. To expand the scope of admixture modeling for a variety of data types, we propose to study Dirichlet Simplex Nest (DSN), a class of probabilistic models that generalizes the LDA, and to develop fast and provably accurate inference algorithms by accounting for the model's convex geometry and its low dimensional simplicial structure.

The generative process given by a DSN is simple to describe: starting from a simplex $\mathscr{B}$ of K vertices embedded in a high-dimensional ambient space $\mathcal{S}$, one draws random points from the $\mathscr{B}$'s relative interior according to a Dirichlet distribution. Given each such point, a data point is generated according to a suitable probability kernel F. For the general simplex nest, $\mathcal{S}$ can be any vector space of dimensions $D \geq K-1$, while the probability kernel F can be taken to be Gaussian, Multinomial, Poisson, etc, depending on the nature of the observed data (continuous, categorical or counts, resp.). If $\mathcal{S}$ is standard probability simplex, and F a Multinomial distribution over categories, then the model is reduced to the familiar LDA model of Blei et al. (2003).

Although several geometric aspects of the DSN can be found in a vast array of well-known models in the literature, they were rarely treated together. First, viewing data as noisy observations from the low-dimensional affine hull that contains $\mathscr{B}$, our model shares an assumption that can be found in both classical factor analysis and non-negative matrix factorization (NMF) models (Lee & Seung, 2001), as well as the work of Anandkumar et al. (2012); Arora et al. (2012b) arising in topic models. Second, the convex constraints (i.e., linear weights of a convex combination are non-negative and sum to one) are present in all latent variable probabilistic modeling, even though most dominant computational approaches to inference such as MCMC sampling (Griffiths & Steyvers, 2004) and variational inference (Blei et al., 2003; Hoffman et al., 2013; Kucukelbir et al., 2017) do not appear to take advantage of the underlying convex geometry.

As is the case with topic models, scalable parameter estimation is a key challenge for the Dirichlet Simplex Nest. Thus, our main contribution is a novel inference algorithm that

[*]Equal contribution [1]IBM Research, Cambridge [2]MIT-IBM Watson AI Lab [3]Department of Statistics, University of Michigan. Correspondence to: Mikhail Yurochkin <mikhail.yurochkin@ibm.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

[1]Code: `https://github.com/moonfolk/VLAD`

accounts for the convex geometry and low dimensionality of the latent simplex structure endowed with a Dirichlet distribution. Starting with an original geometric technique of Yurochkin & Nguyen (2016), we present several new ideas allowing for more effective learning of asymmetric simplicial structures and the Dirichlet's concentration parameter for the general DSN model, thereby expanding its applicability to a broad range of data distributions. We also establish statistical consistency and estimation error bounds for the proposed algorithm.

The paper proceeds as follows. Section 2 describes Dirichlet Simplex Nest models and reviews existing geometric inference techniques. Section 3 elucidates the convex geometry of the DSN via its connection to the Voronoi Tessellation of simplices and the structure of Dirichlet distribution on low-dimensional simplices. This helps motivate the proposed Voronoi Latent Admixture (VLAD) algorithm. Theoretical analysis of VLAD is given in Section 4. Section 5 presents an exhaustive comparative study on simulated and real data. We conclude with a discussion in Section 6.

2. Dirichlet Simplex Nest

We proceed to formally describe Dirichlet Simplex Nest as a generative model. Let $\beta_1, \dots, \beta_K \in \mathcal{S}$ be K elements in a D-dimensional vector space $\mathcal{S}$, and define $\mathscr{B} = \text{Conv}(\beta_1, \dots, \beta_K)$ as their convex hull. When $K \leq D+1$, $\mathscr{B}$ is a simplex in general positions. Next, for each $i = 1, \dots, n$, generate a random vector $\mu_i \in \mathscr{B}$ by taking $\mu_i := \sum_{k=1}^K \theta_{ik}\beta_k$, where the corresponding coefficient vector $\theta_i = (\theta_{i1}, \dots, \theta_{iK}) \in \Delta^{K-1}$ is generated by letting $\theta_i \sim \text{Dir}_K(\alpha)$ for some concentration parameter $\alpha \in \mathbb{R}_+^K$. Now, given μ_i the data point x_i is generated by $x_i|\mu_i \sim F(\cdot \mid \mu_i)$, where F is a given probability kernel such that $\mathbb{E}[x_i \mid \theta_i] = \mu_i$ for any $i = 1, \dots, n$.

Relation to existing models The DSN encompasses several existing models in the literature. If we set $\mathcal{S} := \Delta^{D-1}$ and likelihood kernel $F(\cdot)$ to Multinomial, then we recover the LDA model (Blei et al., 2003). Other specific instances include Gaussian-Exponential (Schmidt et al., 2009) and Poisson-Gamma models (Cemgil, 2009).

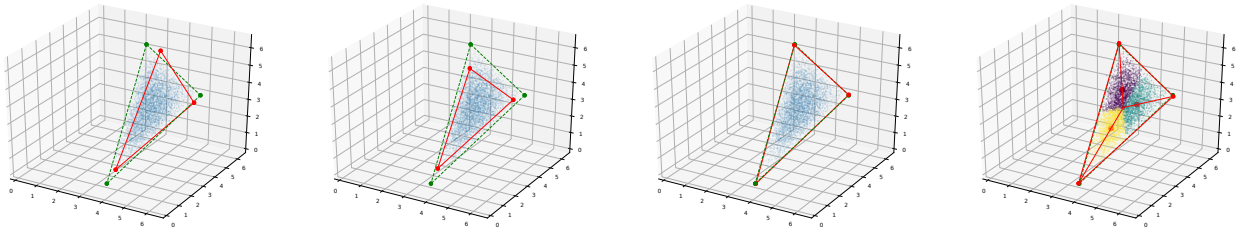
Estimating $\mathscr{B}$ is a challenging task for the general Dirichlet Simplex Nest model. Taking the perspective of Bayesian inference, a standard MCMC implementation for the DSN is likely computationally inefficient. In the case of LDA, as noted in Yurochkin & Nguyen (2016), the inefficiency of posterior inference can be traced to the need for approximating the posterior distributions of the large number of latent variables representing the topic labels. With the DSN model, we bypass the representation of such latent variables by integrating them out, but doing so at the cost of losing conjugacy. An alternative technique is variational inference (cf. Blei et al. (2017); Paisley et al. (2014)). While very fast, this powerful method may be inaccurate in practice and does not carry a strong theoretical guarantee.

Relation to NMF and archetypal analysis The DSN provides a probabilistic justification for these methods, which often impose an additional geometric condition on the model known as *separability* that identifies the model parameters in a way that permits efficient estimation (Donoho & Stodden, 2003; Arora et al., 2012a; Gillis & Vavasis, 2012). Separability is somewhat related to a control on the Dirichlet's concentration parameter α, by setting α be sufficiently small. The DSN allows for a probabilistic description of the nature of the separation. Moreover, by addressing *also* the case where α is large, the DSN modeling provides an arguably more effective approach to archetypal analysis and non-negative matrix factorization for *non-separable* data. We remark that an approach proposed by (Huang et al., 2016) also permits a more general geometric identification condition called sufficiently scattered, but this generality comes at the expense of efficient estimation.

Geometric inference Geometric Dirichlet Means (GDM) algorithm of Yurochkin & Nguyen (2016) is a geometric technique for estimating the (topic) simplex $\mathscr{B}$ that arises in the Latent Dirichlet Allocation model. The basic idea of GDM is simple: performing the K-means clustering algorithm on the n points μ_i (or their estimates) to obtain K centroids. These centroids cannot be a good estimate for $\mathscr{B}$'s vertices, but they provide reasonable directions toward the vertices. Starting from the simplex's estimated centroid, the GDM constructs K line segments connecting to the K centroids and suitably extends the rays to provide an estimate for the K vertices. The GDM method is shown to be accurate when either $\mathscr{B}$ is equilateral, or the Dirichlet concentration parameter α is very small, i.e., most of the points μ_is are concentrated near the vertices. The quality of the estimates deteriorates in the absence of such conditions.

The deficiency of the GDM algorithm can be attributed to several factors: first, for a general simplex, the K-means centroids and the simplex's vertices do not line up. Fortunately, we will see that they may be lined up in a straight line by a suitable affine transformation of the simplex structure. Second, the nature of the Dirichlet distribution on the simplex is not pro-actively exploited, including that of parameter α. Third, typically $K \ll D$, the affine hull of $\mathscr{B}$ is a very low-dimensional structure, a fact not utilized by the GDM algorithm. It turns out that these shortcomings may be overcome by a careful consideration of the geometric structure of the simplex and the Dirichlet distribution.

For illustrations, we consider a toy problem of learning extreme points of simplex $\mathscr{B}$, given Gaussian data likelihood $x_i|\mu_i \sim \mathcal{N}(\mu_i, \sigma^2 I_D)$ and $D = K = 3$. The triangle is



(a) GDM; time ≈ 1s (b) Xray; time < 1s (c) HMC; time ≈ 10m (d) VLAD; time < 1s

Figure 1: Toy simplex learning: $n = 5000, D = 3, K = 3, \alpha = 2.5, \sigma = 0.1$.

chosen to be non-equilateral and Dirichlet concentration parameter is set to $\alpha = 2.5$. Figure 1a illustrates the deteriorating performance of the GDM. In Figure 1b, we also observe Xray (Kumar et al., 2013), another recent NMF algorithm, failing to solve the problem, as the aforementioned separability assumption is violated for large α. On the other hand, Figure 1c demonstrates the high accuracy of the posterior mean obtained by Hamiltonian Monte Carlo (HMC) (Neal et al., 2011; Hoffman & Gelman, 2014) implemented using Stan (Carpenter et al., 2017), albeit at the cost of 10 minutes training time. Lastly our new algorithm (VLAD) in Fig. 1d, exhibits an accuracy comparable to that of the HMC and the run-time of the GDM algorithm.

3. Inference of the Dirichlet Simplex Nest

3.1. Simplicial Geometry

In order to motivate our algorithm, we elucidate the geometry of the DSN through the concept of Centroidal Voronoi Tessellation (CVT) (Du et al., 1999) of a simplex $\mathscr{B}$, a convex subset of D-dimensional metric space $\mathcal{S}$.

Definition 1 (Centroidal Voronoi Tessellation)**.** Let $\Omega \subset \mathcal{S}$ be an open set equipped with a distance function d and a probability density ρ. For a set of K points $c_1, \ldots, c_K$, the Voronoi cell corresponding to c_k is the set

$$V_k = \{x \in \Omega : d(x, c_k) < d(x, c_l) \text{ for any } l \neq k\}.$$

The collection of Voronoi cells $V_1, \ldots, V_K$ is a tessellation of Ω; i.e. the cells are disjoint and $\cup_k V_k = \Omega$. If $c_1, \ldots, c_K$ are also the centroids of their respective Voronoi cells, i.e.,

$$c_k = \frac{1}{\int_{V_k} \rho(x)\mathrm{d}x} \int_{V_k} x\rho(x)\mathrm{d}x$$

the tessellation is a Centroidal Voronoi Tessellation.

CVTs are special: any set of k points induces a Voronoi tessellation, but these points are generally not the centroids of their associated cells. One can check that a CVT minimizes

$$J(c_1, \ldots, c_K) = \int_{V_k} d(x, c_k)^2 \rho(x)\mathrm{d}x.$$

It is a fact that J has a unique global minimizer as long as ρ vanishes on a set of measure zero, the Voronoi cells are convex, and the distance function is convex in each argument (Du et al., 1999). Moreover, it can be seen that the centroids of the CVT of an equilateral simplex equipped with the $\mathrm{Dir}_K(\alpha)$ distribution fall on the line segments between the centroid of the simplex and the extreme points of the simplex, but this is not the case when the simplex shape is non-equilateral (cf. Fig. 1a).

The following lemma formalizes the aforementioned insight to a simplex of arbitrary shape $\mathscr{B}$ by considering a suitably modified distance function $d(\cdot, \cdot)$ of the CVT. (In Fig. 1d, the blue, purple and yellow dots are the sample versions of the Voronoi cells of the CVT under the new distance function and the corresponding centroids are in red.)

Lemma 1. Let $B \in \mathbb{R}^{D \times K}$ denote the matrix form of simplex $\mathscr{B}$. Suppose it has full (column) rank, equipped with distance function $\|\cdot\|_{(BB^T)^\dagger}$ and the probability distribution $\mathbb{P}_B$ defined as

$$\mathbb{P}_B(S) = \text{Prob}(\{\theta \in \Delta^{K-1} : B\theta \in S\}),$$

where θ is distributed by symmetric Dirichlet density $\rho_\alpha := \mathrm{Dir}_K(\alpha)$, for any $S \subset \mathsf{int}(\mathscr{B})$, and $A^\dagger$ denotes a pseudo-inverse of A. The centroids of its CVT fall on the line segments connecting the centroid of $\mathscr{B}$ to $\beta_1, \ldots, \beta_K$.

Proof. Let $c_1, \ldots, c_K$ and $V_1, \ldots, V_K$ be the centroids and cells of the CVT of Δ^{K-1} equipped with Euclidean distance and $\mathrm{Dir}_K(\alpha)$ density ρ_α. It suffices to verify that $Bc_1, \ldots, Bc_K$ and $BV_1, \ldots, BV_K$ are the centroids and cells of the CVT of $\mathscr{B} = B\Delta^{K-1}$. By a change of variables formula,

$$\begin{aligned}
&\text{argmin}\left\{ \frac{\int_{BV_k} \|x - Bv\|^2_{(BB^T)^\dagger} \rho_\alpha(B^\dagger x) |\det(B^\dagger)| \mathrm{d}x}{\int_{V_k} \rho_\alpha(B^\dagger x) |\det(B^\dagger)| \mathrm{d}x} : v \in V_k \right\} \\
&= \text{argmin}\left\{ \frac{\int_{V_k} \|B\theta - Bv\|^2_{(BB^T)^\dagger} \rho_\alpha(\theta)\mathrm{d}\theta}{\int_{V_k} \rho_\alpha(\theta)\mathrm{d}\theta} : v \in V_k \right\} \\
&= \text{argmin}\left\{ \frac{\int_{V_k} \|\theta - v\|_2^2 \rho_\alpha(\theta)\mathrm{d}\theta}{\int_{V_k} \rho_\alpha(\theta)\mathrm{d}\theta} : v \in V_k \right\},
\end{aligned}$$

which we recognize as the centroids of the CVT of Δ^{K-1} under ℓ_2 metric. Since Δ^{K-1} is a standard simplex and therefore equilateral, the centroids of the CVT of equilateral

simplex fall on the line segments connecting the centroid of the simplex to its extreme points. □

Lemma 1 suggests an algorithm to estimate the extreme points of $\mathscr{B}$. First, estimate the centroids of the CVT of $\mathscr{B}$ (equipped with scaled Euclidean norm $\|\cdot\|_{(BB^T)^\dagger}$) and search along the rays extending from the centroid of $\mathscr{B}$ through the CVT centroids for the simplicial vertices.

3.2. The Voronoi Latent Admixture (VLAD) Algorithm

We first consider the noiseless problem, $F(\cdot \mid \mu) = \delta_\mu$. That is, $x_i = \mu_i$s are observed. In this case, Lemma 1 suggests estimating the CVT centroids by scaled K-means optimization:

$$\underset{c_1,\ldots,c_K}{\operatorname{argmin}} \big\{ \tfrac{1}{2} \sum_{k=1}^K \sum_{x_i \in V_k} (x_i - c_k)^T (BB^T)^\dagger (x_i - c_k) \big\}, \tag{1}$$

Unfortunately, the scaled Euclidean norm $\|\cdot\|_{(BB^T)^\dagger}$ is unknown. We propose an equivalent approach that does not depend on knowledge of BB^T.

In the noiseless case, observe that the population covariance matrix of the samples takes the form $\Sigma = BSB^T$, where S is the covariance matrix of a $\mathrm{Dir}(\alpha)$ random variable on Δ^{K-1}. By the standard properties of the $\mathrm{Dir}(\alpha)$ distribution, it can be seen that $S = \frac{1}{K(K\alpha+1)} P$, where $P = I_K - \frac{1}{K} \mathbf{1}_K \mathbf{1}_K^T$ is the centering matrix. Hence, knowledge of Σ will be sufficient because the centered data points x fall in $\mathsf{span}(\Sigma) = \mathsf{span}(BPB^T)$: For each (θ, x) pair,

$$\bar{x} := \underbrace{B\theta}_{x} - \underbrace{\tfrac{1}{K} B\mathbf{1}}_{\mathbb{E}[x]} = B\theta - \tfrac{1}{K} B\mathbf{1}(\underbrace{\mathbf{1}^T\theta}_{=1}) = BP\theta := B\bar{\theta}. \tag{2}$$

This suggests that the centroids of the CVT may be recovered by clustering the centered data points in the $\|\cdot\|_{\Sigma^\dagger}$-norm. This insight is formalized by

Lemma 2. The centroids of the CVT of simplex $\mathscr{B}$ under $\|\cdot\|_{(BB^T)^\dagger}$-norm are given by $\{c_k^* + c_0 | k = 1, \ldots, K\}$, where $(c_1^*, \ldots, c_K^*)$ solves the minimization

$$\min_{\substack{c_1,\ldots,c_K \\ V_1,\ldots,V_K}} \frac{1}{2} \sum_{k=1}^K \int_{x \in BV_k} (\bar{x} - c_k)^T \Sigma^\dagger (\bar{x} - c_k) \rho(x) \mathrm{d}x \tag{3}$$

and $c_0 = \int x\rho(x)\mathrm{d}x$ is the centroid of simplex $\mathscr{B}$.

Proof. We first show that (3) is equivalent to (unscaled) K-means clustering on Δ^{K-1}. Note that $\Sigma = \delta BPB^T$ for some $\delta > 0$. Without loss of generality, we restrict to c_k's in $\mathsf{span}\{BPB^T\}$. Write $c_k = BPv_k$ for $v_k \in \mathbb{R}^K$, for $k = 1, \ldots, K$. Recalling (2) and the fact P is a projector,

$$\begin{aligned} & (1/\delta) \textstyle\sum_{k=1}^K \int_{x \in BV_k} (\bar{x} - c_k)^T \Sigma^\dagger (\bar{x} - c_k) \rho(x) \mathrm{d}x \\ = \; & \textstyle\sum_{k=1}^K \int_{\theta \in V_k} (\bar{\theta} - v_k)^T PB^T \Sigma^\dagger BP (\bar{\theta} - v_k) \rho_\alpha(\theta) \mathrm{d}\theta \\ = \; & \textstyle\sum_{k=1}^K \int_{\theta \in V_k} (\bar{\theta} - v_k)^T P (\bar{\theta} - v_k) \rho_\alpha(\theta) \mathrm{d}\theta \\ = \; & \textstyle\sum_{k=1}^K \int_{\theta \in V_k} \|\bar{\theta} - Pv_k\|_2^2 \rho_\alpha(\theta) \mathrm{d}\theta. \end{aligned} \tag{4}$$

Since θ is distributed by the symmetric Dirichlet $\rho_\alpha = \mathrm{Dir}(\alpha)$ on Δ^{K-1}, the last equality entails that the optimal v_k's are the points which represent the barycentric coordinate of the centroids of the CVT of Δ^{K-1}. Thus, the optimal solution for $c_k = BPv_k$ represents the centroids of the CVT of simplex $\mathscr{B}$ under $\|\cdot\|_{(BB^T)^\dagger}$-norm (using the coordinating system that is centered at origin c_0). □

We proceed to address the optimization (3) applied to empirical data to arrive at Voronoi Latent Admixture (VLAD) algorithm in Algorithm 1. We utilize the singular value decomposition (SVD) of the centered data points to simplify computation. Let $\bar{X} \in \mathbb{R}^{n \times D}$ be the matrix whose rows are the centered data points and $\bar{X} = U\Lambda W^T$ be its SVD. Each term in the objective of (3) is equivalent to, with Σ being replaced by its empirical version, $\Sigma_n = \frac{1}{n} W\Lambda^2 W^T$:

$$\begin{aligned} & (\bar{x}_i - c_k)^T \Sigma_n^\dagger (\bar{x}_i - c_k) = \\ & n(u_i - \eta_k)^T \Lambda W^T W \Lambda^{-2} W^T W \Lambda (u_i - \eta_k) = n\|u_i - \eta_k\|_2^2, \end{aligned}$$

where $\bar{x}_i = W\Lambda u_i$, and set $c_k = W\Lambda\eta_k$. Thus, instead of performing scaled K-means clustering in $\mathcal{S}$, it suffices to perform standard K-means in the low $(K-1)$ dimensional space. This yields a significant computational speed-up. After applying VLAD, the weights θ_i's can be obtained by projecting the data points onto $\mathscr{B}$ and compute the barycentric coordinates of the projected points.

Algorithm 1 Voronoi Latent Admixture (VLAD)

Input: data $x_1, \ldots, x_n$; K; extension parameter γ.
Output: simplex vertices $\beta_1, \ldots, \beta_K$

1: $\widehat{c}_0 \leftarrow \frac{1}{n} \sum_i x_i$ {find data center}
2: $\bar{x}_i \leftarrow x_i - \widehat{c}_0$, $i = 1, \ldots, n$ {centering}
3: compute top $K-1$ singular factors of the centered data matrix $\bar{X} \in \mathbb{R}^{n \times D}$: $\bar{X} = U\Lambda W^T$
4: $\eta_1, \ldots, \eta_K \leftarrow$ K-means$(u_1, \ldots, u_n)$, where the u_i's are the rows of $U \in \mathbb{R}^{n \times (K-1)}$
5: $\widehat{c}_k \leftarrow W\Lambda\eta_k + \widehat{c}_0$
6: $\widehat{\beta}_k \leftarrow \widehat{c}_0 + \gamma(\widehat{c}_k - \widehat{c}_0)$

It remains to estimate the extreme points β_ks given the CVT centroids c_ks. This task is simplified by two observations: First, the CVT centroids reside on the line segment between the centroid of simplex $\mathscr{B}$ and its extreme points, per Lemma 1. Thus we merely need to estimate the ratio of

the distance from the extreme point to the centroids of $\mathscr{B}$ and the distance from the CVT centroids to the centroid of $\mathscr{B}$. Due to the symmetry of $\mathrm{Dir}_K(\alpha)$ distribution on Δ^{K-1}, this ratio is identical for all extreme points – we refer to this ratio as the extension parameter γ. Secondly, γ does not depend on the geometry of $\mathscr{B}$, only that of the Dirichlet distribution. Thus, γ can be easily estimated by appealing to a Monte Carlo technique on Dir_K. This subroutine is summarized in Algorithm 2, provided that α is given.

Algorithm 2 Evaluating extension parameters

1: generate $\theta_1, \ldots, \theta_m \sim \mathrm{Dir}_K(\alpha)$, where m is the number of Monte Carlo samples
2: $v_1, \ldots, v_K \leftarrow$ K-means$(\theta_1, \ldots, \theta_m)$
3: $\gamma \leftarrow \sqrt{K^2 - K} \left(\sum_{l=1}^{K} \| v_l - \frac{1}{K} \mathbf{1}_K \|_2 \right)^{-1}$

3.3. Estimating the Dirichlet Concentration Parameter

Next, we describe how to estimate concentration parameter α from the data, by employing a moment-based approach. Recall from the previous section that there is an one-to-one mapping between α and the extension parameter γ. For each $\alpha > 0$, let $\gamma(\alpha) > 0$ denote the corresponding extension parameter and $B(\gamma) \in \mathbb{R}^{D \times K}$ the estimator of B output by VLAD with extension parameter γ. In the absence of noise, the covariance matrix of the DSN model has the form $BS(\alpha)B^T$, where $S(\alpha) \in \mathbb{R}^{K \times K}$ is the covariance matrix of a $\mathrm{Dir}(\alpha)$ random variable on Δ^{K-1}. This suggests we estimate α by a generalized method of moments approach:

$$\hat{\alpha} = \operatorname*{argmin}_{\alpha > 0} \| \hat{B}(\gamma(\alpha)) S(\alpha) \hat{B}(\gamma(\alpha))^T - \hat{\Sigma} \|, \qquad (5)$$

where $\hat{\Sigma}$ is the sample covariance matrix $\hat{\Sigma} = \frac{1}{n} \bar{X}^T \bar{X}$. We remark that there is no need to run VLAD multiple times to evaluate the objective in (5) at multiple α-values. After VLAD is run once, we may evaluate $\gamma(\alpha)$ for any value of γ by affinely transforming the output of VLAD. Further, (5) is a scalar optimization problem, so the computational cost of solving (5) is negligible.

In the presence of noise, the covariance matrix of the DSN model no longer has the form $BS(\alpha)B^T$. We need to add a correction term to ensure a consistent estimator of $BS(\alpha)B^T$. For example, if the noise is Gaussian, a consistent estimator of $BS(\alpha)B^T$ is

$$\tilde{\Sigma} = \hat{\Sigma} - \hat{\sigma}^2 I_D,$$

where $\hat{\sigma}^2$ is an estimate of the noise variance. In the Supplement, we give consistent estimators of $BS(\alpha)B^T$ for multinomial and Poisson noise. With a good estimator $\tilde{\Sigma}$ of $BS(\alpha)B^T$ in place, we replace $\hat{\Sigma}$ in (5) by $\tilde{\Sigma}$ and then solve (5) to obtain an estimate of α.

4. Consistency and Estimation Error Bounds

In this section we establish consistency properties and error bound guarantees of the VLAD procedure.

For $c = (c_1, \ldots, c_K) \in \mathbb{R}^{K \times D}$, define $\phi_A : \mathbb{R}^D \times \mathbb{R}^{K \times D} \to \mathbb{R}$ as

$$\phi_A(x, c) = \min_{k \in \{1, \ldots, K\}} \| x - c_k \|_{A^\dagger}^2$$

where A is a positive semidefinite matrix. Recall Σ as the covariance matrix of the data generating distribution and Σ_n its empirical counterpart. In the algorithm, we work with the best rank $K-1$ approximation of Σ_n, which we denote by $(\Sigma_n)^K$. Let $\mathbb{Q}$ denote the distribution for μ_is. Recall that $X_i | \mu_i \sim F(\cdot | \mu_i)$. Let $\mathbb{P}$ be the induced distribution corresponding to $\tilde{X}_i$, which is the projection of X_i on the affine space of dimension $K-1$ spanned by the top $K-1$ eigenvectors of Σ. We also use $\mathbb{P}_n$ to denote the empirical distribution of the data represented by random variables X_i.

Since K-means clustering is a subroutine of our algorithm, we expect at least some sort of condition requiring that the K-means clustering routine be well-behaved in some sense. To that end we need the following standard condition on the population K-means objective (cf. Pollard (1981)).

(a.1) Pollard's regularity criterion (PRC): The Hessian matrix of the function $c \mapsto \mathbb{Q}\phi_{BSB^T}(\cdot, c)$ evaluated at c^* for all optimizers c^* of $\mathbb{Q}\phi_{BSB^T}(\cdot, c)$ is positive definite, with minimum eigenvalue $\lambda_0 > 0$.

It turns out that this will be all we need for the following theorem in the noiseless setting, where we have $\Sigma = BSB^T = (\Sigma)^K$ has rank $K-1$ and so, $\mathbb{P} = \mathbb{Q}$ and $\tilde{X}_i \overset{\mathscr{L}}{=} X_i$.

Theorem 1. Consider the noiseless setting, i.e., $F(\cdot \mid \mu) = \delta_\mu$. Suppose that $\mathscr{B} = \mathrm{Conv}(\beta_1, \ldots, \beta_K)$ is the true topic simplex, while $(\beta_{1n}, \ldots, \beta_{Kn})$ are the vertex estimates obtained by VLAD algorithm. Moreover, assume the error due to Monte Carlo estimates of the extension parameter is negligible. Provided that condition (a.1) holds,

$$\min_{\pi} \| (\beta_{\pi_{(1)}n}, \ldots, \beta_{\pi_{(K)}n}) - (\beta_1, \ldots, \beta_K) \| = O_{\mathbb{P}}(n^{-1/2}),$$

where the minimization is taken over all permutations π of $\{1, \ldots, K\}$.

Note that the constant corresponding to the rate $O_{\mathbb{P}}(n^{-1/2})$ is dependent on the Hessian matrix of the function $c \mapsto \mathbb{P}\phi_\Sigma(\cdot, c)$. The proof for Theorem 1 is in the Supplement.

In general, $F(\cdot \mid \mu)$ is not degenerate. Due to the presence of "noise" in the $K-1$ SVD subspace, the estimates of the CVT centroids may be inconsistent, which entails inconsistency of the VLAD's estimate for $\mathscr{B}$. The following theorem provides an error bound in the general setting. We need a

strengthening of Pollard's Regularity Criterion. Let $(\Sigma)^K$ denote the best $K-1$ rank approximation of Σ with respect to the Frobenius norm. Assume:

(a.2) The Hessian matrix of the function $c \mapsto \mathbb{P}\phi_{(\Sigma)^K}(\cdot, c)$ evaluated at c^* for all optimizers c^* of $\mathbb{P}\phi_{(\Sigma)^K}(\cdot, c)$ is uniformly positive definite with minimum eigenvalue $\lambda_0 > 0$, for all $(\Sigma)^K$ such that $(\Sigma - BSB^T) \leq \tilde{\epsilon} I_D$, for some $\tilde{\epsilon} > 0$.

The noise level is formalized by the following conditions:

(b) There is $\epsilon_0 > 0$ such that $\epsilon_0 I_D - Cov(X|\theta)$ is positive semi-definite uniformly over $\theta \in \Delta^{K-1}$.

(c) There exists M_0 such that for all $M > M_0$, $\int_{\mathcal{B}(\sqrt{M}, c_0)^c} \|x - c_0\|_2^2 g(x) \mathrm{d}x \leq \frac{k_1}{M}$, for some universal constant k_1, where $\mathcal{B}(\sqrt{M}, c_0)$ is a ball of radius $\sqrt{M}$ around population centroid c_0 and $g(\cdot)$ is the density of $\mathbb{P}$ with respect to the Lebesgue measure on the $K-1$ dimensional space which contains the top $K-1$ eigenvectors of $BSB^T + \epsilon_0 I_D$.

Theorem 2. Suppose that $\mathscr{B} = \mathrm{Conv}(\beta_1, \ldots, \beta_K)$ is simplex corresponding to extreme points of the DSN. Let $(\beta_{1n}, \ldots, \beta_{Kn})$ be the corresponding extreme point estimates obtained by the VLAD algorithm. Assume the error in the Monte Carlo estimates of the extension parameter is negligible. Provided that (a.2), (b) and (c) hold, then

$$\min_\pi \|(\beta_{\pi_{(1)}n}, \ldots, \beta_{\pi_{(K)}n}) - (\beta_1, \ldots, \beta_K)\|_2 = O\left(\sqrt{\epsilon_0^{1/3}/\lambda_0}\right) + O_{\mathbb{P}}(n^{-1/2}), \quad (6)$$

where π ranges over permutations of $\{1, \ldots, K\}$.

The constant corresponding to the rate $O_{\mathbb{P}}(n^{-1/2})$ in the above theorem, depends on the Hessian matrix of the function $c \mapsto \mathbb{P}\phi_\Sigma(\cdot, c)$. The constant corresponding to the $O\left(\sqrt{\epsilon_0^{1/3}/\lambda_0}\right)$ is dependent on the minimum and maximum eigenvalues of the matrix BSB^T.

The preceding results control the error incurred by the VLAD algorithm when the concentration parameter α is known. When α is unknown, our proposed solution in Section 3.3 performs well in both simulated and real-data experiments. We do not know in theory whether the concentration parameter α is identifiable, we shall present empirical results in the Supplement which suggest identifiability. Assuming a condition which guarantees model identifiability, we can establish that the estimate obtained by the VLAD algorithm via (5) is in fact consistent.

Theorem 3. Assume that function $\varphi(\tilde{\alpha}) = \frac{\gamma(\tilde{\alpha})^2}{K(K\tilde{\alpha}+1)}$ is monotonically increasing in $\tilde{\alpha}$, where $\gamma(\tilde{\alpha})$ is the extension parameter corresponding to $\tilde{\alpha}$. Let $\alpha_0 \in \mathscr{C}$ be the true concentration parameter for some compact set $\mathscr{C}$. Let $\hat{\alpha}_n = \mathrm{argmin}_{\alpha \in \mathscr{C}} \|\hat{B}(\gamma(\alpha))S(\alpha)\hat{B}(\gamma(\alpha))^T - \tilde{\Sigma}_n\|$, where $\tilde{\Sigma}_n$ is a consistent estimator of $BS(\alpha)B^T$. Then,

$$\|\hat{\alpha}_n - \alpha_0\| \xrightarrow{\mathbb{P}} 0. \quad (7)$$

Table 1: Baselines and required conditions

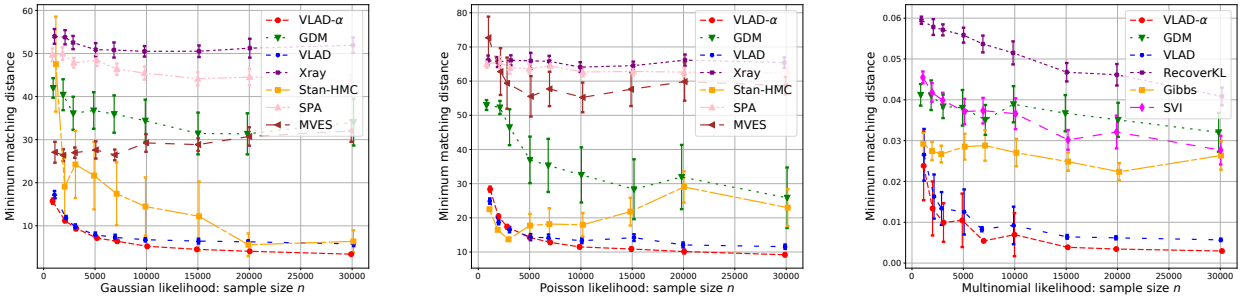
Method	Conjugacy	True α	Separability
VLAD (this work)	$\times$	$\times$	$\times$
VLAD-α (this work)	$\times$	$\surd$	$\times$
Gibbs (2004)	$\surd$	$\surd^\star$	$\times$
Stan-HMC (2017)	$\times$	$\surd^\star$	$\times$
SVI (2013)	$\surd$	$\surd^\star$	$\times$
GDM (2016)	$\times$	$\times$	$\surd^\star$
RecoverKL (2013)	$\times$	$\times$	$\surd$
SPA (2012)	$\times$	$\times$	$\surd$
MVES (2009)	$\times$	$\times$	$\surd$
Xray (2013)	$\times$	$\times$	$\surd$

5. Experiments

The goal of our experimental studies is to demonstrate the applicability and efficiency of our algorithm for a number of choices of the DSN probability kernel: Gaussian, Poisson and Multinomial (i.e. LDA). We summarize all competing estimation procedures in our comparative study and their corresponding underlying assumptions in Table 1.

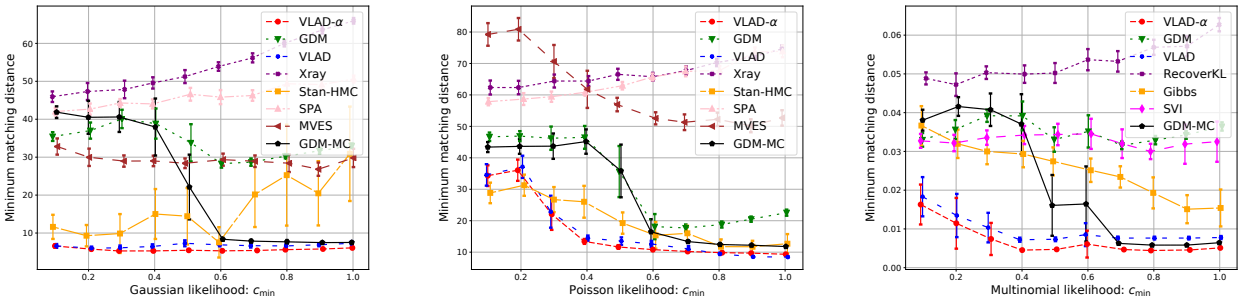
We remark that Gibbs sampler (Griffiths & Steyvers, 2004), Stan implementation of No U-Turn HMC (Hoffman & Gelman, 2014; Carpenter et al., 2017) and Stochastic Variational Inference (SVI) (Hoffman et al., 2013) may be augmented with techniques such as empirical Bayes to estimate hyperparameter α, although it may slow down convergence. We instead allow these baselines to use true values of α in all simulated experiments to their advantage; when latent simplex is of general geometry (i.e. non-equilateral), GDM (Yurochkin & Nguyen, 2016) requires $\alpha \to 0$ to perform well, which is alike separability. Not all baselines are suitable for all three probability kernels, i.e. Gibbs sampler and SVI rely on (local) conjugacy and are only applicable in the LDA scenario; RecoverKL (Arora et al., 2013) is an algorithm that relies on a separability condition (i.e. anchor words) designed for topic models.

In simulated experiments we will consider both VLAD with estimated concentration parameter α following our results in Section 3.3 and VLAD trained with the knowledge of true data generating α (VLAD-α). For real data analysis, we estimate the concentration parameter by (5) and apply VLAD to a text corpus and stock market data set.



(a) Gaussian data (b) Poisson data (c) Categorical data

Figure 2: Minimum matching distance for increasing n



(a) Gaussian data (b) Poisson data (c) Categorical data

Figure 3: Minimum matching distance for varying DSN geometry.

5.1. Comparative Simulation Studies

Convergence behavior We investigate the convergence of the estimates of the DSN extreme points for the three likelihood kernels under the increasing sample size. The hyperparameter settings are $D = 500, K = 10, \alpha = 2$ (for LDA vocabulary size $D = 2000$). To ensure non-trivial geometry of the DSN we rescale extreme points towards their mean by uniform random factors between 0.5 and 1. We use the Minimum Matching distance - a metric previously studied in the context of polytopes estimation (Nguyen, 2015) to compare the quality of the fitted DSN model returned by a variety of inference algorithms. We defer additional details to the supplement.

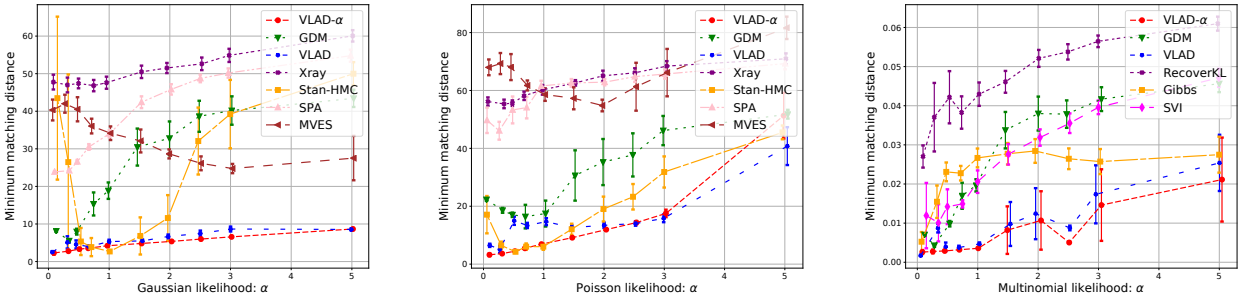
In Fig. 2 we see that VLAD and VLAD-α significantly outperform all baselines. Further, the estimation error reduces with increased sample size verifying statements of Theorems 2 and 3. We note that Stan HMC may also achieve good performance, however it is very costly to fit (e.g., 40 HMC iterations for Poisson case and $n = 30000$ took 14 hours compared to 7 seconds for VLAD), therefore we had to restrict number of iterations, which explains its wider error bars across experiments.

Geometry of the DSN To study the role of geometry of the DSN we rescale extreme points towards their mean by uniform random factors $c_k \sim \text{Unif}(c_{\min}, 1)$ for $k = 1, \ldots, K$ and vary $c_{\min}$ in Fig. 3 (smaller values imply more severe skewness of the latent simplex). To isolate the effect of the geometry of the DSN, we compare to GDM combined with knowledge of true α and extension parameter estimation using Algorithm 2 (GDM-MC). If the underlying simplex is equilateral, GDM-MC will be equivalent to VLAD-α.

In Fig. 3 we see that VLAD and VLAD-α are robust to varying skewness of the DSN. On the contrary, GDM-MC is only accurate when the latent simplex becomes closer to equilateral. This experiment verifies geometric motivation of our work — in practice we can not expect latent geometric structure to be necessarily equilateral and geometrically robust method such as VLAD is more reliable.

Varying Dirichlet prior To complete our simulation studies we verify α estimation procedure proposed in Section 3.3 and analyzed in Theorem 3. It is also interesting to compare performance of other baselines for larger α — scenario often overlooked in the literature.

In Fig. 4 (and in previous experiments) we see that performance gap between VLAD and VLAD-α is very small, supporting effectiveness of our α estimation procedure across probability kernels. Additionally, we see that higher values of α lead to degrading performance of all considered methods, however VLAD degrades more gracefully.



(a) Gaussian data (b) Poisson data (c) Categorical data

Figure 4: Minimum matching distance for increasing α.

Table 2: NYT topic modeling (categorical data) || Stock data factor analysis (continuous data)

	Perplexity	Coherence	Time	Frobenius norm	Volume	Time
VLAD	1767	0.86	**6min**	0.300	**0.14**	**1s**
GDM	1777	0.88	30min	0.294	1499	**1s**
Gibbs || HMC	**1520**	0.80	5.3hours	0.299	1.95	10min
RecoverKL || MVES	2365	0.70	17min	0.287	5.39×10^9	3min
SVI || SPA	1669	0.81	40min	0.392	3.31×10^7	**1s**

5.2. Real Data Analysis

Topic modeling We analyze a collection of news articles from the New York Times. After preprocessing, we have 5320 unique words and 100k training documents with 25k left out for perplexity evaluation. We also compare semantic coherence of the topics (Newman et al., 2010).

In Table 2 (left) we present results for $K = 80$ topics. The Gibbs sampler has the best perplexity score, but it falls behind in topic coherence. VLAD estimated $\alpha = 0.05$ and has approximately same perplexity and coherence as GDM, while being 5 times faster. VLAD identified contextually meaningful topics, as can be seen from good coherence score and by eye-balling the topics — they cover a variety of concepts from fishing and cooking to the Enron scandal and cancer. The top 20 words for each of the VLAD topics are provided along with the code.

Stock market analysis We collect variations (closure minus opening price) for 3400 days and 55 companies. We train several algorithms on data from the first 3000 days and report the average distance between the data points from the last 400 days and fitted simplices (i.e., Frobenius norm). This metric alone might be misleading since stretching any simplex will always reduce the score, therefore we also report the volumes of corresponding simplices. Results are summarized in Table 2 (right) — our method (estimated $\alpha = 0.05$) achieves comparable fit in terms of the Frobenius norm with a more compact simplex. Among the factors identified by VLAD, we notice a growth component related to banks (e.g., Bank of America, Wells Fargo). Another factor suggests that the performance of fuel companies like Valero Energy and Chevron are inversely related to the performance of defense contractors (Boeing, Raytheon).

6. Summary and Discussion

The Dirichlet Simplex Nest model generalizes a number of popular models in machine learning applications, including LDA and several variants of non-negative matrix factorization (NMF). We also develop an algorithm that exploits the geometry of the DSN to perform fast and accurate inference. We demonstrate the superior statistical and computational properties of the algorithm on several real datasets and verify its accuracy through simulations.

One of the key distinctions between the DSN model and NMF models is we replace the separability assumption by a Dirichlet prior on the weights. The main benefit of this approach is it enables us to model data that does not contain archetypal points (Cutler & Breiman, 1994). Among the limitations of our approach is the reliance on the Dirichlet distribution assumption in a crucial way, that the Dirichlet distribution is symmetric on the standard probability simplex Δ^{K-1}. In theory, the algorithm breaks down when the Dirichlet distribution is asymmetric. Surprisingly, in simulations at least, we found that VLAD seems quite robust in recovering the correct direction of extreme points, even as most existing methods break down in such situations. These findings are reported in the Supplement.

Acknowledgement Support provided by NSF grants DMS-1830247, CAREER DMS-1351362, CNS-1409303 and a Margaret and Herman Sokol Faculty Award are gratefully acknowledged.

References

Anandkumar, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Liu, Y.-K. A spectral algorithm for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*, pp. 917–925, 2012.

Arora, S., Ge, R., Kannan, R., and Moitra, A. Computing a nonnegative matrix factorization–provably. In *Proceedings of the 44th annual ACM symposium on Theory of computing*, pp. 145–162. ACM, 2012a.

Arora, S., Ge, R., and Moitra, A. Learning topic models – going beyond SVD. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pp. 1–10. IEEE, 2012b.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 280–288, 2013.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993–1022, March 2003.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

Cemgil, A. T. Bayesian inference for nonnegative matrix factorisation models. *Computational intelligence and neuroscience*, 2009, 2009.

Chan, T.-H., Chi, C.-Y., Huang, Y.-M., and Ma, W.-K. A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing. *IEEE Transactions on Signal Processing*, 57(11):4418–4432, 2009.

Cutler, A. and Breiman, L. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

Donoho, D. and Stodden, V. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems*, 2003.

Du, Q., Faber, V., and Gunzburger, M. Centroidal Voronoi Tessellations: Applications and algorithms. *SIAM Review*, 41(4):637–676, 1999.

Gillis, N. and Vavasis, S. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *arXiv preprint arXiv:1208.1237*, 2012.

Griffiths, T. L. and Steyvers, M. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

Hoffman, M. D. and Gelman, A. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, May 2013.

Huang, K., Fu, X., and Sidiropoulos, N. D. Anchor-Free Correlated Topic Modeling: Identifiability and Algorithm. *arXiv:1611.05010 [cs, stat]*, November 2016.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(1): 430–474, 2017.

Kumar, A., Sindhwani, V., and Kambadur, P. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *International Conference on Machine Learning*, pp. 231–239, 2013.

Lee, D. D. and Seung, H. S. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562, 2001.

Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108. Association for Computational Linguistics, 2010.

Nguyen, X. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, 21(1):618–646, 02 2015.

Paisley, J. W., Blei, D. M., and Jordan, M. I. Bayesian nonnegative matrix factorization with stochastic variational inference., 2014.

Pollard, D. Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135–140, 01 1981.

Pritchard, J. K., Stephens, M., and Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Schmidt, M. N., Winther, O., and Hansen, L. K. Bayesian non-negative matrix factorization. In *International Conference on Independent Component Analysis and Signal Separation*, pp. 540–547. Springer, 2009.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., and Zhang, M. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 190–198, 2014.

Yurochkin, M. and Nguyen, X. Geometric Dirichlet Means Algorithm for topic inference. In *Advances in Neural Information Processing Systems*, pp. 2505–2513, 2016.

Yurochkin, M., Guha, A., and Nguyen, X. Conic Scan-and-Cover algorithms for nonparametric topic modeling. In *Advances in Neural Information Processing Systems*, pp. 3881–3890, 2017.