

MIT Open Access Articles

Small-variance nonparametric clustering on the hypersphere

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Straub, Julian et al. "Small-Variance Nonparametric Clustering on the Hypersphere." IEEE, 2015. 334–342.

As Published: <http://dx.doi.org/10.1109/CVPR.2015.7298630>

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: <http://hdl.handle.net/1721.1/105223>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike



Small-Variance Nonparametric Clustering on the Hypersphere

Julian Straub Trevor Campbell Jonathan P. How John W. Fisher III

CSAIL and LIDS, Massachusetts Institute of Technology

{jstraub, fisher}@csail.mit.edu, {tdjc, jhow}@mit.edu

Abstract

Structural regularities in man-made environments reflect in the distribution of their surface normals. Describing these surface normal distributions is important in many computer vision applications, such as scene understanding, plane segmentation, and regularization of 3D reconstructions. Based on the small-variance limit of Bayesian nonparametric von-Mises-Fisher (vMF) mixture distributions, we propose two new flexible and efficient k -means-like clustering algorithms for directional data such as surface normals. The first, DP-vMF-means, is a batch clustering algorithm derived from the Dirichlet process (DP) vMF mixture. Recognizing the sequential nature of data collection in many applications, we extend this algorithm to DDP-vMF-means, which infers temporally evolving cluster structure from streaming data. Both algorithms naturally respect the geometry of directional data, which lies on the unit sphere. We demonstrate their performance on synthetic directional data and real 3D surface normals from RGB-D sensors. While our experiments focus on 3D data, both algorithms generalize to high dimensional directional data such as protein backbone configurations and semantic word vectors.

1. Introduction

Man-made environments and objects exhibit clear structural regularities such as planar or rounded surfaces. These properties are evident on all scales from small objects such as books, to medium-sized scenes like tables, rooms and buildings and even to the organization of whole cities. Such regularities can be captured in the statistics of surface normals that describe the local differential structure of a shape. These statistics contain valuable information that can be used for scene understanding, plane segmentation, or to regularize a 3D reconstruction.

Inference algorithms in fields such as robotics or augmented reality, which would benefit from the use of surface normal statistics, are not generally provided a single batch of data a priori. Instead, they are often provided a stream of data batches from depth cameras. Thus, capturing the sur-

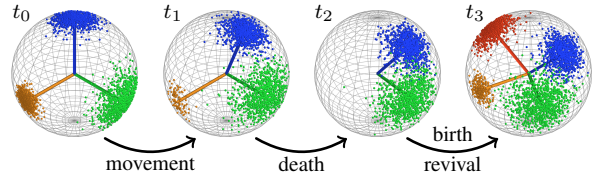


Figure 1: Evolution of the data distribution on the sphere which we model using a dependent Dirichlet process von-Mises-Fisher mixture model (DDP-vMF-MM).

face normal statistics of man-made structures often necessitates the temporal integration of observations from a vast data stream of varying cluster mixtures. Additionally, such applications pose hard constraints on the amount of computational power available, as well as tight timing constraints.

We address these challenges by focusing on flexible Bayesian nonparametric (BNP) Dirichlet process mixture models (DP-MM) which describe the distribution of surface normals in their natural space, the unit sphere in 3D. Taking the small-variance asymptotic limit of this DP-MM of von-Mises-Fisher (vMF) distributions, we obtain a fast k -means-like algorithm, which we call DP-vMF-means, to perform nonparametric clustering of data on the unit hypersphere. Furthermore, we propose a novel dependent DP mixture of vMF distributions to achieve integration of directional data into a temporally consistent streaming model (shown in Fig. 1). Small-variance asymptotic analysis yields the k -means-like DDP-vMF-means algorithm. Finally, we propose a method, inspired by optimistic concurrency control, for parallelizing the inherently sequential labeling process of BNP-derived algorithms. This allows real-time processing of batches of 300k data-points at 30 Hz.

Beyond the aforementioned vision applications, directional data is ubiquitous in many other fields, including protein backbone configurations in computational biology, semantic word vectors in natural language processing, and rotations expressed as quaternions in robotics. Further, many of those sources are observed as a stream. The proposed algorithms directly generalize to these other data sources as well as higher-dimensional spherical data.

2. Related Work

Directional distributions: A variety of distributions [30] has been proposed in the field of directional statistics to model data on the unit sphere. Examples are the antipodal symmetric Bingham distribution [5], the anisotropic Kent distribution [25], and the isotropic von-Mises-Fisher (vMF) distribution [14]. Because of its comparative simplicity, the vMF distribution is most commonly used.

vMF mixture models: vMF mixture models (vMF-MM) are especially popular for modeling and inference purposes. Banerjee *et al.* [3] perform Expectation Maximization (EM) for a finite vMF mixture model to cluster text and genomic data. This method is related to the spherical k -means (spkm) algorithm [11], which can be obtained from a finite vMF-MM by taking the infinite limit of the concentration parameter [3]. Zhong [43] extends the spherical k -means algorithm to an online clustering framework by performing stochastic gradient descent on the spkm objective. This approach requires the number of clusters to be known, does not allow the creation or deletion of clusters, and heuristically weights the contribution of old data. Gopal *et al.* [16] derive variational as well as collapsed Gibbs sampling inference for finite vMF-MMs, for a finite hierarchical vMF-MM and for a finite temporally evolving vMF-MM. The vMF distribution has also been used in BNP and hierarchical MM. Bangert *et al.* [4] formulate an infinite vMF-MM using a Dirichlet process (DP) prior, but their sampling-based inference is known to have convergence issues and to be inefficient [23]. Reisinger *et al.* [34] formulate a finite, latent Dirichlet allocation (LDA) model [7] for directional data using the vMF distribution. To the best of our knowledge there are no other related k -means-like algorithms for directional batch data besides the spkm algorithm. In many applications the spherical manifold of the data is ignored and clustering is performed in the ambient Euclidean space [19].

Surface normal modeling: In the context of object and 3D shape representation, extended Gaussian images (EGI) have been studied [20, 21]. The EGI of a surface is the distribution of surface normals where each surface normal is weighted by the area of the surface it represents. For computations the EGI was approximated via a tessellation of the sphere. A crucial property of EGI is that the representation is invariant to translation. Exploiting this property, Makadia *et al.* [29] extract maxima in the EGI and use the spherical FFT to compute an initial rotation estimate for point-cloud registration. Note that these maxima correspond to the cluster centers which the proposed (D)DP-vMF-means algorithm extracts. Furukawa *et al.* [15] use a tessellation of the unit sphere to extract the dominant directions in a scene as part of a depth regularization algorithm based on the Manhattan World (MW) assumption [10]. Tribel *et al.* [42] employ EM to perform plane segmentation using

surface normals in combination with 3D locations. Assuming a finite vMF-MM and using a hierarchical clustering approach, Hasnat *et al.* [17] model surface normal distributions. Straub *et al.* [39] propose a probabilistic model to describe mixtures of Manhattan frames (MMF) as orthogonally coupled clusters of surface normals on the sphere. In a similar approach, but without MMF constraints, Straub *et al.* [38] introduce efficient inference for a DP-MM of clusters in distinct tangent spaces to the sphere.

Small-variance asymptotics: A principled way to reduce the computational cost of inference, while still drawing on the capabilities of BNP models, is to apply small-variance analysis to the model. This technique was first applied to the DP Gaussian mixture [26], resulting in a k -means-like algorithm with a linear regularization on the number of clusters. Since then, the technique has been extended to develop asymptotic algorithms for mixtures with general exponential family likelihoods [24], HMMs with an unknown number of states [37], and dynamic, time-dependent mixtures [8]. However, small-variance analysis has, to date, been limited to discrete and Euclidean spaces.

3. Von-Mises-Fisher Mixture Models

The von-Mises-Fisher (vMF) [14] distribution with mean direction μ and concentration τ is an isotropic distribution over the D -dimensional unit hypersphere, $\mathbb{S}^{D-1} = \{x \in \mathbb{R}^D : x^T x = 1\}$, (Fig. 2). Its density is defined as [4]

$$\text{vMF}(x; \mu, \tau) = Z(\tau) \exp(\tau \mu^T x) \\ Z(\tau) = \tau^{D/2-1} (2\pi)^{-D/2} \mathbf{I}_{D/2-1}(\tau)^{-1}, \quad (1)$$

where $\mu, x \in \mathbb{S}^{D-1}$, $\tau > 0$ and \mathbf{I}_ν is the modified Bessel function [1] of the first kind of order ν . The conjugate prior for μ given a fixed τ is a vMF distribution $\text{vMF}(\mu; \mu_0, \tau_0)$ and the corresponding posterior given data $\mathbf{x} = \{x_i\}_{i=1}^N$ is

$$p(\mu | \mathbf{x}; \tau, \mu_0, \tau_0) \propto p(\mu; \mu_0, \tau_0) \prod_{i=1}^N p(x_i | \mu; \tau) \quad (2)$$

$$\therefore p(\mu | \mathbf{x}; \tau, \mu_0, \tau_0) = \text{vMF}(\mu; \frac{\mu'}{\|\mu'\|_2}, \|\mu'\|_2),$$

where $\mu' = \tau_0 \mu_0 + \tau \sum_{i=1}^N x_i$.

A finite mixture of K vMF distributions with known concentration τ may be obtained by placing a Dirichlet distribution prior $\text{Dir}(\alpha)$ on the mixture weights π , and a vMF prior on the mean directions μ_k . Data points \mathbf{x} are assigned to clusters via latent indicator variables $\mathbf{z} = \{z_i\}_{i=1}^N$.

$$\pi \sim \text{Dir}(\alpha), \quad \mu_k \sim \text{vMF}(\mu_0, \tau_0) \quad \forall k \in \{1, \dots, K\} \\ z_i \sim \text{Cat}(\pi), \quad x_i \sim \text{vMF}(\mu_{z_i}, \tau) \quad \forall i \in \{1, \dots, N\}. \quad (3)$$

Parallel to the connection between k -means and the Gaussian mixture model in the small-variance asymptotic

limit [26], taking $\tau \rightarrow \infty$ yields deterministic updates as also previously noted in [3]. For completeness of the presentation we give a detailed derivation in the supplement.

4. Dirichlet Process vMF-MM

The Dirichlet process (DP) [13, 41] has been widely used as a prior for mixture models with a countably-infinite set of clusters [2, 4, 9, 32]. Assuming a base distribution $\text{vMF}(\mu; \mu_0, \tau_0)$, the DP is an appropriate prior for a vMF mixture with an unknown number of components and known vMF concentration τ . Gibbs sampling inference only differs from the finite Dirichlet vMF-MM in the label sampling step; the mixing weights π are integrated out, resulting in the Chinese Restaurant Process (CRP) [6, 32]

$$p(z_i = k | \mathbf{z}_{-i}, \boldsymbol{\mu}, \mathbf{x}; \tau, \mu_0, \tau_0) \propto \begin{cases} |\mathcal{I}_k| \text{vMF}(x_i | \mu_k; \tau) & k \leq K \\ \alpha p(x_i; \tau, \mu_0, \tau_0) & k = K + 1, \end{cases} \quad (4)$$

where \mathcal{I}_k is the set of data indices assigned to cluster k . Note that the DP concentration parameter, $\alpha > 0$, influences the likelihood of adding a new clusters. The conjugate prior for μ_k yields $p(x_i; \tau, \mu_0, \tau_0)$ via marginalization:

$$p(x_i; \tau, \mu_0, \tau_0) = \int \text{vMF}(x_i | \mu_k; \tau) \text{vMF}(\mu_k; \mu_0, \tau_0) d\mu_k = \frac{Z(\tau)Z(\tau_0)}{Z(\|\tau x_i + \tau_0 \mu_0\|_2)}. \quad (5)$$

4.1. DP-vMF-means

In this section, we provide a small-variance asymptotic analysis of the label and parameter update steps of the Gibbs sampling algorithm for the DP vMF mixture, yielding deterministic updates. As with k -means, the label assignments are computed sequentially for all datapoints before the means are updated, and the process is iterated until convergence. Pseudocode can be found in the supplement.

Label Update: To derive a hyperspherical analog to DP-means [26], consider the limit of the label sampling step (4) as $\tau \rightarrow \infty$. The normalizer $Z(\|\tau x_i + \tau_0 \mu_0\|_2)$ approaches

$$Z(\|\tau x_i + \tau_0 \mu_0\|_2) \xrightarrow{\tau^{(*)}} \exp(-\tau), \quad (6)$$

where overscript $\tau^{(*)}$ denotes proportionality up to a finite power of τ , and where we have used the fact that as $\tau \rightarrow \infty$, the modified Bessel function of the first kind satisfies [1]

$$I_{D/2-1}(\tau) = \frac{\exp(\tau)}{\sqrt{2\pi\tau}} \left(1 - O\left(\frac{1}{\tau}\right)\right) \xrightarrow{\tau^{(*)}} \exp(\tau). \quad (7)$$

To achieve a nontrivial result, the asymptotic behavior of $Z(\tau)$ must be matched by α , so let $\alpha = \exp(\lambda\tau)$ to obtain

$$\frac{\alpha Z(\tau)Z(\tau_0)}{Z(\|\tau x_i + \tau_0 \mu_0\|_2)} \xrightarrow{\tau^{(*)}} Z(\tau) \exp(\tau(\lambda + 1)). \quad (8)$$

Therefore, as $\tau \rightarrow \infty$, the label sampling step becomes

$$\lim_{\tau \rightarrow \infty} p(z_i = k | \mathbf{z}_{-i}, \boldsymbol{\mu}, \mathbf{x}; \tau, \mu_0, \tau_0) = \lim_{\tau \rightarrow \infty} \begin{cases} \frac{|\mathcal{I}_k| e^{\tau(x_i^T \mu_k - \lambda - 1)}}{\sum_{j=1}^K |\mathcal{I}_j| e^{\tau(x_i^T \mu_j - \lambda - 1) + c(\tau)}} & k \leq K \\ \frac{c(\tau)}{\sum_{j=1}^K |\mathcal{I}_j| e^{\tau(x_i^T \mu_j - \lambda - 1) + c(\tau)}} & k = K + 1, \end{cases} \quad (9)$$

where we have used that the normalizers $Z(\tau)$ of $\text{vMF}(x_i | \mu_k; \tau)$ and Eq. (8) cancel and $c(\tau) \xrightarrow{\tau^{(*)}} 1$. Thus, as $\tau \rightarrow \infty$, sampling from $p(z_i | \mathbf{z}_{-i}, \boldsymbol{\mu}, \mathbf{x}; \tau, \mu_0, \tau_0)$ is equivalent to the following assignment rule:

$$z_i = \arg \max_{k \in \{1, \dots, K+1\}} \begin{cases} x_i^T \mu_k & k \leq K \\ \lambda + 1 & k = K + 1. \end{cases} \quad (10)$$

Since $-1 \leq x_i^T \mu_k \leq 1$, the parameter λ can be restricted to the set $\lambda \in [-2, 0]$ without loss of generality. Intuitively λ defines the maximum angular spread ϕ_λ of clusters about their mean direction, via $\lambda = \cos(\phi_\lambda) - 1$. Note, that upon assigning a datapoint to a new cluster, i.e. $z_i = K + 1$, the mean of that cluster is initialized to $\mu_{K+1} = x_i$. Finally, if an observation x_i is the last one in its cluster, the cluster is removed prior to finding the new label for x_i using Eq. (10).

Parameter Update: Taking $\tau \rightarrow \infty$ in the parameter posterior for cluster k from Eq. (2) causes τ_0 and μ_0 to become negligible. Hence the parameter update becomes:

$$\mu_k = \frac{\sum_{i \in \mathcal{I}_k} x_i}{\|\sum_{i \in \mathcal{I}_k} x_i\|_2} \quad \forall k \in \{1, \dots, K\}. \quad (11)$$

Objective Function: From Eq. (10) we can see that assigning a datapoint x_i to cluster k provides a score of $x_i^T \mu_k$, whereas adding a new cluster provides a score of $\lambda + 1 - x_i^T \mu_{K+1} = \lambda$, since new mean directions are initialized directly to $\mu_{K+1} = x_i$. Hence, the objective function that DP-vMF-means maximizes is

$$J_{\text{DP-vMF}} = \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} x_i^T \mu_k + \lambda K. \quad (12)$$

5. Dependent Dirichlet Process vMF-MM

Suppose now that, in addition to an unknown number K of components, the vMF mixture undergoes temporal evolution in discrete timesteps $t \in \mathbb{N}$ (Fig. 1): mixture components can move, be destroyed, and new ones can be created at each timestep. For such a scenario, the dependent Dirichlet process (DDP) [28, 27, 8] is an appropriate prior over the mixture components and weights. Using intermediate auxiliary DPs F_0 and F_1 , the DDP constructs a Markov chain of DPs G_t , where G_{t+1} is sampled from G_t as follows:

1. **(Death)** For each atom θ in G_t , sample from Bernoulli(q). If the result is 1, add θ to F_0 .

2. **(Motion)** Replace each θ in F_0 with $\theta' \sim T(\theta'|\theta)$.
3. **(Birth)** Sample a DP $F_1 \sim \text{DP}(\alpha, H)$. Let G_{t+1} be a random convex combination of F_0 and F_1 .

There are four parameters in this model: $\alpha > 0$ and $H(\cdot)$, the concentration parameter and base measure of the innovation process; $q \in (0, 1)$, the Bernoulli cluster survival probability; and finally $T(\cdot|\cdot)$, the random walk transition distribution. In the present work, both the base and random transition distributions are von-Mises-Fisher: $H(\mu) = \text{vMF}(\mu; \mu_0, \tau_0)$, and $T(\mu|\nu) = \text{vMF}(\mu; \nu, \xi)$.

Suppose at timestep t , a new batch of data \mathbf{x} is observed. Then Gibbs sampling posterior inference for the DDP mixture, as in the previous sections, iteratively samples labels and parameters. Let the set of tracked mean directions from previous timesteps be $\{\mu_{k0}\}_{k=1}^K$, where $\mu_{k0} \sim \text{vMF}(m_k, \tau_k)$ and Δt_k denotes the number of timesteps since cluster k was last instantiated (i.e. when it last had data assigned to it).¹ Then the label sampling distribution is

$$p(z_i = k | \mu, \mathbf{z}_{-i}, \mathbf{x}) \propto \begin{cases} \alpha^{\frac{1-q^t}{1-q}} p(x_i; \mu_0, \tau_0) & k = K+1 \\ (c_k + |\mathcal{I}_k|) \text{vMF}(x_i | \mu_k; \tau) & k \leq K, |\mathcal{I}_k| > 0 \\ q^{\Delta t_k} c_k p(x_i; m_k, \tau_k) & k \leq K, |\mathcal{I}_k| = 0, \end{cases} \quad (13)$$

where c_k is the number of observations assigned to cluster k in past timesteps. The parameter sampling distribution is

$$p(\mu_k | \mu_{-k}, \mathbf{z}, \mathbf{x}) \propto \begin{cases} \text{vMF}(\mu_k; \frac{\mu'_k}{\|\mu'_k\|_2}, \|\mu'_k\|_2) & c_k = 0 \\ p(\mu_k | \mathbf{x}, \mathbf{z}; m_k, \tau_k) & c_k > 0, \end{cases} \quad (14)$$

where $\mu'_k = \tau_0 \mu_0 + \tau \sum_{i \in \mathcal{I}_k} x_i$, and $p(\mu_k | \mathbf{x}, \mathbf{z}; m_k, \tau_k)$ is the distribution over the current cluster k mean direction μ_k given the assigned data and the old mean direction μ_{k0} .

5.1. DDP-vMF-means

In the following, we analyze the small-variance asymptotics of the DDP-vMF mixture model. We first derive the label assignment rules, followed by the parameter updates.

Label Update: First, let $\alpha = \exp(\lambda\tau)$, $q = \exp(Q\tau)$, $\xi = \exp(\beta\tau)$, and $\tau_k = \tau w_k$, with $\lambda \in [-2, 0]$ as before, $Q \leq 0$, and $\beta, w_k \geq 0$. Note that $\lim_{\tau \rightarrow \infty} \frac{1-q^t}{1-q} = 1$, and thus the asymptotics of the label assignment probability for current and new clusters is the same as in Section 4.1.

Hence, we focus on the assignment of a datapoint to a previously observed, but currently not instantiated, cluster k . During the Δt_k timesteps since cluster k was last observed, the mean direction μ_k underwent a random vMF

walk $\mu_{k0} \rightarrow \mu_{k1} \rightarrow \dots \rightarrow \mu_{k\Delta t_k} = \mu_k$ with initial distribution $\mu_{k0} \sim \text{vMF}(\mu_{k0}; m_k, \tau_k)$. Therefore, the intermediate mean directions $\{\mu_{kn}\}_{n=1}^{\Delta t_k}$ must be marginalized out when computing $p(x_i; m_k, \tau_k)$:

$$\begin{aligned} p(x_i; m_k, \tau_k) &= \int \dots \int_{\mu_{k0}, \dots, \mu_{k\Delta t_k}} p(x_i | \mu_{k\Delta t_k}; \tau) \cdot p(\mu_{k0}; m_k, \tau_k) \\ &\quad \cdot \prod_{n=1}^{\Delta t_k} p(\mu_{kn} | \mu_{k(n-1)}; \xi) \\ &= Z(\tau) Z(\beta\tau)^{\Delta t_k} Z(w_k\tau) \int \dots \int_{\mu_{k0}, \dots, \mu_{k\Delta t_k}} \exp(\tau f) \quad (15) \\ f &= x_i^T \mu_{k\Delta t_k} + \beta \sum_{n=1}^{\Delta t_k} \mu_{kn}^T \mu_{k(n-1)} + w_k \mu_{k0}^T m_k. \end{aligned}$$

The integration in (15) cannot be computed in closed form; however, the value of the integral is only of interest in the limit as $\tau \rightarrow \infty$. Therefore, Theorem 1, an extension of Laplace's approximation to general differentiable manifolds, may be used to obtain an exact formula.

Theorem 1 (Manifold Laplace Approximation). *Suppose $M \subset \mathbb{R}^n$ is a bounded m -dimensional differentiable manifold and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a smooth function on M . Further, suppose f has a unique global maximum on M , $x^* = \arg \max_{x \in M} f(x)$. Then*

$$\lim_{\tau \rightarrow \infty} \frac{\int_M e^{\tau f(x)} dx}{\left(\frac{2\pi}{\tau}\right)^m |\det U^T \nabla^2 f(x^*) U|^{\frac{1}{2}} e^{\tau f(x^*)}} = 1, \quad (16)$$

where $U \in \mathbb{R}^{n \times m}$ is a matrix whose columns are an orthonormal basis for the tangent space of M at x^* .

Proof. See the supplementary material. The general technique of this proof is to transform coordinates between the manifold and its tangent plane using the exponential map [12], and then apply the multidimensional Laplace approximation in the transformed Euclidean space. \square

Corollary 1. *Given a smooth function $f : (\mathbb{R}^D)^N \rightarrow \mathbb{R}$, with a unique global maximum over the N -product of $(D-1)$ -spheres $x^* \in (\mathbb{S}^{D-1})^N$,*

$$\int_{(\mathbb{S}^{D-1})^N} e^{\tau f(x)} \xrightarrow{\tau \rightarrow \infty} e^{\tau f(x^*)}. \quad (17)$$

Proof. This is Theorem 1 applied to $(\mathbb{S}^{D-1})^N$. \square

Using Corollary 1 and the limiting approximation of the modified Bessel function (7) in equation (15) yields the following asymptotic behavior as $\tau \rightarrow \infty$:

$$p(x_i; m_k, \tau_k) \xrightarrow{\tau \rightarrow \infty} \exp(\tau(f^* - 1 - \beta \Delta t_k - w_k)). \quad (18)$$

¹Note that all quantities (\mathbf{x} , μ_{k0} , m_k , τ_k , c_k , n_k , etc.) are now time-varying. This dependence is not shown in the notation for brevity, and all quantities are assumed to be shown for the current timestep t .

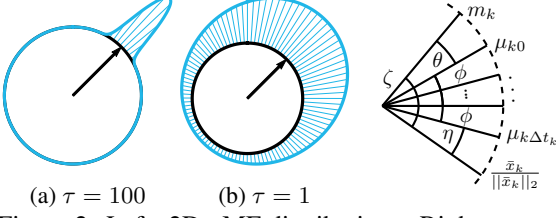


Figure 2: Left: 2D vMF distributions. Right: geometry of the maximum likelihood setting of $\mu_{k0}, \mu_{k1}, \dots, \mu_{k\Delta t_k}$ for the transition distribution.

The only remaining unknown in the asymptotic expression, f^* , can be found via constrained optimization

$$\begin{aligned} \max_{\{\mu_{kn}\}_{n=1}^{\Delta t_k}} & x_i^T \mu_{k\Delta t_k} + \beta \sum_{n=1}^{\Delta t_k} \mu_{kn}^T \mu_{k(n-1)} + w_k \mu_{k0}^T m_k \\ \text{s.t. } & \mu_{kn}^T \mu_{kn} = 1 \quad \forall n \in \{0, \dots, \Delta t_k\}. \end{aligned} \quad (19)$$

The optimization (19) has a closed-form solution:

$$\begin{aligned} \mu_{k0} &= \frac{w_k m_k + \beta \mu_{k1}}{\|w_k m_k + \beta \mu_{k1}\|_2} \\ \mu_{kn} &= \frac{\mu_{k(n+1)} + \mu_{k(n-1)}}{\|\mu_{k(n+1)} + \mu_{k(n-1)}\|_2} \quad \forall n \in \{1, \dots, \Delta t_k - 1\} \\ \mu_{k\Delta t_k} &= \frac{x_i + \beta \mu_{k(\Delta t_k - 1)}}{\|x_i + \beta \mu_{k(\Delta t_k - 1)}\|_2}. \end{aligned} \quad (20)$$

These ternary relationships enforce that the optimal vMF mean directions along the random walk lie on the geodesic between m_k and x_i . Therefore, this walk can be described geometrically by three angles, as shown in Fig. 2 (with $\bar{x}_k = x_i$): the angle ϕ between consecutive μ_{kn} , the angle η between x_i and $\mu_{k\Delta t_k}$, and the angle θ between m_k and μ_{k0} . Given these definitions, standard trigonometry yields a set of three equations in ϕ , η , and θ :

$$\begin{aligned} w_k \sin(\theta^*) &= \beta \sin(\phi^*) = \sin(\eta^*) \\ \zeta &= \theta^* + \Delta t_k \phi^* + \eta^* = \arccos(m_k^T x_i), \end{aligned} \quad (21)$$

where ζ is the full angle between x_i and m_k . Since (21) cannot be solved in closed-form, Newton's method is used to compute ϕ^* , θ^* , and η^* , which in turn determines f^* :

$$f^* = w_k \cos(\theta^*) + \beta \Delta t_k \cos(\phi^*) + \cos(\eta^*). \quad (22)$$

Returning to (18), the transition asymptotics are

$$p(x_i; m_k; \tau_k) \xrightarrow{\tau^*} \exp \left(\frac{\tau w_k (\cos(\theta^*) - 1) + \tau \beta \Delta t_k (\cos(\phi^*) - 1) + \tau (\cos(\eta^*) - 1)}{\tau} \right). \quad (23)$$

Substituting this into Eq. (13) with the earlier definition $q = \exp(\tau Q)$, and taking the limit $\tau \rightarrow \infty$ yields the assignment

rule $z_i = \arg \max_k J_k$, where

$$J_k = \begin{cases} \lambda + 1 & k = K + 1 \\ \mu_k^T x_i & k \leq K, |\mathcal{I}_k| > 0 \\ \begin{pmatrix} \Delta t_k \beta (\cos(\phi^*) - 1) \\ + w_k (\cos(\theta^*) - 1) \\ + \cos(\eta^*) + \Delta t_k Q \end{pmatrix} & k \leq K, |\mathcal{I}_k| = 0. \end{cases} \quad (24)$$

Note that if $Q \Delta t_k < \lambda$, cluster k can be removed permanently as it will never be revived again. Furthermore, for any $Q \leq \lambda$ all clusters are removed after each timestep, and the algorithm reduces to DP-vMF-means.

Parameter Update: The parameter update rule for DDP-vMF-means comes from the asymptotic behavior of (14) as $\tau \rightarrow \infty$. The analysis for any new cluster is the same as that in Section 4.1, so our focus is again on the transitioned mean direction posterior $p(\mu_{k\Delta t_k} | \mathbf{x}, \mathbf{z}; m_k, \tau_k)$ (recall that $\mu_k = \mu_{k\Delta t_k}$ in the definition of the random vMF walk). This distribution can be expanded, similarly to (15), as:

$$\begin{aligned} p(\mu_{k\Delta t_k} | \mathbf{x}, \mathbf{z}; m_k, \tau_k) &= \int \dots \int_{\mu_{k0}, \dots, \mu_{k(\Delta t_k - 1)}} p(\mathbf{x} | \mu_{k\Delta t_k}; \tau) \cdot p(\mu_{k0}; m_k, \tau_k) \\ &\quad \cdot \prod_{n=1}^{\Delta t_k} p(\mu_{kn} | \mu_{k(n-1)}; \xi) \\ &= Z(\tau)^{|\mathcal{I}_k|} Z(\beta \tau)^{\Delta t_k} Z(w_k \tau) \int \dots \int_{\mu_{k0}, \dots, \mu_{k(\Delta t_k - 1)}} \exp(\tau f) \end{aligned} \quad (25)$$

$$f = \sum_{i \in \mathcal{I}_k} x_i^T \mu_{k\Delta t_k} + \beta \sum_{n=1}^{\Delta t_k} \mu_{kn}^T \mu_{k(n-1)} + w_k \mu_{k0}^T m_k.$$

Define $\bar{x}_k = \sum_{i \in \mathcal{I}_k} x_i$. Once again, applying Corollary 1, the limit $\tau \rightarrow \infty$ removes the integrals over the marginalized mean directions. However, in contrast to the label assignment update, $\mu_{k\Delta t_k}$ is not marginalized out. Therefore, an additional maximization with respect to $\mu_{k\Delta t_k}$ to find the concentration point of the posterior yields

$$\begin{aligned} p(\mu_{k\Delta t_k} | \mathbf{x}, \mathbf{z}; m_k, \tau_k) &\xrightarrow{\tau^*} \exp(\tau(f^* - |\mathcal{I}_k| - \Delta t_k \beta - w_k)) \\ f^* &= w_k \cos(\theta^*) + \beta \Delta t_k \cos(\phi^*) + \|\bar{x}_k\|_2 \cos(\eta^*). \end{aligned} \quad (26)$$

Analyzing the geometry of the geodesic between $\bar{x}_k / \|\bar{x}_k\|_2$ and m_k (Fig. 2) there exist ϕ^* , θ^* and η^* such that

$$\begin{aligned} w_k \sin(\theta^*) &= \beta \sin(\phi^*) = \|\bar{x}_k\|_2 \sin(\eta^*) \\ \zeta &= \theta^* + \Delta t_k \phi^* + \eta^* = \arccos(m_k^T \frac{\bar{x}_k}{\|\bar{x}_k\|_2}), \end{aligned} \quad (27)$$

which can be solved via Newton's method. Given the solution, μ_k can be obtained by rotating $\frac{\bar{x}_k}{\|\bar{x}_k\|_2}$ by angle η^* on the geodesic shown in Fig. 2 towards m_k ,

$$\mu_k = R(\eta^*) \frac{\bar{x}_k}{\|\bar{x}_k\|_2}. \quad (28)$$

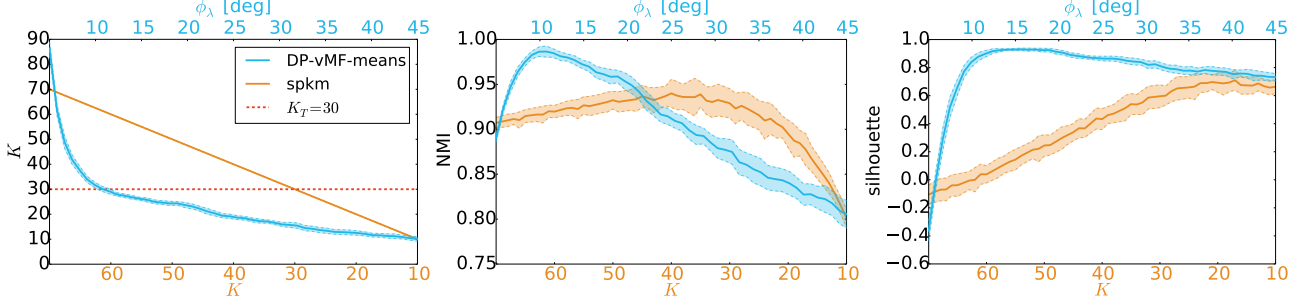


Figure 3: Comparison of the spkm and the DP-vMF-means clustering algorithms on synthetic spherical data with $K_T = 30$ clusters. Note DP-vMF-means’ higher maximum normalized mutual information (NMI) as well as silhouette score.

Weight Update: After the iteration of label and parameter updates has converged, the weight w_k must be updated for all clusters to reflect the new uncertainty in the mean direction of cluster k . This can be done by examining (26): Since at the maximum of a $\text{vMF}(\mu; m_k, w_k \tau)$ density, $\exp(\tau w_k m_k^T \mu) = \exp(\tau w_k)$, w_k is updated to f^* .

6. Optimistic Iterated Restarts (OIR)

In our implementation of the algorithm we pay special attention to speed and parallel execution to enable real-time performance for streaming RGB-D data.

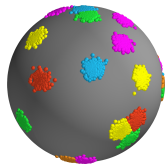
Observe that the main bottleneck of DP-based hard clustering algorithms, such as the proposed (D)DP-vMF-means, DP-means [26] or Dynamic means [8], is the inherently sequential assignment of labels: due to the creation of new clusters, the label assignments depend on all previous assignments. While this is a key feature of the streaming clustering algorithms, it poses a computational hindrance. We address this issue with an optimistic parallel label assignment procedure inspired by techniques for database concurrency control [33].

First, we compute assignments in parallel (e.g. on a GPU). If all datapoints were assigned only to instantiated clusters, we output the labeling. Otherwise, we find the lowest observation id i that modified the number of clusters, apply the modification, and recompute the assignments for all observations $i' > i$ in parallel. Thus, per data-batch, DP-vMF-means restarts once per new cluster, while DDP-vMF-means restarts once for each new or revived cluster.

7. Results

7.1. Evaluation of the DP-vMF-means Algorithm

Synthetic Data: First, we evaluate the behavior of the DP-vMF-means algorithm in comparison to its parametric cousin, the spkm algorithm, on synthetic 3D spherical data sampled from $K_T = 30$ true vMF distributions. All evaluation results are shown as the mean and standard de-



Synthetic Data

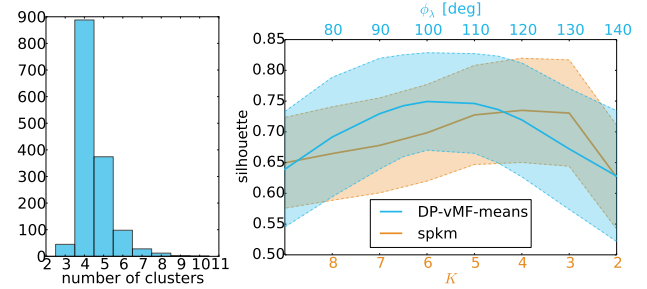


Figure 4: Histogram over the number of clusters found by DP-vMF-means (left) and silhouette values for DP-vMF-means and spkm (right) across the whole NYU dataset [31].

viation over 50 runs. The left plot of Fig. 3 depicts the inferred number of clusters K on the horizontal axis as a function of the respective parameters of the two algorithms: the number of clusters K for spkm and the parameter ϕ_λ for DP-vMF-means (recall that $\phi_\lambda = \cos^{-1}(\lambda + 1)$ as defined in Sec. 4.1). This figure demonstrates the ability of the DP-vMF-means algorithm to discover the correct number of clusters K_T , and the relative insensitivity of the discovered number of clusters with respect to its parameter ϕ_λ .

The middle and right hand plots show two measures for clustering quality. The normalized Mutual Information (NMI) [40], depicted in the middle, is computed using the true labels. DP-vMF-means achieves an almost perfect NMI of 0.99, while spkm only reaches 0.94 NMI even with $K = K_T$. The slightly superior performance of DP-vMF-means stems from its enhanced ability to avoid local optima due to the way labels are initialized: while spkm is forced to initialize K cluster parameters, DP-vMF-means starts with an empty set and adds clusters on the fly as more data are labelled. The NMI results are corroborated by the silhouette score [36], shown to the right in Fig. 3. The silhouette score is an internal measure for clustering quality that can be computed without knowledge of the true clustering, and is used to tune parametric clustering algorithms. With a maximum of 0.92 DP-vMF-means reaches a close to perfect silhouette score, indicating well-separated, concentrated clusters. Again, spkm does not reach the same clus-

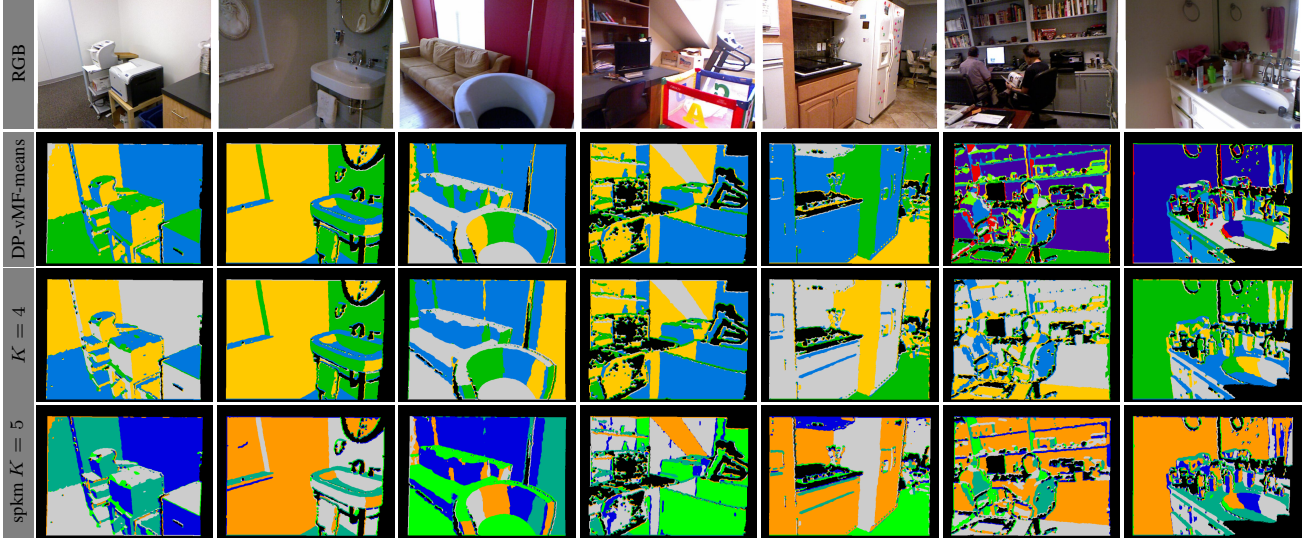


Figure 5: Directional segmentation of scenes from the NYU v2 RGB-D dataset [31] as implied by surface normal clusters. The complexity of the scenes increases from left to right as can be seen from the RGB images. The second row shows the clustering inferred by DP-vMF-Means while the third and fourth show the spherical k -means results. Black denotes missing data due to sensor limitations. Note that DP-vMF-means adapts the number of clusters to the complexity of the scene.

tering performance even for $K = K_T$ for the same aforementioned reasons.

For additional reference we ran two sampling-based inference algorithms for the DP-vMF-MM model. The CRP-based inference of [4] was aborted after running for two days without convergence. This inefficiency has been noted previously [23]. A more efficient alternative is the finite Dirichlet process (FSD) approximation [22] to the DP-vMF-MM. The inference can be parallelized and yielded results within minutes. After 5000 iterations the sampler converged to an incorrect number of 23.0 ± 1.67 clusters on average over 10 sampler runs. These convergence issues have also been noted before [9]. The last sample of the different runs achieves an average NMI of 0.94 ± 0.01 and a silhouette score of 0.74 ± 0.04 .

NYU v2 depth dataset: In this experiment, the DP-vMF-means and spkm algorithms were compared on the NYU v2 RGB-D dataset [31]. Surface normals were extracted from the depth images [19] and preprocessed with total variation smoothing [35]. We quantify the clustering quality in terms of the average silhouette score over the clusterings of the 1449 scenes of the NYU v2 depth dataset. Since we do not possess the true scene labeling, we use the silhouette quality metric as a proxy for the NMI metric; this was motivated by results of the synthetic experiment.

Across the whole NYU v2 dataset, the DP-vMF-means algorithm achieves the highest average silhouette score of 0.75 for $\phi_\lambda^* = 100^\circ$ as depicted in Fig. 4. The histogram over the number of inferred clusters by DP-vMF-means for ϕ_λ^* indicates the varying complexity of the scenes ranging

from three to eleven. The clear peak at $K = 4$ coincides with the highest silhouette score for spkm (0.73) and explains the only slightly lower silhouette score of spkm: most scenes in the dataset exhibit four primary directions.

Figure 5 shows a qualitative comparison of the scene segmentation implied by the clustering of surface normals. In comparison to spkm, the DP-vMF-means clustering results show the ability of the algorithm to adapt the number of clusters to the scene at hand. If the right number of clusters is selected for the spkm clustering, the results have similar quality; however, the number of clusters is generally not known a priori and varies across scenes. This demonstrates two major advantages of DP-vMF-means over spkm: (1) DP-vMF-means is less sensitive to the parameter setting (see Fig. 3, left) and (2) it is easier to choose ϕ_λ than K since it intuitively corresponds to the maximum angular radius of a cluster, which can be gauged from the type of data and its noise characteristics. For this experiment $\phi_\lambda = 100^\circ$ is justified by the typical Manhattan structure [10] of the indoor environment plus 10° to account for sensor noise.

7.2. Evaluation of the DDP-vMF-means Algorithm

Real-time Directional Segmentation: In fields such as mobile robotics or augmented reality, it is uncommon to observe just a single RGB-D frame of a scene; more typically, the sensor will observe a temporal sequence of frames. The following experiment demonstrates the temporally consistent clustering capability of the DDP-vMF-means algorithm on surface normals extracted from a sequence of depth images recorded in an indoor environment. Each frame is pre-

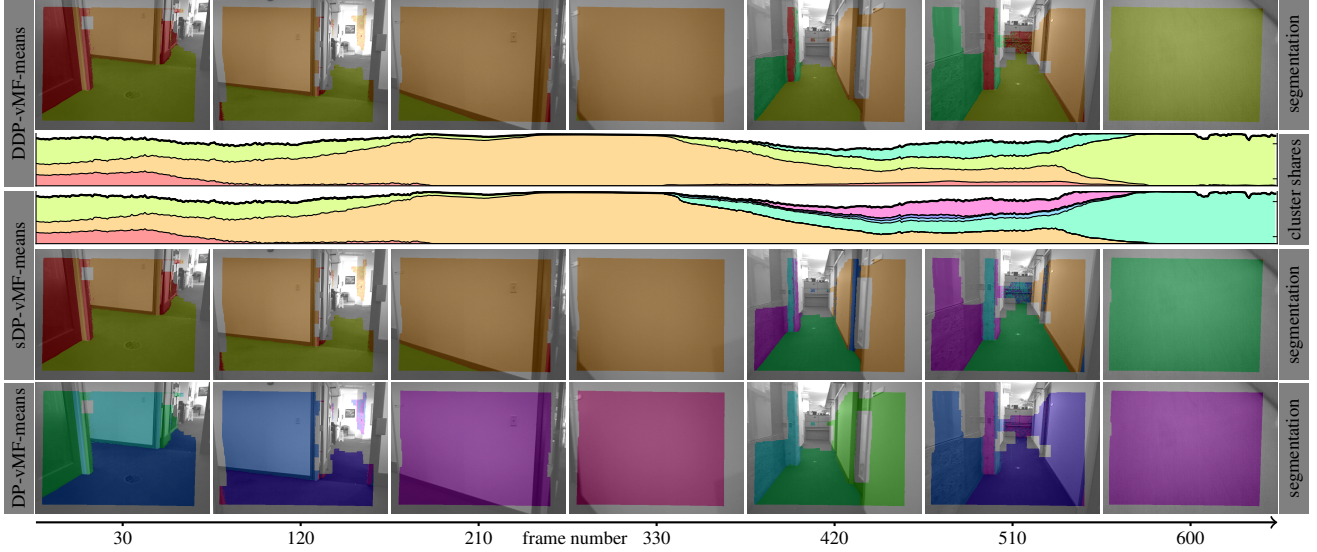


Figure 6: Clustering of a surface normal stream recorded when walking a 90° turn in an office environment. We depict key-frames color-coded with the implied surface-normal clustering for three clustering algorithms. The plots in the second and third row depict the percentage of normals associated to the respective cluster for DDP-vMF-means and sequential DP-vMF-means. Note that only the clustering obtained via the DDP-vMF-means algorithm is consistent across the whole run.

processed in 11ms using edge-preserving smoothing with a hybrid CPU-GPU guided filter [18].

We compare against the ad-hoc approaches of clustering on a frame-by-frame basis using DP-vMF-means, both with and without initializing the algorithm from the previous frame’s clusters. The former is referred to as sequential DP-vMF-means (sDP-vMF-means). sDP-vMF-means achieves a greedy frame-to-frame label consistency, but, unlike DDP-vMF-means, it cannot reinstantiate previous clusters after multiframe lapses. Motivated by the DP-vMF-means evaluation, all algorithms were run with $\phi_\lambda = 100^\circ$. For DDP-vMF-means $\beta = 10^5$ and $Q = \frac{\lambda}{400}$.

The differences in labeling consistency can be observed in rows two and three of Fig. 6, which shows the percentage of normals associated with a specific cluster. While DDP-vMF-means is temporally consistent and reinstantiates the lime-green and red clusters, observed in the first half of the run, DP-vMF-means erroneously creates new clusters. We do not depict the percentages of surface normals associated with the clusters for the batch DP-vMF-means algorithm, since there is no label consistency between time-steps as can be observed in the last row of Fig. 6.

The average run-time per frame was 28.4 ms for batch DP-vMF-means, 12.8 ms for sDP-vMF-means, 20.4 ms for DDP-vMF-means, and 13.6 ms for spkm with $K = 5$. The increased running time of batch DP-vMF-means is a result of clustering each batch of surface normals in isolation; OIR label assignment needs several restarts to assign labels to all surface normals. By initializing the clusters from a previ-

ous frame, sDP-vMF-means only incurs labeling restarts if a new cluster is observed, and hence has significantly lower run time. DDP-vMF-means is slightly slower than sDP-vMF-means since it is keeping track of both observed and unobserved clusters.

8. Conclusion

Taking the small-variance asymptotic limit of the Bayesian nonparametric DP-vMF and DDP-vMF mixture models, we have derived two novel spherical kmeans-like algorithms for efficient batch and streaming clustering on the unit hypersphere. The performance and flexibility of DP-vMF-means was demonstrated on both synthetic data and the NYU v2 RGB-D dataset. For DDP-vMF-means, Optimistic Iterated Restarts (OIR) parallelized label assignments, enable real-time temporally consistent clustering of batches of 300k surface normals collected at 30 Hz from a RGB-D camera.

We envision a large number of potential applications for the presented algorithms in computer vision and in other realms where directional data is encountered. Implementations are available at <http://people.csail.mit.edu/jstraub/>.

Acknowledgements

This work was partially supported by ONR MURI N00014-11-1-0688 and ARO MURI W911NF-11-1-0391.

References

- [1] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions*. Dover Books on Mathematics. Dover Publications, 1965.
- [2] C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 1974.
- [3] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra, and G. Ridgeway. Clustering on the unit hypersphere using von Mises-Fisher distributions. *JMLR*, 6(9), 2005.
- [4] M. Bangert, P. Hennig, and U. Oelfke. Using an infinite von Mises-Fisher mixture model to cluster treatment beam directions in external radiation therapy. In *ICMLA*, 2010.
- [5] C. Bingham. An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, 2(6):1201–1225, 1974.
- [6] D. Blackwell and J. B. MacQueen. Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, 1973.
- [7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.
- [8] T. Campbell, M. Liu, B. Kulis, J. P. How, and L. Carin. Dynamic clustering via asymptotics of the dependent Dirichlet process mixture. In *NIPS*, 2013.
- [9] J. Chang and J. W. Fisher III. Parallel sampling of dp mixture models using sub-clusters splits. In *NIPS*, 2013.
- [10] J. M. Coughlan and A. L. Yuille. Manhattan world: Compass direction from a single image by Bayesian inference. In *ICCV*, 1999.
- [11] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2):143–175, 2001.
- [12] M. P. do Carmo. *Riemannian Geometry*. Birkhäuser Verlag, Boston, MA, 1992.
- [13] T. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1973.
- [14] N. I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, 1995.
- [15] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Manhattan-world stereo. In *CVPR*, 2009.
- [16] S. Gopal and Y. Yang. von Mises-Fisher clustering models. In *ICML*, 2014.
- [17] M. A. Hasnat, O. Alata, and A. Trémeau. Hierarchical 3-d von Mises-Fisher mixture model. In *Workshop on Divergences and Divergence Learning, ICML*, 2013.
- [18] K. He, J. Sun, and X. Tang. Guided image filtering. In *ECCV*, 2010.
- [19] D. Holz, S. Holzer, and R. B. Rusu. Real-Time Plane Segmentation using RGB-D Cameras. In *Proceedings of the RoboCup Symposium*, 2011.
- [20] B. K. P. Horn. Extended Gaussian images. *Proceedings of the IEEE*, 72(12):1671–1686, 1984.
- [21] K. Ikeuchi. Recognition of 3-d objects using the extended Gaussian image. In *In IJCAI Conference*, 1981.
- [22] H. Ishwaran and M. Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.
- [23] S. Jain and R. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13:158–182, 2000.
- [24] K. Jiang, B. Kulis, and M. Jordan. Small-variance asymptotics for exponential family Dirichlet process mixture models. In *NIPS*, 2012.
- [25] J. T. Kent. The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society*, pages 71–80, 1982.
- [26] B. Kulis and M. I. Jordan. Revisiting k-means: New algorithms via Bayesian nonparametrics. In *ICML*, 2012.
- [27] D. Lin, E. Grimson, and J. Fisher. Construction of dependent Dirichlet processes based on Poisson processes. *NIPS*, 2010.
- [28] S. N. MacEachern. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, 1999.
- [29] A. Makadia, A. Patterson, and K. Daniilidis. Fully automatic registration of 3D point clouds. In *CVPR*, 2006.
- [30] K. V. Mardia and P. E. Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.
- [31] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012.
- [32] R. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.
- [33] X. Pan, J. E. Gonzalez, S. Jegelka, T. Broderick, and M. Jordan. Optimistic concurrency control for distributed unsupervised learning. In *NIPS*, 2013.
- [34] J. Reisinger, A. Waters, B. Silverthorn, and R. J. Mooney. Spherical topic models. In *ICML*, 2010.
- [35] G. Rosman, Y. Wang, X.-C. Tai, R. Kimmel, and A. M. Bruckstein. Fast regularization of matrix-valued images. In *ECCV*, 2012.
- [36] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.
- [37] A. Roychowdhury, K. Jiang, and B. Kulis. Small-variance asymptotics for hidden Markov models. In *NIPS*, 2013.
- [38] J. Straub, J. Chang, O. Freifeld, and J. W. Fisher III. A Dirichlet process mixture model for spherical data. In *AISTATS*, 2015.
- [39] J. Straub, G. Rosman, O. Freifeld, J. J. Leonard, and J. W. Fisher III. A mixture of Manhattan frames: Beyond the Manhattan world. In *CVPR*, 2014.
- [40] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *JMLR*, 3:583–617, 2003.
- [41] Y. W. Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 2010.
- [42] R. Triebel, W. Burgard, and F. Dellaert. Using hierarchical EM to extract planes from 3D range scans. In *ICRA*, 2005.
- [43] S. Zhong. Efficient online spherical k-means clustering. In *IJCNN*, volume 5, pages 3180–3185. 2005.