A Dirichlet Process Mixture Model for Spherical Data

Julian Straub, Jason Chang, Oren Freifeld, John W. Fisher III
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{jstraub, jchang7, freifeld, fisher}@csail.mit.edu

Abstract

Directional data, naturally represented as points on the unit sphere, appear in many applications. However, unlike the case of Euclidean data, flexible mixture models on the sphere that can capture correlations, handle an unknown number of components and extend readily to high-dimensional data have yet to be suggested. For this purpose we propose a Dirichlet process mixture model of Gaussian distributions in distinct tangent spaces (DP-TGMM) to the sphere. Importantly, the formulation of the proposed model allows the extension of recent advances in efficient inference for Bayesian nonparametric models to the spherical domain. Experiments on synthetic data as well as real-world 3D surface normal and 20-dimensional semantic word vector data confirm the expressiveness and applicability of the DP-TGMM.

1 Introduction

Many applications of interest involve measurements of directional data. In 3D scenes, unit-length surface normals extracted from point clouds [17, 24, 44] reside in a 2D manifold (i.e., the unit sphere in $\mathbb{R}^3$). In biology, protein backbone measurements are described and classified based on their angular configurations in the so-called Ramachandran plots [37]. Directional data also exists outside of the 3D world. E.g., the words counts in a corpus of documents can be viewed as directional data once normalized to have unit ℓ_2 norm. Word-frequency vectors are often clustered using the cosine similarity [10], which measures the cosine of the angle formed by two vectors. This measure essentially treats the word-frequency vectors as directional data, and has been shown to be superior to Euclidean distance for document clustering [46]. Another example of directional data is semantic word vectors [32], which associate a high-dimensional vector with each word in a given corpus. The semantic word vectors capture the semantic context of the associated words, and should not to be confused with the word-frequency vectors of documents. Again, cosine similarity is used as the distance measure to find words with similar meaning.

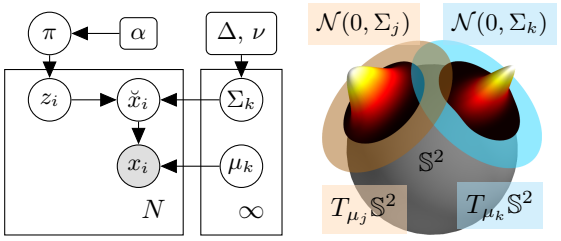


Figure 1: The graphical model of the proposed Dirichlet process tangential Gaussian mixture model (DP-TGMM) and an illustrative drawing for $K = 2$ clusters, k and j, in their respective tangent spaces to the sphere $\mathbb{S}^2$. For more details refer to Sec. 3.2.

One common task in many of these applications is to group the data into similar clusters. Due to the non-linearity of the hyper-sphere, clustering on the spherical manifold is often treated in an ad-hoc manner by either ignoring the geometry of the sphere or using overly-restricted models. In this work, we present a flexible Bayesian nonparametric (BNP) model for data residing on a hyper-sphere that respects the inherent geometry of the manifold. As shown in Fig. 1, our approach draws on the Dirichlet process Gaussian mixture model (DP-GMM), and models full covariance matrices on (linear) tangent spaces to the sphere, as opposed to the isotropic covariances associated with a von-Mises-Fisher (vMF) distribution [1, 2, 38, 49]. Importantly, the covariances of the Gaussians, capturing intra-cluster correlations, have analytical conjugate

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

Table 1: Properties of different typically-used clustering algorithms for directional data.

	k-means [21]	spkm [10, 49]	vMF-MM [1]	TGMM [12, 42]	DP-vMF-MM [2]	DP-GMM [6]	DP-TGMM (proposed)
Spherical geometry	·	✓	✓	✓	✓	·	✓
Bayesian inference	·	·	✓	✓	✓	✓	✓
Anisotropic covariance	·	·	·	✓	·	✓	✓
Bayesian nonparametric	·	·	·	·	✓	✓	✓
Parallelizeable	✓	✓	✓	✓	·	✓	✓

priors that enable efficient inference. Additionally, the approach transparently scales to high-dimensional data. We extend the efficient inference method of [6], a parallelized restricted Gibbs sampler with sub-cluster split/merge moves, to account for the geometry of the sphere. Moreover, we show how to combine sufficient statistics from tangent spaces around *different points of tangency* to propose merges efficiently.

To highlight the differences of the proposed Dirichlet process tangential Gaussian mixture model (DP-TGMM), we quantitatively compare it with four other methods on synthetic directional data with ground-truth labels. We demonstrate the scalability and efficiency of the inference algorithm as well as the applicability of the DP-TGMM to real-world directional data by modeling 3D surface normals extracted from point clouds. Furthermore, we show its scalability to higher dimensions by clustering the 20-dimensional semantic word vectors of 41k words extracted from the English Wikipedia corpus [48].

2 Related Work

We now discuss relevant work related to clustering directional data on a hyper-sphere. Many directional distributions exist (e.g., [3, 28, 31]). However, we focus our discussion on work using the von-Mises-Fisher (vMF) distribution [14] due to its popularity.

Several algorithms model directional data using a finite mixture of vMF distributions. Banerjee et al. [1] perform Expectation Maximization (EM) for a finite vMF mixture model to cluster text and genomic data. In the limit, when the vMF concentration parameter approaches infinity, this method simplifies to the spherical k-means (spkm) algorithm [10]. Zhong [49] extends the spkm algorithm to online clustering.

The vMF distribution has also been used in BNP mixture models. Bangert et al. [2] formulate a Dirichlet process (DP) [13] vMF mixture model. Their inference relies heavily on the conjugacy of the prior of the vMF mean and is difficult to generalize to non-conjugate priors. Furthermore, scaling this method to large datasets is problematic because the inference procedure is based on the Chinese Restaurant Process [36], which cannot be parallelized. Reisinger et al. [38] formulate a finite latent Dirichlet allocation model [5] for directional data using vMF distributions with fixed vMF concentration parameters. This is akin to using a GMM with a fixed variance, which is known to perform poorly if the model variance does not match the noise characteristics.

In general, using vMF distributions has two major flaws. First, the lack of a closed-form conjugate prior for the concentration parameter in the vMF distribution complicates posterior inference. Slice sampling methods [2] partially address this issue, but at the cost of extra computation. Often, the concentration parameter is still arbitrarily fixed and not inferred from the data (e.g., [38]). More importantly though, the vMF distribution is isotropic. That is, similar to a spherical Gaussian distribution, a vMF distribution cannot capture different variances in each dimension of the data or correlations between the dimensions. We note that the Fisher-Bingham distribution [28] generalizes the vMF distribution to anisotropic (i.e., elliptical) distributions on the sphere. While Peel et al. [34] propose an EM-based inference for finite mixtures of Fisher-Bingham distributions in 3D, extensions to higher dimensions are difficult due to the normalizer of the probability density function.

In other applications, the inherent geometry of the problem is ignored and, without taking the spherical geometry into account, algorithms developed for Euclidean geometry are used; e.g., k-means [21], the finite Gaussian mixture model (GMM) [4], as well as the Dirichlet process GMM [13, 30].

The work in protein-configuration modeling from Ramachandran plots [37] exemplifies this well. First Dahl et al. [9] introduced modeling the angular data as a DP-GMM, ignoring the spherical manifold of the angular data. To solve this issue Lennox et al. [29] model the data on the 3D sphere as a DP-vMF mixture, but require an approximation to the vMF posterior. Work by Ting et al. [47] uses an HDP with normal-inverse-Wishart base measure to share data between proteins, but does not respect the manifold of the data.

Approaches utilizing a single tangent space to define distributions over the hyper-sphere have been pro-

posed for rotation estimation and tracking [8, 19]. Finite mixture models of Gaussians in separate tangent spaces have been explored to estimate rigid-body motion in robotics [12] and for human body-pose regression [42]. While Simo-Serra et al. [42] show a way to reduce (but not increase) the number of clusters within their EM inference framework, both models are finite mixture models in contrast to our DP-based infinite mixture model.

In contrast to previous approaches, the proposed DP-TGMM allows for anisotropic distributions on the sphere, lends itself to consistent Bayesian inference, and adapts the model complexity to the observations. We develop a corresponding inference algorithm that can be parallelized and respects the geometry of the unit sphere. Table 1 highlights the differences between the proposed and previous approaches.

3 BNP Mixtures of Spherical Data

Classical Statistics rely on the Euclidean structure of $\mathbb{R}^{\mathbb{D}}$. Thus, due to the nonlinearity of the sphere, the statistical analysis of spherical data requires special care [15, 31]. In this section we introduce the Dirichlet process tangential Gaussian mixture model (DP-TGMM), a mixture model for data lying on the unit sphere, $\mathbb{S}^{D-1} = \{x : x^T x = 1\,; x \in \mathbb{R}^D\}$. Importantly, the model (as well as the inference algorithm; cf. Sec. 4) respects that the sphere is a $(D-1)$-dimensional nonlinear Riemannian manifold. Before introducing the probabilistic model, we now give a brief introduction to the geometric concepts used in the DP-TGMM. The interested reader can consult [11] for a more detailed discussion.

While $\mathbb{S}^{D-1}$ is nonlinear, every point, $p \in \mathbb{S}^{D-1}$, is associated with a linear *tangent space*, denoted $T_p\mathbb{S}^{D-1}$:

$$T_p\mathbb{S}^{D-1} \triangleq \{\breve{x} : p^T\breve{x} = 0\}\,. \tag{1}$$

Elements of $T_p\mathbb{S}^{D-1}$ are called *tangent vectors* and may be viewed as "arrows" based at p and tangent to $\mathbb{S}^{D-1}$. Note that $\dim(T_p\mathbb{S}^{D-1}) = D-1$ and that the point of tangency, p, may be identified with the origin of $T_p\mathbb{S}^{D-1}$ (i.e., a zero-length tangent vector).

Due to their linearity, tangent spaces often provide a convenient way to model spherical data. In fact, this is also true for more general manifolds [16, 23, 35, 43]. This linearity, together with mappings between $\mathbb{S}^{D-1}$ and $T_p\mathbb{S}^{D-1}$ (cf. Sec. 3.1), enables the modeling and clustering of data points via a zero-mean Gaussian distribution in a *cluster-dependent* tangent space. We illustrate this model in Fig. 2, where $\breve{x}$ denotes the point $x \in \mathbb{S}^{D-1}$ mapped to $T_p\mathbb{S}^{D-1}$.

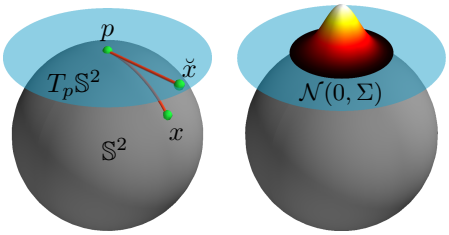


Figure 2: Left: The blue plane illustrates $T_p\mathbb{S}^2$, the tangent space to the sphere $\mathbb{S}^2$ at $p \in \mathbb{S}^2$. A tangent vector $\breve{x} \in T_p\mathbb{S}^2$ is mapped to $x \in \mathbb{S}^2$ via Exp_p. Right: We describe the data as zero-mean Gaussian in $T_\mu\mathbb{S}^2$.

3.1 Geometric Properties of $\mathbb{S}^{D-1}$ & $T_p\mathbb{S}^{D-1}$

In statistical modeling, distances are of paramount importance. On $\mathbb{S}^{D-1}$, rather than using the Euclidean distance of the ambient space, $\mathbb{R}^D$, an appropriate measure is the *geodesic distance* between points, which is simply the angle between them:

$$d_G : (p, q) \mapsto \arccos(p^T q)\,, \tag{2}$$

where $p, q \in \mathbb{S}^{D-1}$. The probability measure we will define on $\mathbb{S}^{D-1}$ will exploit this distance measure. A point, $x \in \mathbb{S}^{D-1} \setminus \{-p\}$, is mapped to a point, $\breve{x} \in T_p\mathbb{S}^{D-1}$, via the Riemannian logarithm:

$$\mathrm{Log}_p : x \mapsto \breve{x} = (x - p\cos\theta)\tfrac{\theta}{\sin\theta} \text{ (with } \tfrac{0}{\sin 0} = 1) \tag{3}$$

where $\theta = d_G(p, x)$. Conversely, $\breve{x} \in T_p\mathbb{S}^{D-1}$ is mapped to $x \in \mathbb{S}^{D-1}$ by the Riemannian exponential:

$$\mathrm{Exp}_p : \breve{x} \mapsto x = p\cos(||\breve{x}||_2) + \tfrac{\breve{x}}{||\breve{x}||_2}\sin(||\breve{x}||_2)\,. \tag{4}$$

The ℓ_2 norm $||\breve{x}||_2$ in $T_p\mathbb{S}^{D-1}$ is equal to the distance between p and x: $||\breve{x}||_2 = \theta = d_G(p, x)$. However, this is true only since p is the *point of tangency*. In general, the distance between two *other* points in $T_p\mathbb{S}^{D-1}$ is not equal to the geodesic distance between their corresponding points in $\mathbb{S}^{D-1}$. Thus, while a single zero-mean Gaussian in the tangent space around a point, p, provides an effective model for the within-cluster deviations from p (provided it is the *Karcher mean* of this cluster – see below), using a Gaussian mixture model whose (non-zero mean) components live on the *same* tangent space is a poor choice. Thus, in the proposed model, each mixture component exists in its own tangent space, and each tangent space is unique with certainty due to the continuous base measure.

Of special importance is the issue of selecting points of tangency. In this context we utilize the notion of the (so-called[1]) Karcher mean [20, 26], which generalizes

[1] See a recent discussion by Karcher [27].

the notion of the sample mean from $\mathbb{R}^D$ to Riemannian manifolds. Particularly for $\mathbb{S}^{D-1}$, it is defined as a local minimizer of the following function:

$$\langle x \rangle = \arg\min_{p \in \mathbb{S}^{D-1}} \sum_{i=1}^{N} d_G^2(p, x_i) \tag{5}$$

where $d_G(\cdot,\cdot)$ is given by Eq. (2). In practice, an iterative approach [35] (also included in our supplemental material) efficiently computes the Karcher mean.

3.2 Probabilistic DPMM for Spherical Data

The well known Dirichlet process [13] has been extensively used to model data in *Euclidean* spaces. A DP mixture model (DPMM) uses the DP as a prior to weight a countably-infinite set of clusters, where the distribution of weights is controlled by a concentration parameter, α. Here, we formulate the Dirichlet process tangential Gaussian mixture model (DP-TGMM) which extends DPMMs to data on the *unit sphere*, $\mathbb{S}^{D-1}$, in a manner that explicitly respects the intrinsic geometry. The graphical model is depicted in Fig. 1.

The generative DPMM first samples the infinite-length cluster proportions, π, from a stick-breaking process [41]. Then cluster assignments, $\mathbf{z} = \{z_i\}_{i=1}^N$, are sampled from the categorical distribution defined by π:

$$\pi \sim \mathrm{GEM}(1, \alpha)\,, \qquad z_i \sim \mathrm{Cat}(\pi_1, \pi_2, \dots)\,. \tag{6}$$

Associated with each cluster, $k \in \{1, \dots, \infty\}$, is a mean location on the sphere, μ_k, and a covariance, Σ_k, in the corresponding tangent space, $T_{\mu_k}\mathbb{S}^{D-1}$. These parameters are drawn from the following priors:

$$\mu_k \sim \mathrm{Unif}(\mathbb{S}^{D-1})\,, \qquad \Sigma_k \sim \mathrm{IW}(\Delta, \nu)\,, \tag{7}$$

where Unif and IW are the uniform and inverse-Wishart distributions, respectively. Note that $\mathbb{S}^{D-1}$ has a finite surface area which is used as the normalizing constant of Unif. This will later play a role in the inference procedure.

An observation, x_i, is drawn by sampling from a zero-mean Gaussian with covariance Σ_k in its corresponding tangent space, $T_{\mu_k}\mathbb{S}^{D-1}$, followed by mapping the point to $\mathbb{S}^{D-1}$ via the exponential map (Eq. (4)):

$$x_i \sim \mathrm{Exp}_{\mu_{z_i}}\left(\mathcal{N}(0, \Sigma_{z_i})\right) \quad \forall i \in \{1, \dots, N\}\,; \tag{8}$$

see Fig. 2. The Gaussian on $T_{\mu_{z_i}}\mathbb{S}^{D-1}$ induces a probability measure on $\mathbb{S}^{D-1}$. Note this statement is valid despite the fact that the Gaussian has infinite support in $T_{\mu_{z_i}}\mathbb{S}^{D-1}$ while the sphere is compact and the fact that $\mathrm{Exp}_{\mu_{z_i}}$ is not injective (note, however, that it is injective when restricted to $\mathrm{Log}_{\mu_{z_i}}(\mathbb{S}^{D-1} \setminus \{p\})$).

We now describe efficient MCMC inference for aforementioned geometry-respecting model.

4 Manifold-Aware MCMC Inference

Markov chain Monte Carlo (MCMC) techniques [39] provide a computational mechanism for sampling from complex Bayesian models. Unfortunately, in DP mixture models, MCMC methods are often slow. When parameters are marginalized, inference scales poorly because algorithms cannot be parallelized. When parameters are instantiated, the algorithm is parallelizable, but typically requires approximations and exhibits slow convergence. The recent DP sub-cluster algorithm of [6] addresses these issues by combining Metropolis-Hastings (MH) split/merge moves with a restricted Gibbs sampler, which is not allowed to add or remove clusters. The resulting Markov chain is guaranteed to converge to the desired posterior distribution. Additionally, this approach allows parallelization and the support of non-conjugate priors.

The DP sub-cluster algorithm proposes splits effectively via the MH framework [22] by exploiting an inferred auxiliary two-component, "sub-cluster" model for each regular cluster. The sub-clusters are inferred within the restricted Gibbs sampler. Excluding the varying complexity of posterior parameter sampling ($O(KD^2)$ for a GMM), the computational complexity per MCMC iteration is $O(NK + K^2)$, K is the maximum number of non-empty clusters. While the algorithm from [6] was originally suggested for DP models in $\mathbb{R}^D$, we show here that it can be extended to the DP-TGMM. This extension requires: (1) respecting the geometry of $\mathbb{S}^{D-1}$ when computing posterior distributions; and (2) combining sufficient statistics from different tangent spaces to propose splits efficiently. Additionally, we propose a MH algorithm to sample from the true posterior of the mean location on $\mathbb{S}^{D-1}$.

4.1 Restricted Gibbs Sampling

We now discuss restricted Gibbs sampling of the labels $\mathbf{z} \triangleq \{z_i\}_{i=1}^N$, means $\boldsymbol{\mu} \triangleq \{\mu_k\}_{k=1}^K$ and covariances $\boldsymbol{\Sigma} \triangleq \{\Sigma_k\}_{k=1}^K$ for K clusters. We note that each Σ_k is defined over a separate tangent space, $T_{\mu_k}\mathbb{S}^{D-1}$.

The covariances for each cluster are first sampled. Conditioned on the mean, μ_k, the data, $\mathbf{x} \triangleq \{x_i\}_{i=1}^N$, are modeled via a zero-mean Gaussian distribution in the tangent plane, $T_{\mu_k}\mathbb{S}^{D-1}$, as defined in Eq. (8). Hence, the same analysis as in the Euclidean space applies, and we sample $\boldsymbol{\Sigma}$ from the IW posterior [18]:

$$\Sigma_k \sim p(\Sigma_k | \mathbf{x}, \mathbf{z}, \hat{\mu}_k) = \mathrm{IW}(\Delta + S_k, \nu + N_k) \tag{9}$$

where $\mathcal{I}_k \triangleq \{i : z_i = k\}$ is the set of indices with label k, $N_k \triangleq |\mathcal{I}_k|$ counts the points assigned to cluster k, and S_{μ_k} is the scatter matrix at $T_{\mu_k}\mathbb{S}^{D-1}$, defined as:

$$S_{\mu_k} \triangleq \sum\nolimits_{i \in \mathcal{I}_k} \mathrm{Log}_{\mu_k}(x_i)\mathrm{Log}_{\mu_k}(x_i)^T. \tag{10}$$

Note, however, that the geometry of $\mathbb{S}^{D-1}$ renders the frequently-required computation of S_{μ_k} inefficient. The bottleneck of the calculation is attributed to the computationally-intensive evaluation of $\{\text{Log}_{\mu_k}(x_i)\}_{i\in\mathcal{I}_k}$ that depends on the point of tangency, μ_k, which constantly changes during inference.

To circumvent this issue, we exploit the fact that $\text{Log}_{\mu_k}(x_i) \approx \text{Log}_{\mu_k}(\langle x\rangle_k) + \text{Log}_{\langle x\rangle_k}(x_i)$, and make the following approximation for the scatter matrix:

$$S_{\mu_k} \approx S_{\langle x\rangle_k} + N_k \text{Log}_{\mu_k}(\langle x\rangle_k)\text{Log}_{\mu_k}(\langle x\rangle_k)^T, \quad (11)$$

where $\langle x\rangle_k$ is the Karcher mean of $\mathbf{x}_{\mathcal{I}_k}$ and $S_{\langle x\rangle_k}$ is the scatter matrix computed in the tangent plane of $\langle x\rangle_k$. This approximation is more efficient because the computation of $S_{\langle x\rangle_k}$ can be reused when μ_k changes. We will also use this approximation for proposing merges.

Conditioned on the sampled covariance matrix, Σ_k, we then sample μ_k. Ideally, we would sample directly from the following posterior distribution of μ_k:

$$\begin{aligned} p(\mu_k|\mathbf{x},\mathbf{z},\Sigma_k) &\propto p(\mu_k)p(\mathbf{x}|\mu_k,\mathbf{z},\Sigma_k) \quad (12)\\ &= p(\mu_k)\prod\nolimits_{i\in\mathcal{I}_k}\mathcal{N}(\text{Log}_{\mu_k}(x_i);0,\Sigma_k). \end{aligned}$$

Unfortunately, due to the nonlinearity of $\mathbb{S}^{D-1}$, this distribution cannot be expressed in a closed form. Instead, we utilize the MH framework to sample μ_k. It is well known in the literature that the closer the proposal distribution is to the target posterior distribution, the faster the convergence. We therefore use the following proposal as an approximation to Eqn. (12):

$$q(\mu_k|\mathbf{x},\mathbf{z},\Sigma_k) = p(\mu_k)\,\mathcal{N}(\text{Log}_{\langle x\rangle_k}(\mu_k);0,\tfrac{\Sigma_k}{N_k}) \quad (13)$$

See the supplemental material for more details. The proposed mean $\hat{\mu}_k$ is accepted according to the MH algorithm, with probability $\Pr(\text{accept}) = \min(1,r)$, where the Hastings ratio, r, is

$$\begin{aligned} r &= \tfrac{p(\mathbf{x}|\mathbf{z},\hat{\mu}_k,\Sigma_k)p(\hat{\mu}_k)q(\mu_k|\mathbf{x},\mathbf{z},\Sigma_k)}{p(\mathbf{x}|\mathbf{z},\mu_k,\Sigma_k)p(\mu_k)q(\hat{\mu}_k|\mathbf{x},\mathbf{z},\Sigma_k)} \quad (14)\\ &= \tfrac{\mathcal{N}(\text{Log}_{\langle x\rangle_k}(\mu_k);0,\Sigma_k/N_k)}{\mathcal{N}(\text{Log}_{\langle x\rangle_k}(\hat{\mu}_k);0,\Sigma_k/N_k)}\prod\nolimits_{i\in\mathcal{I}_k}\tfrac{\mathcal{N}(\text{Log}_{\hat{\mu}_k}(x_i);0,\Sigma_k)}{\mathcal{N}(\text{Log}_{\mu_k}(x_i);0,\Sigma_k)}. \end{aligned}$$

We have used the fact that the distribution of means, $p(\mu_k)$, is uniform over the sphere.

Finally, given means, $\boldsymbol{\mu}$, and covariances, $\boldsymbol{\Sigma}$, we sample new labels, $\mathbf{z}$, for all data, $\mathbf{x}$, as

$$z_i \overset{\propto}{\sim} \sum\nolimits_{k=1}^{K}\pi_k\,\mathcal{N}(\text{Log}_{\mu_k}(x_i);0,\Sigma_k)\mathbb{1}_{[z_i=k]}, \quad (15)$$

where $\overset{\propto}{\sim}$ denotes sampling from the distribution proportional to the right side, and the indicator function $\mathbb{1}_{[z_i=k]}$ is 1 if $z_i = k$ and 0 otherwise.

4.2 Sub-Cluster Split/Merge Proposals

We now describe the MH split-and-merge proposals that are specialized to the geometry of $\mathbb{S}^{D-1}$. The previously-defined posterior distributions for directional data uniquely define posterior inference in the sub-clusters [6]. When constructing split-and-merge moves, joint proposals over the entire latent space, $\{\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Sigma}\}$, must be constructed. The proposed labels, $\hat{\mathbf{z}}$, will be constructed from the inferred sub-clusters. Ideally, the parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, will be proposed from the true posteriors. However, as discussed previously, no conjugate prior exists for $\boldsymbol{\mu}$. Hence we propose the parameters from

$$\hat{\mu}_a \sim q(\mu_a|\mathbf{x},\mathbf{z}) = \mathcal{N}(\text{Log}_{\langle x\rangle_a}(\mu_a);0,\Sigma_a^\star), \quad (16)$$

$$\Sigma_a \sim p(\Sigma_a|\mathbf{x},\mathbf{z},\hat{\mu}_a), \quad (17)$$

where $\Sigma_a^\star \triangleq \arg\max_\Sigma \text{IW}(\Delta+S_{\langle x\rangle_a},\nu+N_a)$. The scatter matrix, $S_{\langle x\rangle_a}$, in $T_{\langle x\rangle_a}\mathbb{S}^{D-1}$ is computed according to Eq. (11), and $p(\Sigma_a|\mathbf{x},\mathbf{z},\hat{\mu}_a)$ is the true posterior denoted in Eq. (9). We note that Eqn. (16) uses the covariance $\Sigma_a^\star$ instead of Σ_a because Σ_a depends on $\hat{\mu}_a$ through the point of tangency.

The DP Sub-Cluster algorithm then deterministically constructs moves from the sub-clusters. We extend the algorithm to the DP-TGMM by splitting cluster a into clusters b and c with:

$$\begin{aligned} \hat{\mathbf{z}}_{\mathcal{I}_a} &= \text{split-}a(\mathbf{z}), \quad (\hat{\mu}_b,\hat{\mu}_c) = (\hat{\mu}_{a\ell},\hat{\mu}_{ar}),\\ (\hat{\Sigma}_b,\hat{\Sigma}_c) &\sim p(\hat{\Sigma}_b|\mathbf{x},\mathbf{z},\hat{\mu}_b)p(\hat{\Sigma}_c|\mathbf{x},\mathbf{z},\hat{\mu}_c), \quad (18) \end{aligned}$$

and by merging clusters b and c into cluster a with:

$$\begin{aligned} \hat{\mathbf{z}}_{\mathcal{I}_b\cup\mathcal{I}_c} &= \text{merge-}bc(\mathbf{z}), \quad \hat{\mu}_a \sim q(\mu_a|\mathbf{x},\mathbf{z}),\\ \hat{\Sigma}_a &\sim p(\hat{\Sigma}_a|\mathbf{x},\mathbf{z},\hat{\mu}_a), \quad (19) \end{aligned}$$

where split-$a(\mathbf{z})$ splits cluster a into clusters b and c deterministically based on the sub-cluster labels, and merge-$bc(\mathbf{z})$ merges clusters b and c into cluster a. We show in the supplement that this choice of proposals results in the following Hastings ratio for a split:

$$r_{\text{split}} = \tfrac{\alpha\Gamma(\hat{N}_b)\Gamma(\hat{N}_c)}{\Gamma(N_a)}\tfrac{p(\mathbf{x}|\hat{\mathbf{z}},\hat{\mu})p(\hat{\mu}_b)}{p(\mathbf{x}|\mathbf{z},\mu)}q(\mu_a|\mathbf{x},\mathbf{z}). \quad (20)$$

As shown in [6], any proposed deterministic merge move will be rejected with very high probability. As such, the DP Sub-Cluster incorporates a set of randomized split/merge proposals that are generated from a data-independent, two-dimensional Dirichlet distribution. We use this formulation as well, and introduce the following randomized split/merge proposals. The random splits are proposed as:

$$\begin{aligned} \hat{\mathbf{z}}_{\mathcal{I}_a} &\sim \text{DirMult}(\alpha/2,\alpha/2),\\ (\hat{\mu}_b,\hat{\mu}_c) &\sim q(\hat{\mu}_b|\mathbf{x},\hat{\mathbf{z}})q(\hat{\mu}_c|\mathbf{x},\hat{\mathbf{z}}),\\ (\hat{\Sigma}_b,\hat{\Sigma}_c) &\sim p(\hat{\Sigma}_b|\mathbf{x},\mathbf{z},\hat{\mu}_b)p(\hat{\Sigma}_c|\mathbf{x},\mathbf{z},\hat{\mu}_c), \quad (21) \end{aligned}$$

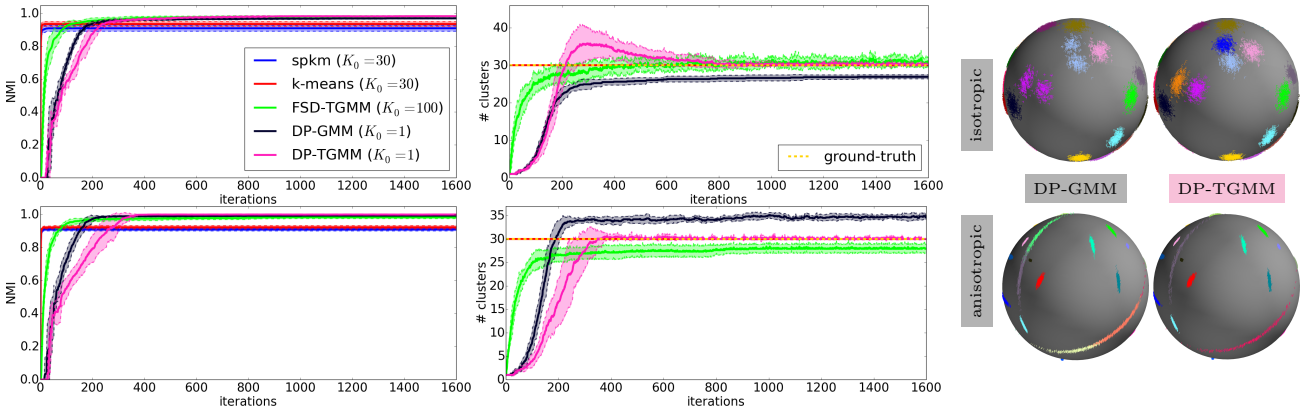


Figure 3: Mean and standard deviation over ten sampler runs of normalized mutual information (NMI) and cluster-count for synthetic datasets of 30 isotropic (top) and anisotropic (bottom) clusters on $\mathbb{S}^2$. The colors for the different algorithms are consistent across all plots. In the sphere-plots to the right it can be observed that, in contrast to the DP-TGMM, the DP-GMM fails to separate (top) or incorrectly splits (bottom) clusters.

and the random merges according to:

$$\hat{\mathbf{z}}_{\mathcal{I}_b \cup \mathcal{I}_c} = \text{merge-}bc(\mathbf{z}), \quad \mu_a \sim q(\hat{\mu}_a|\mathbf{x},\mathbf{z}),$$
$$\Sigma_a \sim p(\hat{\Sigma}_a|\mathbf{x},\mathbf{z},\hat{\mu}_a)\,. \tag{22}$$

We show in the supplementary material that these result in the following Hastings ratios:

$$r^{\text{rand}}_{\text{split}} = \frac{\alpha\Gamma(\alpha/2)^2\Gamma(\alpha+N_a)\Gamma(\hat{N}_b)\Gamma(\hat{N}_c)}{\Gamma(\alpha)\Gamma(N_a)\Gamma(\alpha/2+\hat{N}_b)\Gamma(\alpha/2+\hat{N}_c)} \cdot \frac{p(\mathbf{x}|\hat{\mathbf{z}},\hat{\mu})p(\hat{\mu}_b)}{p(\mathbf{x}|\mathbf{z},\mu)} \frac{q(\mu_a|\mathbf{x},\mathbf{z})}{q(\hat{\mu}_b|\mathbf{x},\hat{\mathbf{z}})q(\hat{\mu}_c|\mathbf{x},\hat{\mathbf{z}})} \tag{23}$$

$$r^{\text{rand}}_{\text{merge}} = \frac{\Gamma(\alpha)\Gamma(\hat{N}_a)\Gamma(\alpha/2+N_b)\Gamma(\alpha/2+N_c)}{\alpha\Gamma(\alpha/2)^2\Gamma(\alpha+\hat{N}_a)\Gamma(N_b)\Gamma(N_c)} \cdot \frac{p(\mathbf{x}|\mathbf{z},\hat{\mu})}{p(\mathbf{x}|\hat{\mathbf{z}},\mu)p(\mu_b)} \frac{q(\mu_b|\mathbf{x},\mathbf{z})q(\mu_c|\mathbf{x},\mathbf{z})}{q(\hat{\mu}_a|\mathbf{x},\hat{\mathbf{z}})}\,. \tag{24}$$

Note that while $\langle x\rangle_b$, $\langle x\rangle_c$, S_{μ_b}, and S_{μ_c} must be recomputed for the random split proposal, we efficiently approximate these quantities for the deterministic split and the random merge from the statistics of clusters b and c as described in the supplemental.

5 Experimental Results

In the following we compare the DP-TGMM inference algorithm on synthetic data with ground-truth labels against four related algorithms. Subsequently, we evaluate the clustering on real data, namely, surface normals extracted from Kinect depth images, and 20-dimensional semantic word vectors [32]. All MCMC inference algorithms are evaluated based on one sample from the Markov chain after burn-in.

5.1 Comparisons on Synthetic Data

We generate ground-truth data on the 3D unit sphere by sampling from a 30-component mixture model with equi-probable classes. The cluster means are drawn from a uniform distribution on the unit sphere and covariances from an IW prior. As depicted to the right in Fig. 3 the datasets for evaluation encompass an isotropic as well as an anisotropic dataset.

We compare the DP-TGMM with two commonly-used optimization-based clustering algorithms, k-means [21] and spherical k-means (spkm) [10], as well as with the finite symmetric Dirichlet approximation (FSD-TGMM) [25] to the DP-TGMM. Additionally, we show the performance of the DP-GMM, a BNP infinite GMM, that does not exploit the geometry of the sphere. DP-GMM inference uses the sub-cluster-split algorithm [6]. All algorithms are initialized with a random labeling of the data. We use normalized mutual information (NMI) [45] between the groundtruth and the inferred labels as a measure for clustering quality which penalizes the use of superfluous clusters. Additionally, we show the number of clusters per iteration, which changes only for the BNP models.

From the middle plots in Fig. 3 it can be seen that the inference for the DP-TGMM finds the true number of clusters in both cases, while the DP-GMM does not. The FSD-TGMM method gives an incorrect estimate of the number of clusters, which is consistent with what was observed in [7]. This motivates the need for our proposed sub-cluster inference algorithm. The depiction of the clustering results on the sphere to the right of Fig. 3, shows that the DP-GMM fails to separate isotropic clusters and splits anisotropic clusters incorrectly. The parametric algorithms, k-means and spkm, were set to the true number of clusters, which is unknown in many problems of interest.

The evolution of the NMI with iterations, depicted in the left plots of Fig. 3, shows that the optimization-based methods quickly converge to a (sub-optimal)

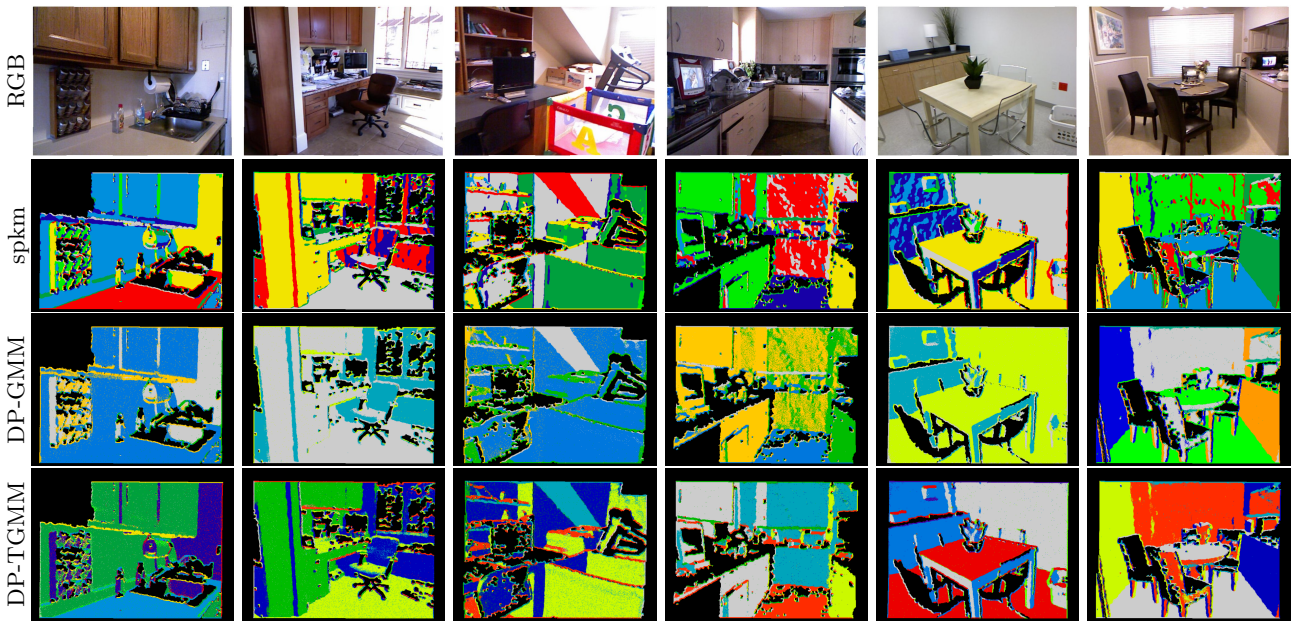


Figure 4: In the top row we show RGB images of the scene (only) for reference. We visualize the segmentation implied *solely* by the clustering of normals using the spkm algorithm with $k = 7$ (2nd row), using the DP-GMM (3rd row) and the proposed DP-TGMM (bottom row). Note how spkm and DP-GMM fail to properly segment the scenes. The colors encode the association to inferred clusters and black denotes missing depth data.

solution. The sampling based algorithms generally achieve better solutions. The DP-TGMM finds the best fit to the data as it explicitly allows for anisotropic distributions and respects the geometry of the sphere.

5.2 Surface Normals in Point-cloud Data

Surface normals extracted from point-clouds exhibit clusters on the unit sphere since planes in a scene create sets of normals pointing into the same direction. Hence, clustering these normals amounts to segmenting the scene into planes with similar orientation. We extract surface normals from raw Kinect depth images of the NYU V2 dataset [33] using the algorithm described in [24] and apply total variation regularization [40] to smooth the initial normal estimate.

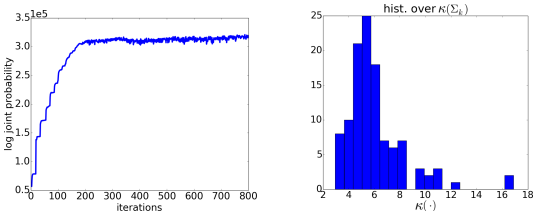
We show a set of examples scenes in Fig. 4 and the segmentation into planes with equal orientation as implied by the clustering of normals on the unit sphere obtained using three different algorithms. In the second row we display results from clustering with spkm where $k = 7$, in the third clusterings obtained using the DP-GMM sub-cluster algorithm and in the last row the segmentation obtained with the DP-TGMM inference algorithm. Note that the scene images in the first row are only for reference – only surface normals were used as input to the algorithms. The DP-GMM as well as the DP-TGMM inference was initialized to two clusters with the hyper-parameters of the IW prior set to $\nu = 10k$ and $\Delta = (12^\circ)^2\nu \mathrm{I}_{3\times 3}$. Each scene contains around $300k$ data points on $\mathbb{S}^2$.

The differences in segmentation illuminate the shortcomings of spkm and DP-GMM. The spkm algorithm finds a decent segmentation, but we get spurious clusters since the number of clusters is generally unknown. The inferred DP-GMM tends to under-segment the data because it ignores the manifold of the data and hence does not properly split clusters of normals in the presence of significant noise in the real data. For example in column two and five of Fig. 4 the floor and the wall are not separated into distinct clusters. By respecting the manifold as well as adopting a BNP model the DP-TGMM infers the intuitively correct segmentation as can be seen in the last row of Fig. 4.

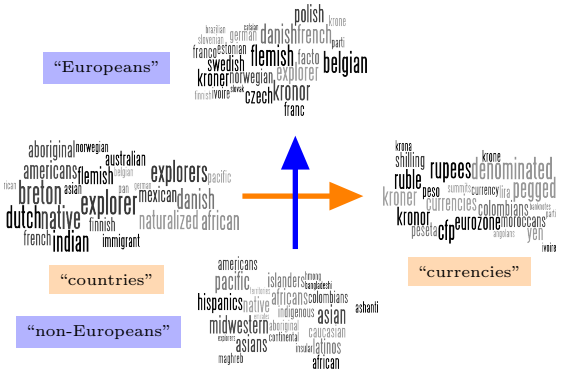
5.3 Clustering of Semantic Word-Vectors

We extract 20-dimensional semantic word vectors [32] from the English Wikipedia corpus and filter out all words with less than 100 counts to arrive at a set of 41k semantic word vectors for English words. Note that we normalize the word vectors to unit length before clustering. This is motivated by the fact that [32] utilizes the cosine similarity to find the semantically closest word to a given location in the vector space. The use of the cosine similarity is equivalent to the assumption that all the information about semantic proximity resides in the angular difference.

"finance"	"music"	"religion "	"leisure"	"government"	"food"
funding	symphonic	orthodoxy	malls	parliamentary	tomatoes
prospective	operatic	orthodox	hotels	enacted	edible
loans	soloists	evangelical	dining	parliament	fruit
financing	orchestral	christians	nightlife	delegation	meats
funds	music	primacy	outdoor	unanimously	meat
contracts	waltz	preaching	upscale	granting	vegetables
compensation	trios	doctrines	shopping	mandate	juice
regulations	lute	rabbis	restaurants	constitutional	baked
assets	soloist	clergy	taverns	citizenship	corn
investors	flute	catholicism	shops	committee	tasting



(a) Top: most likely words for 6 clusters. Bottom: log probability and histogram over condition numbers for clusters.



(b) Dimensionality reduction in the tangent space by projecting onto the eigenvectors of the two largest eigenvalues of the cluster's covariance matrix.

Figure 5: Evaluation of DP-TGMM inference on 20D semantic word vectors trained on the Wikipedia corpus.

Therefore, by clustering the semantic word-vectors by their directions we obtain clusters of semantically similar words as can be observed in the table of Fig. 5a. The table lists the ten most likely words of a subset of the clusters obtained when running the DP-TGMM inference algorithm. We start the algorithm from 20 centroids and run it for 800 iterations. After about 250 iterations the algorithm converges to 96 clusters. Note that this clustering is different from conventional topic modeling, which relies on document-level word counts. Semantic word-vectors depend on nearby words, and our clustering disregards document groupings.

To validate our hypothesis that real directional data exhibits anisotropic distributions on the sphere, we compute the condition number of the inferred cluster covariance matrices $\kappa(\Sigma_k) = \frac{\max[\sigma(\Sigma_K)]}{\min[\sigma(\Sigma_K)]}$ where $\sigma(\Sigma_K)$ is the set of all eigenvalues of Σ_k. The condition number thus serves as a measure for how elliptical the clusters are: $\kappa(\Sigma_k) \approx 1$ means the cluster is isotropic whereas $\kappa(\Sigma_k) \gg 1$ indicates an elliptical or anisotropic distribution. The spread-out histogram over condition numbers shown in Fig. 5a indicates that the inferred covariances are indeed anisotropic.

To demonstrate the analysis our model affords, we show in Fig. 5b the conceptual distribution of words in an inferred cluster when projected onto the 2D coordinate system defined by the two largest eigenvalues of the cluster's covariance matrix in the tangent space. We show 20 words that are at the extremes of both axes. In one direction the meaning of the words changes from non-European to European countries. On the other coordinate axes we find a progression from a mix of different countries to their currencies. We note that such analysis is made possible by modeling the data as a mixture of anisotropic distributions.

6 Conclusion

In this work we introduce the DP-TGMM, a DP mixture model over Gaussian distributions in multiple tangent spaces to the unit sphere in $\mathbb{R}^D$. Aimed at modeling directional data, this Bayesian nonparametric model does not only adapt to the complexity of the data but also describes anisotropic distributions on the sphere. Experiments on synthetic data demonstrate that the proposed model is more expressive in describing directional data than other commonly-used models. Moreover, we have shown the scalability and effectiveness of the inference algorithm as well as the applicability and versatility of the model on batches of 300k real-world 3D surface normals and on 20-dimensional semantic word-vectors for 41k English words. All inference code can be found at `http://people.csail.mit.edu/jstraub/`.

Future work should investigate the extension of this model to other Riemannian manifolds. Another promising research direction is embedding DP-TGMM in a hierarchical structure (akin to the Hierarchical DP-GMM for $\mathbb{R}^D$-valued data) to allow information sharing between batches of data in applications such as protein backbone configuration modeling.

Acknowledgements

We thank Karthik Narasimhan and Ardavan Saeedi for pointing us to the work on semantic word vectors. This work was partially supported by the ONR MURI N00014-11-1-0688, ARO MURI W911NF-11-1-0391, and the AFRL MIMFA program.

References

[1] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra, and G. Ridgeway. Clustering on the unit hypersphere using von Mises-Fisher distributions. *JMLR*, 2005.

[2] M. Bangert, P. Hennig, and U. Oelfke. Using an infinite von Mises-Fisher mixture model to cluster treatment beam directions in external radiation therapy. *ICMLA*, 2010.

[3] C. Bingham. An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, 1974.

[4] C. M. Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 2003.

[6] J. Chang and J. W. Fisher III. Parallel sampling of dp mixture models using sub-clusters splits. *NIPS*, 2013.

[7] J. Chang and J. W. Fisher III. MCMC sampling in HDPs using sub-clusters. *NIPS*, 2014.

[8] S. B. Choe. *Statistical analysis of orientation trajectories via quaternions with applications to human motion*. PhD thesis, The University of Michigan, 2006.

[9] D. B. Dahl, Z. Bohannan, Q. Mo, M. Vannucci, and J. Tsai. Assessing side-chain perturbations of the protein backbone: A knowledge-based classification of residue ramachandran space. *JMB*, 2008.

[10] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine learning*, 2001.

[11] M. P. do Carmo. *Riemannian Geometry*. Birkhäuser Verlag, Boston, MA, 1992.

[12] W. Feiten, M. Lang, and S. Hirche. Rigid motion estimation using mixtures of projected gaussians. *FUSION*, 2013.

[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] R. Fisher. Dispersion on a sphere. *Proc. R. Soc. A: Math. & Phy. Sci.*, 1953.

[16] O. Freifeld, S. Hauberg, and M. J. Black. Model transport: Towards scalable transfer learning on manifolds. *CVPR*, 2014.

[17] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing building interiors from images. *ICCV*, 2009.

[18] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

[19] J. S. Goddard and M. A. Abidi. Pose and motion estimation using dual quaternion-based extended kalman filtering. *Photonics West'98 Electronic Imaging*, 1998.

[20] K. Grove and H. Karcher. How to conjugatec 1-close group actions. *Mathematische Zeitschrift*, 1973.

[21] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, 1979.

[22] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 1970.

[23] S. Hauberg, F. Lauze, and K. Pedersen. Unscented Kalman filtering on Riemannian manifolds. *JMIV*, 2012.

[24] D. Holz, S. Holzer, and R. B. Rusu. Real-Time Plane Segmentation using RGB-D Cameras. *Proc. of the RoboCup Symposium*, 2011.

[25] H. Ishwaran and M. Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 2002.

[26] H. Karcher. Riemannian center of mass and mollifier smoothing. *Commun. Pure and Appl. Math.*, 1977.

[27] H. Karcher. Riemannian center of mass and so called karcher mean. *arXiv:1407.2087*, 2014.

[28] J. T. Kent. The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society*, 1982.

[29] K. P. Lennox, D. B. Dahl, M. Vannucci, and J. W. Tsai. Density estimation for protein conformation angles using a bivariate von Mises distribution and Bayesian nonparametrics. *JASA*, 2009.

[30] S. N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Commun. Stat. - Simul. and Comp.*, 1994.

[31] K. V. Mardia and P. E. Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *NIPS*, 2013.

[33] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from RGBD images. *ECCV*, 2012.

[34] D. Peel, W. J. Whiten, and G. J. McLachlan. Fitting mixtures of kent distributions to aid in joint set identification. *JASA*, 2001.

[35] X. Pennec. Probabilities and statistics on Riemannian manifolds: Basic tools for geometric measurements. *NSIP*, 1999.

[36] J. Pitman. Combinatorial stochastic processes. Technical Report 621, U.C. Berkeley Department of Statistics, 2002.

[37] C. Ramakrishnan and G. Ramachandran. Stereochemical criteria for polypeptide and protein chain conformations: Ii. allowed conformations for a pair of peptide units. *Biophysical Journal*, 1965.

[38] J. Reisinger, A. Waters, B. Silverthorn, and R. J. Mooney. Spherical topic models. *ICML*, 2010.

[39] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 1999.

[40] G. Rosman, Y. Wang, X.-C. Tai, R. Kimmel, and A. M. Bruckstein. Fast regularization of matrix-valued images. *ECCV*, 2012.

[41] J. Sethuraman. A constructive definition of Dirichlet priors. Technical report, DTIC Document, 1991.

[42] E. Simo-Serra, C. Torras, and F. Moreno-Noguer. Geodesic finite mixture models. *BMVC*, 2014.

[43] S. Sommer, F. Lauze, S. Hauberg, and M. Nielsen. Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximation. *ECCV*, 2010.

[44] J. Straub, G. Rosman, O. Freifeld, J. J. Leonard, and J. W. Fisher III. A mixture of Manhattan frames: Beyond the Manhattan world. *CVPR*, 2014.

[45] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *JMLR*, 2003.

[46] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. *AAAI*, 2000.

[47] D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan, and R. L. Dunbrack Jr. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLoS computational biology*, 2010.

[48] Wikipedia. Database download, June 2014.

[49] S. Zhong. Efficient online spherical k-means clustering. *IJCNN*, 2005.