Supplemental Material for
A Dirichlet Process Mixture Model for Spherical Data

Julian Straub, Jason Chang, Oren Freifeld, John W. Fisher III
{jstraub, jchang7, freifeld, fisher}@csail.mit.edu

1 Karcher Mean Algorithm

The Karcher mean [1, 2, 3] is a generalization of the standard sample mean to arbitrary manifolds M with associated an distance measure $d(\cdot,\cdot)$. It is defined as (local) minimizer to the following cost function:

$$\langle x \rangle = \arg\min_{p \in M} \sum_{i=1}^{N} w_i d^2(p, x_i)\,. \tag{1}$$

For our purposes $w_i = 1/N$ and $M = \mathbb{S}^{D-1}$, which implies the use of the geodesic distance metric.

On the unit sphere we can find the Karcher mean $\langle x \rangle$ by the following iterative algorithm:

- project data points $\{x_i\}_{i=1}^N$ into $T_p\mathbb{S}^{D-1}$ and compute mean $\langle \breve{x} \rangle = \frac{1}{N}\sum_{i=1}^N \mathrm{Log}_p(q_i)$
- project $\langle \breve{x} \rangle$ back onto the sphere to obtain updated $p' = \mathrm{Exp}_p(\langle \breve{x} \rangle)$. Set $p = p'$.
- iterate until $||\langle \breve{x} \rangle||_2$ close to 0 and then set the Karcher mean $\langle x \rangle = p$.

This algorithm takes the geometry of the sphere into account and exhibits fast convergence.

1.1 Weighted Mean of two Points on the Sphere

When merging two clusters a and b that have two different Karcher means $\langle x \rangle_a$ and $\langle x \rangle_b$, we want to compute the Karcher mean of the merged cluster efficiently without having to run the Karcher mean algorithm on the joint set of data points. Let cluster a contain N_a and cluster b N_b data points.

We approximate the Karcher mean of the merged cluster $\langle x \rangle_c$ as the weighted Karcher mean, of $\langle x \rangle_a$ and $\langle x \rangle_b$ with weights N_a and N_b respectively. Using Eq. (1) the optimization problem that will yield $\langle x \rangle_c$ becomes:

$$\langle x \rangle_c = \arg\min_{p \in \mathbb{S}^{D-1}} N_a \arccos(p^T \langle x \rangle_a)^2 + N_b \arccos(p^T \langle x \rangle_b)^2. \tag{2}$$

Since the geodesic between any two points is the shortest path on the manifold between the two of them, we know that p has to lie on the geodesic. On the unit sphere we can describe the location on the geodesic as a rotation about the axis defined by the cross product of the two vectors $\langle x \rangle_a$ and $\langle x \rangle_b$ by an angle θ_a, which we define such that the location of $\langle x \rangle_a$ on the geodesic has $\theta_a = 0$. This implies that the location of $\langle x \rangle_b$ on the geodesic has angle $\theta_b = \arccos(\langle x \rangle_a^T \langle x \rangle_b)$. With this intuition we can reformulate the optimization problem in terms of angles on the geodesic as:

$$\theta_a^\star = \arg\min_{\theta_a} N_a \theta_a^2 + N_b(\theta_a - \theta_b)^2\,. \tag{3}$$

The minimizer of this function is $\theta_a^\star = \frac{N_b}{N_a + N_b}\theta_b$. Hence we can compute $\langle x \rangle_c$ by rotating $\langle x \rangle_a$ by an angle of $\theta_a^\star$ about the aforementioned axis. The reader is referred to [4] for details on differentiable manifolds and geodesics.

2 Proposal Distribution for the Mean of a DP-TGMM Cluster

Since the sphere is a non-linear manifold it is, to our knowledge, not possible to derive a closed-form posterior distribution for the means μ_k. Instead, we utilize the Metropolis-Hastings framework to sample means μ_k from the true posterior using the following proposal distribution $q(\mu_k|\mathbf{x},\mathbf{z},\Sigma_k)$, which approximates the true posterior:

$$\begin{aligned} p(\mu_k|\mathbf{x},\mathbf{z},\Sigma_k) &\propto p(\mathbf{x}|\mu_k,\mathbf{z},\Sigma_k)p(\mu_k) = p(\mu_k)\prod\nolimits_{i\in\mathcal{I}_k}\mathcal{N}(\mathrm{Log}_{\mu_k}(x_i);0,\Sigma_k) \\ &\approx p(\mu_k)\mathcal{N}(\mathrm{Log}_{\langle x\rangle_k}(\mu_k);0,\Sigma_k/N_k) = q(\mu_k|\mathbf{x},\mathbf{z},\Sigma_k) \end{aligned} \tag{4}$$

where $\langle x\rangle_k$ is the Karcher mean of the data points $\mathbf{x}_{\mathcal{I}_k}$. The approximation lies in the assumption, that the data $\mathbf{x}_{\mathcal{I}_k}$ has a small spread which implies that $\theta_i = d_G(x_i,\mu_k) \approx \bar{\theta} = d_G(\langle x\rangle_k,\mu_k)$. This can be seen when looking more closely at the product of Gaussians:

$$\begin{aligned} &\textstyle\prod_{i\in\mathcal{I}_k}\mathcal{N}(\mathrm{Log}_{\mu_k}(x_i);0,\Sigma_k) \propto \\ &\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{N}\frac{\theta_i^2}{\sin^2\theta_i}x_i^T\Sigma^{-1}x_i - 2\sum_{i=1}^{N}\frac{\theta_i^2\cos\theta_i}{\sin^2\theta_i}x_i^T\Sigma^{-1}\mu + \sum_{i=1}^{N}\frac{\theta_i^2\cos^2\theta_i}{\sin^2\theta_i}\mu^T\Sigma^{-1}\mu\right]\right\} \\ &\approx \exp\left\{-\frac{1}{2}\left[N\frac{\bar{\theta}^2}{\sin^2\bar{\theta}}\langle x\rangle^T\Sigma^{-1}\langle x\rangle - 2N\frac{\bar{\theta}^2\cos\bar{\theta}}{\sin^2\bar{\theta}}\langle x\rangle^T\Sigma^{-1}\mu + N\frac{\bar{\theta}^2\cos^2\bar{\theta}}{\sin^2\bar{\theta}}\mu^T\Sigma^{-1}\mu\right]\right\} \\ &\propto \mathcal{N}(\mathrm{Log}_{\mu_k}(\langle x\rangle);0,\Sigma_k/N_k) = \mathcal{N}(\mathrm{Log}_{\langle x\rangle}(\mu_k);0,\Sigma_k/N_k) \end{aligned} \tag{5}$$

The equality in the last line stems from the fact, that both Gaussian densities live in the respective tangents spaces (around μ_k and $\langle x\rangle$) and have the same covariances. Therefore,$\mathrm{Log}_{\mu_k}(\langle x\rangle) = -\mathrm{Log}_{\langle x\rangle}(\mu_k)$ in the tangent planes and hence their pdf value will be equal due to the radial symmetry of the Gaussian density.

3 Sufficient Statistics in the Tangent Space

As defined in our paper, the distribution of the covariances $\boldsymbol{\Sigma}$ in the tangent spaces is inverse-Wishart (IW) conditioned on the associated means $\boldsymbol{\mu}$. Since the IW distribution is in the exponential family [5] of probability distributions, we only need the associated sufficient statistics of the data to evaluate the joint probability as well as to compute the posterior distribution in the tangent space. As described in the paper we compute the sufficient statistics, namely the Karcher mean $\langle x\rangle_k$, the scatter matrix S_k, and the number of data points in the cluster N_k for the Gibbs sampling of covariances $\boldsymbol{\Sigma}$ from the IW posterior.

3.1 Sufficient Statistics in a single Tangent Space

For a single D-TGMM cluster around the mean μ_k we would ideally bring all associated data $x_{\mathcal{I}_k}$ into the tangent space $T_{\mu_k}\mathbb{S}^{D-1}$ using $\mathrm{Log}_{\mu_k}(x_i)$ and compute the scatter matrix S_{μ_k} as

$$S_{\mu_k} = \sum_{i\in\mathcal{I}_k}\mathrm{Log}_{\mu_k}(x_i)\mathrm{Log}_{\mu_k}(x_i)^T\,. \tag{6}$$

The number of data points N_k and the scatter matrix S_{μ_k} are sufficient statics for the posterior covariance matrix of a zero-mean Gaussian distribution with IW prior on the covariance.

As pointed out in the paper, the issue with this approach is that whenever the point of tangency μ_k changes all $\{\mathrm{Log}_{\mu_k}(x_i)\}_{i\in\mathcal{I}_k}$ as well as S_{μ_k} have to be recomputed. To circumvent this problem, we use that

$$\mathrm{Log}_{\mu_k}(x_i) \approx \mathrm{Log}_{\mu_k}(\langle x\rangle_k) + \mathrm{Log}_{\langle x\rangle_k}(x_i)\,, \tag{7}$$

where $\langle x\rangle_k$ is the Karcher mean of $\{x_i\}_{\mathcal{I}_k}$ computed as described in Sec. 1. Under this approximation and

starting from Eq. (6) the scatter matrix S_{μ_k} can be approximated as

$$S_{\mu_k} \approx \sum_{i\in\mathcal{I}_k} \left[\mathrm{Log}_{\mu_k}(\langle x\rangle_k) + \mathrm{Log}_{\langle x\rangle_k}(x_i)\right] \left[\mathrm{Log}_{\mu_k}(\langle x\rangle_k) + \mathrm{Log}_{\langle x\rangle_k}(x_i)\right]^T \tag{8}$$

$$\approx \sum_{i\in\mathcal{I}_k} \mathrm{Log}_{\mu_k}(\langle x\rangle_k)\mathrm{Log}_{\mu_k}(\langle x\rangle_k)^T + 2\mathrm{Log}_{\langle x\rangle_k}(x_i)\mathrm{Log}_{\mu_k}(\langle x\rangle_k)^T + \mathrm{Log}_{\langle x\rangle_k}(x_i)\mathrm{Log}_{\langle x\rangle_k}(x_i)^T \tag{9}$$

$$\approx N_k\mathrm{Log}_{\mu_k}(\langle x\rangle_k)\mathrm{Log}_{\mu_k}(\langle x\rangle_k)^T + 2\sum_{i\in\mathcal{I}_k} \left[\mathrm{Log}_{\langle x\rangle_k}(x_i)\right] \mathrm{Log}_{\mu_k}(\langle x\rangle_k)^T + \sum_{i\in\mathcal{I}_k} \mathrm{Log}_{\langle x\rangle_k}(x_i)\mathrm{Log}_{\langle x\rangle_k}(x_i)^T \tag{10}$$

$$\approx N_k\mathrm{Log}_{\mu_k}(\langle x\rangle_k)\mathrm{Log}_{\mu_k}(\langle x\rangle_k)^T + \sum_{i\in\mathcal{I}_k} \mathrm{Log}_{\langle x\rangle_k}(x_i)\mathrm{Log}_{\langle x\rangle_k}(x_i)^T \tag{11}$$

$$\approx N_k\mathrm{Log}_{\mu_k}(\langle x\rangle_k)\mathrm{Log}_{\mu_k}(\langle x\rangle_k)^T + S_{\langle x\rangle_k}\,. \tag{12}$$

From Eq. (10) to Eq. (11) we have used the definition of the Karcher mean namely that $\sum_{i\in\mathcal{I}_k} \mathrm{Log}_{\langle x\rangle_k}(x_i) = 0$.

This approximation has the desired advantage that unless the set, $\mathcal{I}_k$, of data points associated with cluster k changes, we do not have to recompute $\langle x\rangle_k$ and $S_{\langle x\rangle_k}$. If the mean μ_k changes we can quickly update the scatter S_{μ_k} without having to iterate through all associated data points again since the computation of $N_k\mathrm{Log}_{\mu_k}(\langle x\rangle_k)\mathrm{Log}_{\mu_k}(\langle x\rangle_k)^T$ involves just one outer product and neither N_k nor $\langle x\rangle_k$ changes.

3.2 Merging Sufficient Statistics between Tangent Spaces

When we propose merges, and to compute the sufficient statistics of the "upper" cluster consisting of the left and right sub-clusters, we use the following approach to efficiently compute the needed sufficient statistics solely from the already computed sufficient statistics. While, we describe the process in the context of merging two clusters b and c into cluster a, the approach is the same for computing the sufficient statistics of the "upper" cluster from left and right sub-cluster.

Assume we want to merge cluster b with cluster c to obtain the merged cluster a. We need to obtain its Karcher mean $\langle x\rangle_a \in \mathbb{S}^{D-1}$ as well as the sufficient statistics $N_a = N_b + N_c$, $\langle \breve{x}\rangle_a = \frac{1}{N_a}\sum_{i:z_i=} \mathrm{Log}_{\mu_a}(x_i)$ and $S_a = \sum_{i:z_i=a} \mathrm{Log}_{\mu_a}(x_i)\mathrm{Log}_{\mu_a}(x_i)^T$.

Clearly, we could just compute the Karcher mean and the sufficient statistics from scratch each time we propose a merge. Instead, in order to save computations we want to reuse the already computed statistics and Karcher means for clusters b and c. The Karcher mean $\langle x\rangle_a$ can be computed as described in Sec. 1.1 from $\langle x\rangle_b$ to $\langle x\rangle_c$ together with the counts N_b and N_c.

The sufficient statistics for clusters b and c are computed in the tangent spaces around their respective Karcher means. Therefore their sample means $\langle \breve{x}\rangle_{b,c}$ in the tangent space will be very close to zero. However, the sample mean in $T_{\langle x\rangle_a}\mathbb{S}^{D-1}$ is generally non-zero and we compute it as the weighted mean between the sample means of cluster b and c mapped into $T_{\langle x\rangle_a}\mathbb{S}^{D-1}$:

$$\langle \breve{x}\rangle_a = \frac{1}{N_a}\left(N_b\mathrm{Log}_{\langle x\rangle_a}(\mathrm{Exp}_{\langle x\rangle_b}(\langle \breve{x}\rangle_b)) + N_c\mathrm{Log}_{\langle x\rangle_a}(\mathrm{Exp}_{\langle x\rangle_c}(\langle \breve{x}\rangle_c))\right) \tag{13}$$

Similarly, we can map the scatter matrices $S_{b,c}$ into $T_{\langle x\rangle_a}\mathbb{S}^{D-1}$ to obtain $\tilde{S}_{b,c}$ by making the following approximation:

$$\tilde{S}_b = \sum\nolimits_{\mathcal{I}_b} \mathrm{Log}_{\langle x\rangle_a}(x_i)\mathrm{Log}_{\langle x\rangle_a}(x_i)^T \tag{14}$$

$$\approx \sum\nolimits_{\mathcal{I}_b} \left(\mathrm{Log}_{\langle x\rangle_a}(\mu_b) + \mathrm{Log}_{\langle x\rangle_b}(x_i)\right) \left(\mathrm{Log}_{\langle x\rangle_a}(\mu_b) + \mathrm{Log}_{\langle x\rangle_b}(x_i)\right)^T \tag{15}$$

$$= \sum\nolimits_{\mathcal{I}_b} \mathrm{Log}_{\langle x\rangle_b}(x_i)\mathrm{Log}_{\langle x\rangle_b}(x_i)^T + 2\left(\sum\nolimits_{\mathcal{I}_b} \mathrm{Log}_{\langle x\rangle_b}(x_i)\right) \mathrm{Log}_{\langle x\rangle_a}(\mu_b)^T + +N_b\mathrm{Log}_{\langle x\rangle_a}(\mu_b)\mathrm{Log}_{\langle x\rangle_a}(\mu_b)^T \tag{16}$$

$$= S_b + N_b\mathrm{Log}_{\langle x\rangle_a}(\mu_b)\mathrm{Log}_{\langle x\rangle_a}(\mu_b)^T\,, \tag{17}$$

where we have used the fact, that the Karcher mean algorithm gives us $\langle x\rangle_b$ such that $\sum_{\mathcal{I}_b} \mathrm{Log}_{\langle x\rangle_b}(x_i) = 0$. Equation (17) gives us a approximate way of computing the statistics $\tilde{S}_b$ in $T_{\langle x\rangle_a}\mathbb{S}^{D-1}$ using only the already

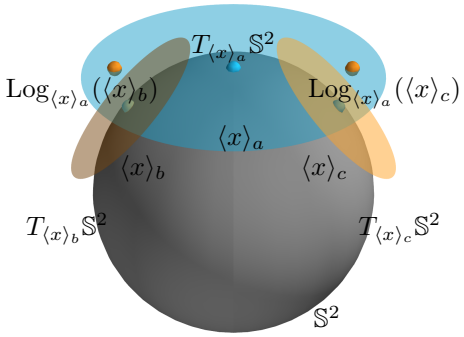

Figure 1: Illustration of the problem of computing sufficient statistics for cluster a from clusters b and c. Depicted are the tangent plane around the Karcher mean $\langle x\rangle_a$ of cluster a in blue and the two tangent planes by clusters b and c to the left and right in orange.

computed statistics S_b in $T_{\langle x\rangle_b}\mathbb{S}^{D-1}$ and the mean $\langle x\rangle_b$ of cluster b. Note that we can do exactly the same computation for cluster c to obtain $\tilde{S}_c$.

The approximation we made lies in the fact that $\{x_i\}_{\mathcal{I}_b}$ were linearized around $\langle x\rangle_b$ and hence the deviations from $\langle x\rangle_b$ which they describe are only valid in $T_{\langle x\rangle_b}\mathbb{S}^{D-1}$. By approximating

$$\mathrm{Log}_{\langle x\rangle_a}(x_i) \approx \mathrm{Log}_{\langle x\rangle_a}(\mu_b) + \mathrm{Log}_{\langle x\rangle_b}(x_i) \tag{18}$$

we make a small error that stems from the different linearizations. However, if the spread of cluster b (or c) is small, the approximation error is small. Using $\tilde{S}_{b,c}$ we compute the scatter matrix S_a of the merged cluster in $T_{\langle x\rangle_a}\mathbb{S}^{D-1}$ as

$$S_a = S_b + N_b\left(\langle\breve{x}\rangle_b\langle\breve{x}\rangle_b^T + \mathrm{Log}_{\langle x\rangle_a}(\langle x\rangle_b)\mathrm{Log}_{\langle x\rangle_a}(\langle x\rangle_b)^T\right) \tag{19}$$

$$+ S_c + N_c\left(\langle\breve{x}\rangle_c\langle\breve{x}\rangle_c^T + \mathrm{Log}_{\langle x\rangle_a}(\langle x\rangle_c)\mathrm{Log}_{\langle x\rangle_a}(\langle x\rangle_c)^T\right) - N_a\langle\breve{x}\rangle_a\langle\breve{x}\rangle_a^T \tag{20}$$

3.3 Metropolis-Hastings Ratio for Deterministic Split Proposal

Here we derive the Hastings ratio for a deterministic split proposal based on the sub-clusters.

In general the Hastings ratio for the DP-TGMM model is:

$$r = \frac{p(\mathbf{x},\hat{\mathbf{z}},\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})}{p(\mathbf{x},\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Sigma})}\frac{q(\mathbf{z},\boldsymbol{\Sigma},\boldsymbol{\mu})}{q(\hat{\mathbf{z}},\hat{\boldsymbol{\Sigma}},\hat{\boldsymbol{\mu}})}\,. \tag{21}$$

Since we can propose covariances $\hat{\boldsymbol{\Sigma}}$ from the posterior given the means $\hat{\boldsymbol{\mu}}$, we factor the Hastings ratio as follows:

$$r_{\text{split}} = \frac{p(\hat{\mathbf{z}})p(\hat{\boldsymbol{\mu}})p(\mathbf{x}|\hat{\mathbf{z}},\hat{\boldsymbol{\mu}})p(\hat{\boldsymbol{\Sigma}}|\mathbf{x},\hat{\mathbf{z}},\hat{\boldsymbol{\mu}})}{p(\mathbf{z})p(\boldsymbol{\mu})p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu})p(\boldsymbol{\Sigma}|\mathbf{x},\mathbf{z},\boldsymbol{\mu})}\frac{q(\mathbf{z})q(\boldsymbol{\mu}|\mathbf{x},\mathbf{z})p(\boldsymbol{\Sigma}|\mathbf{x},\mathbf{z},\boldsymbol{\mu})}{q(\hat{\mathbf{z}})q(\hat{\boldsymbol{\mu}}|\mathbf{x},\hat{\mathbf{z}})p(\hat{\boldsymbol{\Sigma}}|\mathbf{x},\hat{\mathbf{z}},\hat{\boldsymbol{\mu}})}\,. \tag{22}$$

We can further expand and simplify this to

$$r_{\text{split}} = \frac{p(\hat{\mathbf{z}})p(\hat{\boldsymbol{\mu}})p(\mathbf{x}|\hat{\mathbf{z}},\hat{\boldsymbol{\mu}})}{p(\mathbf{z})p(\boldsymbol{\mu})p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu})}\frac{q(\mathbf{z})q(\boldsymbol{\mu}|\mathbf{x},\mathbf{z})}{q(\hat{\mathbf{z}})q(\hat{\boldsymbol{\mu}}|\mathbf{x},\hat{\mathbf{z}})} = \frac{p(\hat{\mathbf{z}})q(\mathbf{z})}{p(\mathbf{z})q(\hat{\mathbf{z}})}\frac{p(\hat{\mu}_b)p(\hat{\mu}_c)p(\mathbf{x}|\hat{\mathbf{z}},\hat{\mu}_b)p(\mathbf{x}|\hat{\mathbf{z}},\hat{\mu}_c)q(\mu_a|\mathbf{x},\mathbf{z})}{p(\mu_a)p(\mathbf{x}|\mathbf{z},\mu_a)q(\hat{\mu}_b|\mathbf{x},\hat{\mathbf{z}})q(\hat{\mu}_c|\mathbf{x},\hat{\mathbf{z}})}\,. \tag{23}$$

Since the labels $\hat{\mathbf{z}}$ and the parameters $\hat{\boldsymbol{\mu}}_{b,c}$ are proposed analogous to [6] and because the prior distribution on the means $\boldsymbol{\mu}$ are uniform we have

$$r_{\text{split}} = \frac{\alpha\Gamma(\hat{N}_b)\Gamma(\hat{N}_c)}{\Gamma(N_a)}\frac{p(\hat{\mu})p(\mathbf{x}|\hat{\mathbf{z}},\hat{\mu}_b)p(\mathbf{x}|\hat{\mathbf{z}},\hat{\mu}_c)}{p(\mathbf{x}|\mathbf{z},\mu_a)}q(\mu_a|\mathbf{x},\mathbf{z})\,. \tag{24}$$

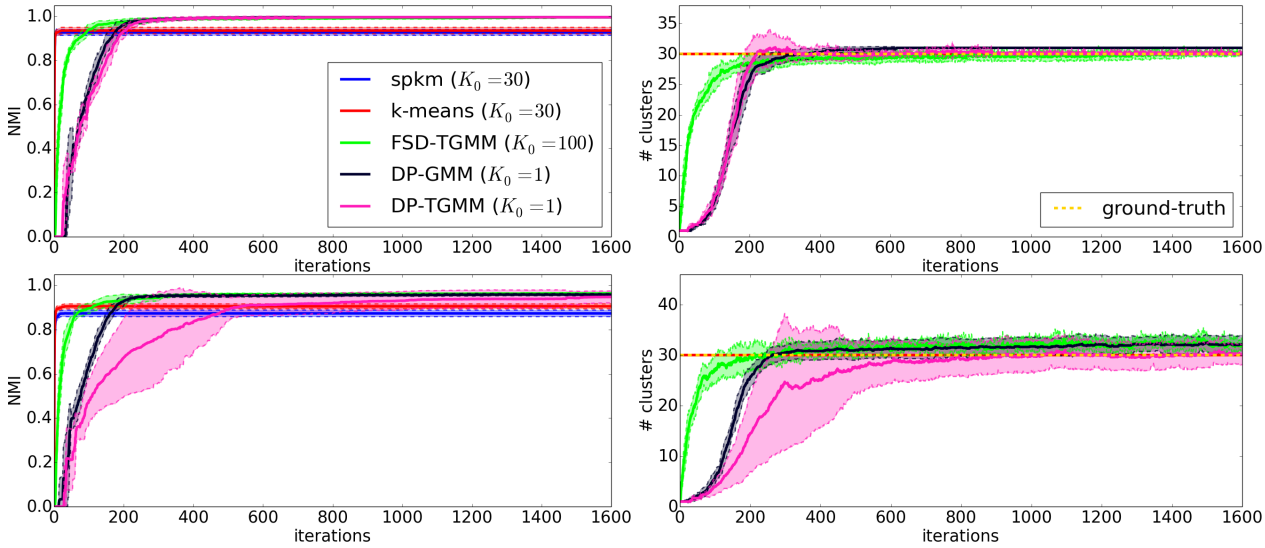

Figure 2: Mean and standard deviation over ten sampler runs of normalized mutual information (NMI) and cluster-count for synthetic datasets of 30 mixed isotropic and anisotropic clusters on $\mathbb{S}^2$. The colors for the different algorithms are consistent across all plots.

3.4 Metropolis-Hastings Ratio for Random Split/Merge Proposals

We derive the Metropolis-Hastings ratio for the random split/merge proposals starting from Eq. (24). Using derivations in [6] we arrive at

$$r^{\text{rand}}_{\text{split}} = \frac{\alpha\Gamma(\alpha/2)^2\Gamma(\alpha+N_a)\Gamma(\hat{N}_b)\Gamma(\hat{N}_c)}{\Gamma(\alpha)\Gamma(N_a)\Gamma(\alpha/2+\hat{N}_b)\Gamma(\alpha/2+\hat{N}_c)} \frac{p(\hat{\mu}_b)p(\mathbf{x}|\hat{\mathbf{z}},\hat{\mu}_b)p(\mathbf{x}|\hat{\mathbf{z}},\hat{\mu}_c)}{p(\mathbf{x}|\mathbf{z},\mu_a)} \frac{q(\mu_a|\mathbf{x},\mathbf{z})}{q(\hat{\mu}_b|\mathbf{x},\hat{\mathbf{z}})q(\hat{\mu}_c|\mathbf{x},\hat{\mathbf{z}})} \tag{25}$$

Similarly, we can derive the expression for a random merge as

$$r^{\text{rand}}_{\text{merge}} = \frac{\Gamma(\alpha)\Gamma(\hat{N}_a)\Gamma(\alpha/2+N_b)\Gamma(\alpha/2+N_c)}{\alpha\Gamma(\alpha/2)^2\Gamma(\alpha+\hat{N}_a)\Gamma(N_b)\Gamma(N_c)} \frac{p(\mathbf{x}|\hat{\mathbf{z}},\hat{\mu}_a)}{p(\mu_b)p(\mathbf{x}|\mathbf{z},\mu_b)p(\mathbf{x}|\mathbf{z},\mu_c)} \frac{q(\mu_b|\mathbf{x},\mathbf{z})q(\mu_c|\mathbf{x},\mathbf{z})}{q(\hat{\mu}_a|\mathbf{x},\hat{\mathbf{z}})} \tag{26}$$

4 Additional Synthetic Results

In Fig. 2 we show additional results for the synthetic data experiment. The two rows show the NMI and cluster-counts for two different mixed isotropic and anisotropic datasets. The dataset in the second row is more difficult since it contains more spread-out clusters with overlaps.

All the algorithms converge to close to the true number of clusters for the dataset in the first row. For the more difficult dataset in the second row, the DP-GMM and the FSD-TGMM both overestimate the number of clusters while the DP-TGMM converges to the true number of clusters on average. The larger standard deviation of the DP-TGMM and the overestimation of the number of clusters for the DP-GMM and the FSD-TGMM are likely due to the spread-out and overlapping clusters in the dataset.

References

[1] Karsten Grove and Hermann Karcher. How to conjugatec 1-close group actions. *Mathematische Zeitschrift*, 132(1):11–20, 1973.

[2] Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.

[3] Hermann Karcher. Riemannian center of mass and so called karcher mean. *arXiv preprint arXiv:1407.2087*, 2014.

[4] Manfredo Perdigao do Carmo. *Riemannian Geometry*. Birkhäuser Verlag, Boston, MA, 1992.

[5] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

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