

Supplementary Material: Small-Variance Nonparametric Clustering on the Hypersphere

Julian Straub Trevor Campbell Jonathan P. How John W. Fisher III
Massachusetts Institute of Technology

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

1. Dirichlet von-Mises-Fisher Mixture Model

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 ,

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

Let $\mathbf{z} = \{z_i\}_{i=1}^N$, and $\boldsymbol{\mu} = \{\mu_k\}_{k=1}^K$. Further, let negative subscript j , $\mathbf{u}_{-j} = \mathbf{u} \setminus u_j$ denote removal of item j from a set, and \mathcal{I}_k denote the set $\{i : z_i = k\}$. The Gibbs sampling inference algorithm for the finite vMF mixture iterates between sampling the label z_i given $\{\mathbf{z}_{-i}, \boldsymbol{\mu}\}$ and the mean direction μ_k given $\{\mathbf{z}, \boldsymbol{\mu}_{-k}\}$. These two steps are summarized as

$$p(z_i = k | \mathbf{z}_{-i}, \boldsymbol{\mu}, \mathbf{x}) \propto \mathbb{E} [\pi_k | \mathbf{z}_{-i}] \text{vMF}(x_i; \mu_k, \tau), \quad (2)$$

where $\mathbb{E} [\pi_k | \mathbf{z}_{-i}] = \frac{|\mathcal{I}_k| + \alpha_k}{N - 1 + \sum_k \alpha_k}$ and

$$p(\mu_k | \mathbf{z}, \boldsymbol{\mu}_{-k}, \mathbf{x}) = \text{vMF}(\mu_k; \frac{\vartheta_k}{\|\vartheta_k\|_2}, \|\vartheta_k\|_2), \quad (3)$$

where $\vartheta_k = \tau_0 \mu_0 + \tau \sum_{i \in \mathcal{I}_k} x_i$.

Parallel to the connection between k-means and the Gaussian mixture model in the small-variance asymptotic limit [2], taking $\tau \rightarrow \infty$ yields deterministic updates as also previously noted in [1]:

$$\begin{aligned} &\lim_{\tau \rightarrow \infty} p(z_i = k | \mathbf{z}_{-i}, \boldsymbol{\mu}, \mathbf{x}) \\ &= \lim_{\tau \rightarrow \infty} \frac{\mathbb{E} [\pi_k | \mathbf{z}_{-i}] \text{vMF}(x_i; \mu_k, \tau)}{\sum_{\kappa=1}^K \mathbb{E} [\pi_\kappa | \mathbf{z}_{-i}] \text{vMF}(x_i; \mu_\kappa, \tau)} \\ &= \lim_{\tau \rightarrow \infty} \frac{\mathbb{E} [\pi_k | \mathbf{z}_{-i}] \exp(\tau \mu_k^T x_i)}{\sum_{\kappa=1}^K \mathbb{E} [\pi_\kappa | \mathbf{z}_{-i}] \exp(\tau \mu_\kappa^T x_i)} \\ &= \begin{cases} 1 & x_i^T \mu_k \geq x_i^T \mu_\kappa \quad \forall \kappa \in \{1, \dots, K\} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (4)$$

and

$$\lim_{\tau \rightarrow \infty} p(\mu_k | \mathbf{z}, \boldsymbol{\mu}_{-k}, \mathbf{x}) = \delta \left(\mu_k = \frac{\sum_{i \in \mathcal{I}_k} x_i}{\|\sum_{i \in \mathcal{I}_k} x_i\|_2} \right). \quad (5)$$

These two updates together form the spherical k-means algorithm [1],

$$\begin{aligned} z_i &\leftarrow \arg \max_k \mu_k^T x_i \quad \forall i \in \{1, \dots, N\} \\ \mu_k &\leftarrow \frac{\sum_{i \in \mathcal{I}_k} x_i}{\|\sum_{i \in \mathcal{I}_k} x_i\|_2} \quad \forall k \in \{1, \dots, K\}. \end{aligned} \quad (6)$$

2. Proof of Theorem 1

Lemma 1 (Integration on a Manifold). *Suppose $M \subset \mathbb{R}^n$ is an m -dimensional differentiable manifold given by the parametric form $g : \mathbb{R}^m \rightarrow \mathbb{R}^n$, where $g \in C^1$, and $g(A) = M$ for some measurable $A \subset \mathbb{R}^m$, and let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be an integrable function on M . Then*

$$\int_M f(x) = \int_A f(g(a)) \sqrt{\det Dg^T Dg}$$

$$Dg = \begin{pmatrix} \frac{\partial g_1}{\partial a_1} & \cdots & \frac{\partial g_1}{\partial a_m} \\ \vdots & & \vdots \\ \frac{\partial g_n}{\partial a_1} & \cdots & \frac{\partial g_n}{\partial a_m} \end{pmatrix}. \quad (7)$$

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*

$$\lim_{\tau \rightarrow \infty} \frac{\int_M e^{\tau f(x)}}{\left(\frac{2\pi}{\tau}\right)^m |\det U^T \nabla_{x^*}^2 f(x^*) U|^{\frac{1}{2}} e^{\tau f(x^*)}} = 1 \quad (8)$$

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

Proof. Suppose $\{u_i\}_{i=1}^m$, $u_i \in \mathbb{R}^n$ is an orthonormal basis for $T_{x^*} M$, and add to it any orthonormal completion $\{u_i\}_{i=m+1}^n$ to \mathbb{R}^n . Then $U = [u_1 \ \cdots \ u_m] \in \mathbb{R}^{n \times m}$ is a matrix that maps $\mathbb{R}^m \rightarrow T_{x^*} M$. Finally, define $g : V \rightarrow M$ as the transformed exponential map $g(v) = \exp_{x^*}(Uv)$, where $UV \subset T_{x^*} M$ is the local domain of validity of the exponential map. Then by Lemma 1,

$$\int_M e^{\tau f(x)} = \int_{g(V)} e^{\tau f(x)} + \int_{M \setminus g(V)} e^{\tau f(x)} = \int_V e^{\tau h(v)} \sqrt{\det Dg^T Dg} + \int_{M \setminus g(V)} e^{\tau f(x)}. \quad (9)$$

where $h(v) \equiv f(g(v))$. First, note that

$$[Dg]_{ij}(0) = \frac{\partial g_i}{\partial v_j}(0) = \lim_{h \rightarrow 0} \frac{\exp_{x^*}(U1_j h) - \exp_{x^*}(0)}{h} \quad (10)$$

$$= \lim_{h \rightarrow 0} \frac{\exp_{x^*}(u_j h) - \exp_{x^*}(0)}{h} \quad (11)$$

$$= U_{ij}, \quad (12)$$

and thus by the fact that U is unitary and hence $U^T U = I$, $\sqrt{\det Dg(0)^T Dg(0)} = 1$. Next, using a second-order Taylor expansion of h ,

$$h(v) \simeq h(0) + \nabla_v h(0)^T v + \frac{1}{2} v^T \nabla_v^2 h(0) v \quad (13)$$

Note that $h(0) = f(x^*)$,

$$\left. \frac{\partial h}{\partial v_j} \right|_0 = \sum_{k=1}^n \left. \frac{\partial f}{\partial x_k} \frac{\partial g_k}{\partial v_j} \right|_0 = \sum_{k=1}^n U_{kj} \left. \frac{\partial f}{\partial x_k} \right|_{x^*} = u_j^T \nabla_{x^*} f(x^*) \quad (14)$$

and since f reaches a maximum at x^* on M , and u_j span $T_{x^*} M$,

$$\nabla_v h(0) = 0. \quad (15)$$

Further,

$$\frac{\partial^2 h}{\partial v_i \partial v_j} = \sum_{k=1}^n \frac{\partial f}{\partial x_k} \frac{\partial^2 g_k}{\partial v_i \partial v_j} + \sum_{l=1}^n \frac{\partial^2 f}{\partial x_k \partial x_l} \frac{\partial g_k}{\partial v_j} \frac{\partial g_l}{\partial v_i} \quad (16)$$

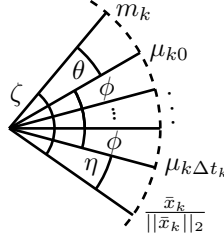


Figure 1: Geometry of the maximum likelihood setting of $\mu_{k0}, \mu_{k1}, \dots, \mu_{k\Delta t_k}$ for the transition distribution.

once again, since f reaches a maximum at x^* on M , and $\frac{\partial^2 g_k}{\partial v_i \partial v_j}$ has columns contained in $T_{x^*}M$, the first term is zero and thus

$$\frac{\partial^2 h}{\partial v_i \partial v_j} \Big|_0 = \sum_{k=1}^n \sum_{l=1}^n \frac{\partial^2 f}{\partial x_k \partial x_l} U_{kj} U_{li} \Big|_{x^*} \quad (17)$$

so

$$\nabla_v^2 h(0) = U^T \nabla_x^2 f(x^*) U. \quad (18)$$

Returning to the integral from earlier,

$$\lim_{\tau \rightarrow \infty} \frac{\int_M e^{\tau f(x)}}{\left(\frac{2\pi}{\tau}\right)^m |\det U^T \nabla_x^2 f(x^*) U|^{\frac{1}{2}} e^{\tau f(x^*)}} \quad (19)$$

$$= \lim_{\tau \rightarrow \infty} \frac{\int_V e^{\tau h(v)} \sqrt{\det Dg^T Dg} + \int_{M \setminus g(V)} e^{\tau f(x)}}{\left(\frac{2\pi}{\tau}\right)^m |\det U^T \nabla_x^2 f(x^*) U|^{\frac{1}{2}} e^{\tau f(x^*)}} \quad (20)$$

$$= \lim_{\tau \rightarrow \infty} \frac{\int_V e^{\tau(h(v)-f(x^*))} \sqrt{\det Dg^T Dg} + \int_{M \setminus g(V)} e^{\tau(f(x)-f(x^*))}}{\left(\frac{2\pi}{\tau}\right)^m |\det U^T \nabla_x^2 f(x^*) U|^{\frac{1}{2}}}. \quad (21)$$

By assumption, $\max_{x \in M \setminus g(V)} f(x) \leq f(x^*) - \epsilon$ for some $\epsilon > 0$, so using the Laplace approximation for multivariate Euclidean spaces,

$$\lim_{\tau \rightarrow \infty} \frac{\int_M e^{\tau f(x)}}{\dots} \leq \lim_{\tau \rightarrow \infty} \frac{\int_V e^{\tau(h(v)-f(x^*))} \sqrt{\det Dg^T Dg} + \text{vol}(M \setminus g(V)) e^{-\tau\epsilon}}{\left(\frac{2\pi}{\tau}\right)^m |\det U^T \nabla_x^2 f(x^*) U|^{\frac{1}{2}}} \quad (22)$$

$$= \lim_{\tau \rightarrow \infty} \frac{\left(\frac{2\pi}{\tau}\right)^m |\det U^T \nabla_x^2 f(x^*) U|^{\frac{1}{2}} \sqrt{\det Dg(0)^T Dg(0)}}{\left(\frac{2\pi}{\tau}\right)^m |\det U^T \nabla_x^2 f(x^*) U|^{\frac{1}{2}}} \quad (23)$$

$$= 1 \quad (24)$$

and since $\int_{M \setminus g(V)} e^{\tau f(x)} \geq 0$, the lower bound is also 1, and the result follows. \square

3. Solving for State Transition Geometry

As depicted in Fig. 1 the setting of $\mu_{k0}, \mu_{k1}, \dots, \mu_{k\Delta t_k}$ for the transition distribution can be described in terms of the angles $\phi, \eta,$ and θ . As derived in the paper the maximum likelihood parameters ϕ^*, η^* and θ^* for the transition distribution satisfy the following equations:

$$w_k \sin(\theta^*) = \beta \sin(\phi^*) = \|\bar{x}_k\|_2 \sin(\eta^*)$$

$$\zeta = \theta^* + \Delta t_k \phi^* + \eta^* = \arccos \left(m_k^T \frac{\bar{x}_k}{\|\bar{x}_k\|_2} \right) \quad (25)$$

where ζ is defined as the full angle between $\frac{\bar{x}_k}{\|\bar{x}_k\|_2}$ and m_k . Note that for the label assignment $\bar{x}_k = x_i$ while for the weight update $\bar{x}_k = \sum_{i \in \mathcal{I}_k} x_i$.

This set of equations can be solved exactly and efficiently using Euler’s method which in practice converges very quickly for this problem. Starting from $\phi = 0$, Euler’s method iterates over the following steps until convergence of ϕ :

1. compute $f(\phi) = \arcsin\left(\frac{\beta}{w_k} \sin(\phi)\right) + \Delta t_k \phi + \arcsin\left(\frac{\beta}{\|\bar{x}_k\|_2} \sin(\phi)\right) - \zeta$
2. compute $df(\phi) = \frac{\partial f(\phi)}{\partial \phi} = \Delta t_k + \frac{\beta \cos(\phi)}{\sqrt{\|\bar{x}_k\|_2^2 - \beta^2 \sin^2(\phi)}} + \frac{\beta \cos(\phi)}{\sqrt{w_k^2 - \beta^2 \sin^2(\phi)}}$
3. $\phi \leftarrow \phi - \frac{f(\phi)}{df(\phi)}$

A second faster but approximate approach is to assume that all angles are small. For a small angle α the following approximation is often used $\sin(\alpha) \approx \alpha$. With this we obtain the following closed form solutions:

$$\phi^* \approx \zeta \left[\beta \left(1 + \frac{1}{w_k} \right) + \Delta t_k \right]^{-1} \quad (26)$$

$$\theta^* \approx \zeta \left[1 + w_k \left(1 + \frac{\Delta t_k}{\beta} \right) \right]^{-1} \quad (27)$$

$$\eta^* \approx \zeta \left[1 + \frac{1}{w_k} + \frac{\Delta t_k}{\beta} \right]^{-1} \quad (28)$$

4. DP-vMF-means algorithm

The DP-vMF-means algorithm is given in full detail in Alg. 1. For the label assignment step both algorithms 2 and 3 can be used. The sequential label assignment algorithm 2 is directly derived from the Gibbs sampling posterior. Alg. 3 details the optimistic iterated restarts (OIR) algorithm which can be massively parallelized both in CPU and GPU.

Algorithm 1 DP-vMF-means algorithm

```

1:  $J_{\text{DP-vMF}} \leftarrow \infty$ 
2:  $\mu \leftarrow \emptyset$ 
3: while  $J_{\text{DP-vMF}}$  not converged do
4:    $\{z_i\}_{i=1}^N, \mu \leftarrow \text{DP-vMF-MEANS LABEL ASSIGNMENTS}(\{x_i\}_{i=1}^N, \mu, \lambda)$ 
5:   for  $k \in \{1, \dots, |\mu|\}$  do
6:     if  $n_k > 0$  then
7:        $\mu_k \leftarrow \frac{\sum_{i \in \mathcal{I}_k} x_i}{\|\sum_{i \in \mathcal{I}_k} x_i\|_2}$ 
8:     else
9:        $\mu \leftarrow \mu \setminus \mu_k$  ▷ remove cluster  $k$ 
10:    end if
11:  end for
12:   $J_{\text{DP-vMF}} \leftarrow \sum_{k=1}^{|\mu|} \sum_{i \in \mathcal{I}_k} x_i^T \mu_k + \lambda |\mu|$ 
13: end while

```

5. DDP-vMF-means algorithm

Algorithm 4 outlines the necessary operations of the DDP-vMF-means algorithm per timestep. The sequential label assignment algorithm for DDP-vMF-means is shown in Alg. 5. The OIR label assignment algorithm for DDP-vMF-means follows the same pattern as Alg. 3 with the additional possibility of reviving a cluster. Reviving a cluster changes the number of active clusters and thus requires a restart as well.

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*, 6(9), 2005.
- [2] B. Kulis and M. I. Jordan. Revisiting k-means: New algorithms via Bayesian nonparametrics. In *ICML*, 2012.

Algorithm 2 DP-vMF-means sequential label assignments algorithm

```
1: function DP-vMF-MEANSSEQUENTIALLABELASSIGNMENT( $\{x_i\}_{i=1}^N, \mu, \lambda$ )
2:   for  $i \in \{1, \dots, N\}$  do
3:      $z_i \leftarrow \arg \max_{k \in \{1, \dots, |\mu|+1\}} \begin{cases} x_i^T \mu_k & k \leq |\mu| \text{ and } |\mathcal{I}_k \setminus z_i| > 0 \\ \lambda + 1 & k = |\mu| + 1 \end{cases}$   $\triangleright$  only consider clusters that contain more than  $x_i$ 
4:     if  $z_i = |\mu| + 1$  then
5:        $\mu \leftarrow \mu \cup \{\mu_{|\mu|+1} \leftarrow x_i\}$   $\triangleright$  add cluster  $K + 1$  and initialize to  $x_i$ 
6:     end if
7:   end for
8:   return  $\{z_i\}_{i=1}^N, \mu$ 
9: end function
```

Algorithm 3 DP-vMF-means OIR label assignments algorithm

```
1: function DP-vMF-MEANSOIRLABELASSIGNMENT( $\{x_i\}_{i=1}^N, \mu, \lambda$ )
2:    $I \leftarrow N$ 
3:   repeat
4:     for  $i \in \{I, \dots, N\}$  in parallel do
5:        $z_i \leftarrow \arg \max_{k \in \{1, \dots, |\mu|+1\}} \begin{cases} x_i^T \mu_k & k \leq |\mu| \text{ and } |\mathcal{I}_k \setminus z_i| > 0 \\ \lambda + 1 & k = |\mu| + 1 \end{cases}$   $\triangleright$  only consider clusters that contain more than  $x_i$ 
6:       if  $z_i = |\mu| + 1$  or any( $|\mathcal{I}_k| = 0$ ) then
7:         atomic:  $I = \min(I, i)$   $\triangleright$  obtain the first index of cluster number change
8:       end if
9:     end for
10:    if  $I < N$  then
11:      if  $z_I = |\mu| + 1$  then
12:         $\mu \leftarrow \mu \cup \{\mu_{|\mu|+1} \leftarrow x_I\}$   $\triangleright$  add cluster  $K + 1$  and initialize to  $x_I$ 
13:      else
14:         $\mu \leftarrow \mu \setminus \mu_{k:|\mathcal{I}_k|=0}$   $\triangleright$  remove empty cluster
15:      end if
16:    end if
17:  until  $I = N$ 
18:  return  $\{z_i\}_{i=1}^N, \mu$ 
19: end function
```

Algorithm 4 DDP-vMF-means algorithm for a single time-step

```
1:  $\mu \leftarrow \emptyset$ 
2:  $\{z_i\}_{i=1}^N \leftarrow$  unassigned
3: while not converged do
4:    $\{z_i\}_{i=1}^N, \mu \leftarrow$  DDP-VMF-MEANSLABELASSIGNMENTS( $\{x_i\}_{i=1}^N, \mu, \lambda$ )
5:   for  $k \in \{1, \dots, |\mu|\}$  do
6:     if  $|\mathcal{I}_k| > 0$  then ▷ if cluster is instantiated in current timestep
7:       if  $k < \mu_{t-1}$  then ▷ cluster is not a novel cluster from this timestep
8:          $\mu_k \leftarrow R(\eta^*) \frac{\bar{x}_k}{\|\bar{x}_k\|_2}$  ▷ reinstantiate the cluster
9:       else
10:         $\mu_k \leftarrow \frac{\sum_{\mathcal{I}_k} x_i}{\|\sum_{\mathcal{I}_k} x_i\|_2}$  ▷ update the cluster center of the novel cluster
11:      end if
12:    end if
13:  end for
14: end while
15: for  $k \in \{1, \dots, |\mu|\}$  do
16:   if  $|\mathcal{I}_{z_i}| > 0$  then ▷ if cluster is instantiated in current timestep
17:     Solve for  $\phi^*, \theta^*, \eta^*$  in Eq. (25) with  $\bar{x}_k = \sum_{i \in \mathcal{I}_k} x_i$  as described in Sec. 3
18:      $w_k = w_k \cos(\theta^*) + \beta \Delta t_k \cos(\phi^*) + \|\bar{x}_k\|_2 \cos(\eta^*)$  ▷ update weight
19:   end if
20:   if  $Q \Delta t_k < \lambda$  then ▷ cluster cannot be revived again
21:      $\mu \leftarrow \mu \setminus \mu_k$  ▷ remove the cluster
22:   end if
23: end for
```

Algorithm 5 DDP-vMF-means sequential label assignments

```
1: function DDP-VMF-MEANSSEQUENTIALLABELASSIGNMENT( $\{x_i\}_{i=1}^N, \mu, \lambda$ )
2:   for  $i \in \{1, \dots, N\}$  do
3:     Solve for  $\phi^*, \theta^*, \eta^*$  in Eq. (25) with  $\bar{x}_k = x_i$  as described in Sec. 3
4:      $z_i \leftarrow \arg \max_{k \in \{1, \dots, |\mu|+1\}} \begin{cases} \lambda + 1 & k = |\mu| + 1 \\ \mu_k^T x_i & k \leq |\mu|, |\mathcal{I}_k \setminus z_i| > 0 \\ \Delta t_k \beta (\cos(\phi^*) - 1) + \Delta t_k Q \\ + w_k (\cos(\theta^*) - 1) + \cos(\eta^*) & k \leq |\mu|, |\mathcal{I}_k| = 0. \end{cases}$ 
5:     if  $z_i = |\mu| + 1$  then
6:        $\mu \leftarrow \mu \cup \{|\mu|+1 \leftarrow x_i\}$  ▷ add cluster  $K + 1$  and initialize to  $x_i$ 
7:     else
8:       if  $|\mathcal{I}_{z_i}| = 0$  then ▷ cluster  $z_i$  is not instantiated yet since it does not have data associated
9:          $\mu_k \leftarrow R(\eta^*) x_i$  ▷ reinstantiate the cluster
10:      end if
11:    end if
12:  end for
13:  return  $\{z_i\}_{i=1}^N, \mu$ 
14: end function
```
