Geometric Data Analysis Reading Group

Kernel Smoothing, Mean Shift, and Their Learning Theory with Directional Data

Yikun Zhang
Joint work with *Yen-Chi Chen*

Department of Statistics,
University of Washington

November 23, 2020

UNIVERSITY · OF · WASHINGTON · 1861 ·
LVX SIT

Outline

1. Introduction
2. Mean Shift Algorithm with Directional Data
3. Statistical Learning Theory of Directional KDE and its Derivatives
4. Computational Learning Theory of Directional Mean Shift Algorithm
5. Real-World Applications
6. Conclusion and Future Directions

Introduction

Why Do We Study Directional Data?

In some real-world applications, observations are inherently directional (or L_2 normalized) in nature. Data of this type arise in:

Why Do We Study Directional Data?

In some real-world applications, observations are inherently directional (or L_2 normalized) in nature. Data of this type arise in:

- **Astronomy**: locations of galaxies or stars, measurements (such as craters, lakes, etc.) on Earth or other planets.
- **Seismology**: epicenters of earthquakes.
- **Biology**: yeast gene expression analysis, studies of animal navigation.
- **Text mining**: similarities between text documents are effectively measured by *cosine similarity*.
- **Geology**, **Meteorology**, **Electromagnetism**,...

Why Do We Study Directional Data?

In some real-world applications, observations are inherently directional (or L_2 normalized) in nature. Data of this type arise in:

- **Astronomy**: locations of galaxies or stars, measurements (such as craters, lakes, etc.) on Earth or other planets.
- **Seismology**: epicenters of earthquakes.
- **Biology**: yeast gene expression analysis, studies of animal navigation.
- **Text mining**: similarities between text documents are effectively measured by *cosine similarity*.
- **Geology**, **Meteorology**, **Electromagnetism**,...

Challenge and Difficulty: Statistical models in the Euclidean space are inadequate for analyzing such data, which are assumed to lie on a (unit) hypersphere, a non-linear manifold.

Directional Data

Typically, a directional dataset is assumed to be

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} f,$$

where f is a directional density supported on

$$\Omega_q := \left\{\boldsymbol{x} \in \mathbb{R}^{q+1} : ||\boldsymbol{x}||_2 = 1\right\}$$

with $\int_{\Omega_q} f(\boldsymbol{x})\, \omega_q(d\boldsymbol{x}) = 1$ and $||\cdot||_2$ is the L_2-norm in $\mathbb{R}^{q+1}$. Here, ω_q is the Lebesgue measure on Ω_q.

Directional Data

Typically, a directional dataset is assumed to be

$$\boldsymbol{X}_1, ..., \boldsymbol{X}_n \overset{i.i.d.}{\sim} f,$$

where f is a directional density supported on

$$\Omega_q := \left\{\boldsymbol{x} \in \mathbb{R}^{q+1} : ||\boldsymbol{x}||_2 = 1\right\}$$

with $\int_{\Omega_q} f(\boldsymbol{x})\, \omega_q(d\boldsymbol{x}) = 1$ and $||\cdot||_2$ is the L_2-norm in $\mathbb{R}^{q+1}$. Here, ω_q is the Lebesgue measure on Ω_q.

When $q = 1$, $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \Omega_1$ can also be represented by angles, e.g., in $[-\pi, \pi]$. In this case, the (circular) density f is 2π-periodic and $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ is called circular data.

von Mises Distribution

One of the most notable circular distributions is the so-called *von Mises* distribution (or circular normal distribution), whose density is

$$f_{vM}(\theta;\mu,\nu) = \frac{1}{2\pi \mathcal{I}_0(\kappa)} \exp\left[\kappa \cos(\theta-\mu)\right],$$

where

von Mises Distribution

One of the most notable circular distributions is the so-called *von Mises* distribution (or circular normal distribution), whose density is

$$f_{vM}(\theta;\mu,\nu) = \frac{1}{2\pi\mathcal{I}_0(\kappa)} \exp\left[\kappa\cos(\theta-\mu)\right],$$

where

- μ is the location parameter,
- κ is a measure of concentration ($\frac{1}{\kappa}$ is analogous to σ^2 in $\mathcal{N}(\mu,\sigma^2)$),
- and $\mathcal{I}_\alpha$ is the modified Bessel function of order α.

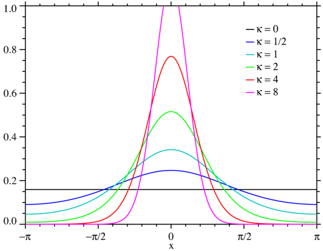


Figure: von Mises densities (cited from Wikipedia (2020))

von Mises-Fisher Distribution

The preceding von Mises density can be generalized to the directional density on Ω_q. It becomes the *von Mises-Fisher* distribution, whose density is given by:

$$f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}, \kappa) = C_q(\kappa) \cdot \exp\left(\kappa \boldsymbol{\mu}^T \boldsymbol{x}\right) \quad \text{with } C_q(\kappa) = \frac{\kappa^{\frac{q-1}{2}}}{(2\pi)^{\frac{q+1}{2}} \mathcal{I}_{\frac{q-1}{2}}(\kappa)}.$$

von Mises-Fisher Distribution

The preceding von Mises density can be generalized to the directional density on Ω_q. It becomes the *von Mises-Fisher* distribution, whose density is given by:

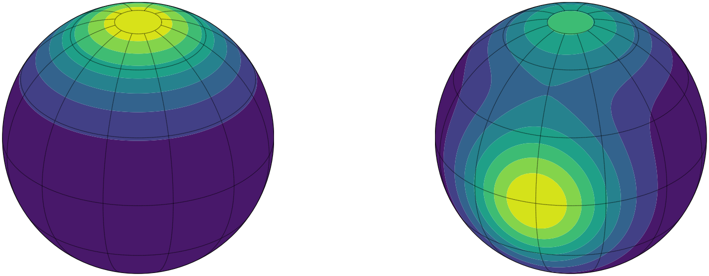
$$f_{\text{vMF}}(\boldsymbol{x};\boldsymbol{\mu},\kappa) = C_q(\kappa)\cdot \exp\left(\kappa\boldsymbol{\mu}^T\boldsymbol{x}\right) \qquad \text{with } C_q(\kappa) = \frac{\kappa^{\frac{q-1}{2}}}{(2\pi)^{\frac{q+1}{2}}\mathcal{I}_{\frac{q-1}{2}}(\kappa)}.$$

It is related to the normal distribution in $\mathbb{R}^{q+1}$ because

$$\mathcal{N}_{q+1}\left(\boldsymbol{x};\boldsymbol{\mu},\frac{1}{\kappa}\boldsymbol{I}_{q+1}\right) \sim \left(\sqrt{\frac{\kappa}{2\pi}}\right)^{q+1} \exp\left(-\frac{\kappa\left|\left|\boldsymbol{x}-\boldsymbol{\mu}\right|\right|^2}{2}\right)$$
$$\propto \exp\left(\kappa\boldsymbol{\mu}^T\boldsymbol{x}\right) \propto f_{vMF}(\boldsymbol{x};\boldsymbol{\mu},\kappa),$$

using the fact that $\left|\left|\boldsymbol{x}-\boldsymbol{\mu}\right|\right|^2 = 2-2\boldsymbol{\mu}^T\boldsymbol{x}$ on Ω_q.

von Mises-Fisher Density and its Mixture



(a) $f_{\text{vMF},2}(\boldsymbol{x};\boldsymbol{\mu},\nu)$ with $\boldsymbol{\mu}=(0,0,1)$ and $\nu=4.0$

(b) $\frac{2}{5}\cdot f_{\text{vMF},2}(\boldsymbol{x};\boldsymbol{\mu}_1,\nu_1)+\frac{3}{5}\cdot f_{\text{vMF},2}(\boldsymbol{x};\boldsymbol{\mu}_2,\nu_2)$ with $\boldsymbol{\mu}_1=(0,0,1)$, $\boldsymbol{\mu}_2=(1,0,0)$, and $\nu_1=\nu_2=5.0$

Figure: Contour plots of a 2-von Mises-Fisher density and a mixture of 2-vMF densities

Statistical Problems with Directional Data

Given a directional random sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \Omega_q$, the following problems are often of research interest:

Statistical Problems with Directional Data

Given a directional random sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \Omega_q$, the following problems are often of research interest:

1. estimating the underlying directional density f (as well as its derivatives), and

Statistical Problems with Directional Data

Given a directional random sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \Omega_q$, the following problems are often of research interest:

1. estimating the underlying directional density f (as well as its derivatives), and
2. identifying the local modes of the (estimated) directional density on Ω_q.

Density Estimation Problem on Ω_q

The first problem can be addressed by the directional kernel density estimator (KDE), which is often written as (Hall et al., 1987; Bai et al., 1988; García-Portugués, 2013):

$$\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right), \tag{1}$$

Density Estimation Problem on Ω_q

The first problem can be addressed by the directional kernel density estimator (KDE), which is often written as (Hall et al., 1987; Bai et al., 1988; García-Portugués, 2013):

$$\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right), \tag{1}$$

where $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \Omega_q \subset \mathbb{R}^{q+1}$ are random directional observations, L is a directional kernel (a rapidly decaying function with nonnegative values on $[0, \infty)$), $h > 0$ is the bandwidth parameter, and $c_{h,q}(L)$ is a normalizing constant satisfying

$$c_{h,q}(L)^{-1} = \int_{\Omega_q} L\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{y}}{h^2}\right) \omega_q(d\boldsymbol{y}) = h^q \lambda_{h,q}(L) \asymp h^q \lambda_q(L) \tag{2}$$

with $\lambda_q(L) = 2^{\frac{q}{2}-1} \omega_{q-1} \int_0^\infty L(r) r^{\frac{q}{2}-1} dr$.

Density Estimation Problem on Ω_q

Using the von Mises kernel $L(r) = e^{-r}$, the directional KDE $\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$ becomes a mixture of von Mises-Fisher densities as follows:

$$\begin{aligned}\widehat{f}_h(\boldsymbol{x}) &= \frac{1}{n}\sum_{i=1}^{n} f_{\text{vMF}}\left(\boldsymbol{x}; \boldsymbol{X}_i, \frac{1}{h^2}\right) \\ &= \frac{1}{n(2\pi)^{\frac{q+1}{2}} \mathcal{I}_{\frac{q-1}{2}}(1/h^2) h^{q-1}} \sum_{i=1}^{n} \exp\left(\frac{\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right).\end{aligned}$$

Mode-Seeking Problem on Ω_q

The second problem is more challenging from two aspects:

Mode-Seeking Problem on Ω_q

The second problem is more challenging from two aspects:

1. How can we generalize the regular mean shift algorithm with Euclidean KDEs (Cheng, 1995; Comaniciu and Meer, 2002) to the directional data setting?

Mode-Seeking Problem on Ω_q

The second problem is more challenging from two aspects:

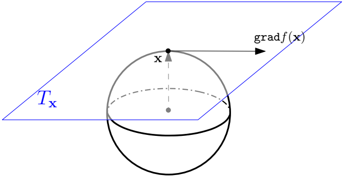
1. How can we generalize the regular mean shift algorithm with Euclidean KDEs (Cheng, 1995; Comaniciu and Meer, 2002) to the directional data setting?
2. How can we conduct the gradient ascent procedure on Ω_q?

Mode-Seeking Problem on Ω_q

The second problem is more challenging from two aspects:

1. How can we generalize the regular mean shift algorithm with Euclidean KDEs (Cheng, 1995; Comaniciu and Meer, 2002) to the directional data setting?
2. How can we conduct the gradient ascent procedure on Ω_q?

(Note that the usual gradient ascent procedure $\boldsymbol{x}^{(t+1)} \leftarrow \boldsymbol{x}^{(t)} + \eta \cdot \mathtt{grad}\, f(\boldsymbol{x}^{(t)})$ is problematic on Ω_q, because Ω_q has a nonzero curvature and the (Riemannian) gradient of f at $\boldsymbol{x} \in \Omega_q$ is defined on the tangent space $T_{\boldsymbol{x}}$. Moving along the gradient direction will deviate from Ω_q.)

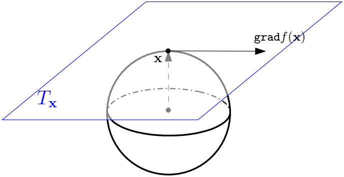


Mode-Seeking Problem on Ω_q

The second problem is more challenging from two aspects:

1. How can we generalize the regular mean shift algorithm with Euclidean KDEs (Cheng, 1995; Comaniciu and Meer, 2002) to the directional data setting?
2. How can we conduct the gradient ascent procedure on Ω_q?

(Note that the usual gradient ascent procedure $\boldsymbol{x}^{(t+1)} \leftarrow \boldsymbol{x}^{(t)} + \eta \cdot \texttt{grad} f(\boldsymbol{x}^{(t)})$ is problematic on Ω_q, because Ω_q has a nonzero curvature and the (Riemannian) gradient of f at $\boldsymbol{x} \in \Omega_q$ is defined on the tangent space $T_{\boldsymbol{x}}$. Moving along the gradient direction will deviate from Ω_q.)



The above two questions are related in our directional mean shift algorithm!

In this talk, we will discuss both computational and statistical learning theory of the mean shift algorithm with directional data.

Highlights of the Talk

In this talk, we will discuss both computational and statistical learning theory of the mean shift algorithm with directional data.

- We will derive our directional mean shift algorithm.

Highlights of the Talk
Mean Shift and Learning Theory with Directional Data

In this talk, we will discuss both computational and statistical learning theory of the mean shift algorithm with directional data.

- We will derive our directional mean shift algorithm.
  - **Ascending property**: $\left\{\widehat{f}_h\left(\widehat{\boldsymbol{y}}_s\right)\right\}_{s=0,1,\dots}$ is monotonically increasing along any mean shift path $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\dots}$ on Ω_q.

Highlights of the Talk

Mean Shift and Learning Theory with Directional Data

In this talk, we will discuss both computational and statistical learning theory of the mean shift algorithm with directional data.

- We will derive our directional mean shift algorithm.
  - **Ascending property**: $\left\{\widehat{f}_h\left(\widehat{\boldsymbol{y}}_s\right)\right\}_{s=0,1,\dots}$ is monotonically increasing along any mean shift path $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\dots}$ on Ω_q.
  - **Algorithmic Convergence**: the mean shift path $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\dots}$ converges (linearly) to an estimated local mode $\widehat{\boldsymbol{m}}_k$ if it is initialized in its small neighborhood.

In this talk, we will discuss both computational and statistical learning theory of the mean shift algorithm with directional data.

- We will derive our directional mean shift algorithm.
    - **Ascending property**: $\left\{\widehat{f}_h\left(\widehat{\boldsymbol{y}}_s\right)\right\}_{s=0,1,\dots}$ is monotonically increasing along any mean shift path $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\dots}$ on Ω_q.
    - **Algorithmic Convergence**: the mean shift path $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\dots}$ converges (linearly) to an estimated local mode $\widehat{\boldsymbol{m}}_k$ if it is initialized in its small neighborhood.
- We will formulate (Riemannian) gradient and Hessian estimators from the directional KDE $\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n}\sum_{i=1}^{n} L\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$ and prove

Highlights of the Talk

In this talk, we will discuss both computational and statistical learning theory of the mean shift algorithm with directional data.

- We will derive our directional mean shift algorithm.
  - **Ascending property**: $\left\{\widehat{f}_h\left(\widehat{\boldsymbol{y}}_s\right)\right\}_{s=0,1,\dots}$ is monotonically increasing along any mean shift path $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\dots}$ on Ω_q.
  - **Algorithmic Convergence**: the mean shift path $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\dots}$ converges (linearly) to an estimated local mode $\widehat{\boldsymbol{m}}_k$ if it is initialized in its small neighborhood.

- We will formulate (Riemannian) gradient and Hessian estimators from the directional KDE $\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n}\sum_{i=1}^{n} L\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$ and prove
  - **Pointwise and uniform consistency** of $\mathtt{grad}\,\widehat{f}_h$ and $\mathcal{H}\widehat{f}_h$.
  - **Mode consistency**: $\mathbb{P}\left(\widehat{K}_n \neq K\right)$ and $\mathtt{Haus}\left(\mathcal{M}, \widehat{\mathcal{M}}_n\right)$ can be arbitrarily small when h is sufficiently small and n is sufficiently large.

Mean Shift Algorithm with Directional Data

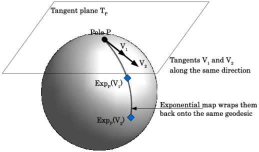
Background Knowledge in Differential Geometry

Definition (Tangent space of Ω_q)

The *tangent space* of the sphere Ω_q at $\boldsymbol{x} \in \Omega_q$ is given by

$$T_{\boldsymbol{x}} \equiv T_{\boldsymbol{x}}(\Omega_q) = \left\{ \boldsymbol{u} - \boldsymbol{x} \in \mathbb{R}^{q+1} : \boldsymbol{x}^T(\boldsymbol{u} - \boldsymbol{x}) = 0 \right\} \simeq \left\{ \boldsymbol{v} \in \mathbb{R}^{q+1} : \boldsymbol{x}^T \boldsymbol{v} = 0 \right\},$$

where $V_1 \simeq V_2$ signifies that the two vector spaces are isomorphic. In what follows, $\boldsymbol{v} \in T_{\boldsymbol{x}}$ indicates that $\boldsymbol{v}$ is a vector tangent to Ω_q at $\boldsymbol{x}$.


Tangent plane T_p
Pole P
V_1
V_2
Tangents V_1 and V_2 along the same direction
$Exp_p(V_1)$
$Exp_p(V_2)$
Exponential map wraps them back onto the same geodesic

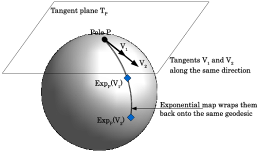
Background Knowledge in Differential Geometry

Definition (Tangent space of Ω_q)

The *tangent space* of the sphere Ω_q at $\boldsymbol{x} \in \Omega_q$ is given by

$$T_{\boldsymbol{x}} \equiv T_{\boldsymbol{x}}(\Omega_q) = \left\{ \boldsymbol{u} - \boldsymbol{x} \in \mathbb{R}^{q+1} : \boldsymbol{x}^T(\boldsymbol{u} - \boldsymbol{x}) = 0 \right\} \simeq \left\{ \boldsymbol{v} \in \mathbb{R}^{q+1} : \boldsymbol{x}^T \boldsymbol{v} = 0 \right\},$$

where $V_1 \simeq V_2$ signifies that the two vector spaces are isomorphic. In what follows, $\boldsymbol{v} \in T_{\boldsymbol{x}}$ indicates that $\boldsymbol{v}$ is a vector tangent to Ω_q at $\boldsymbol{x}$.


Tangent plane T_P
Pole P
V_1
V_2
Tangents V_1 and V_2 along the same direction
$Exp_P(V_1)$
Exponential map wraps them back onto the same geodesic
$Exp_P(V_2)$

Definition (Geodesic)

A *geodesic* on Ω_q is a non-constant, parametrized curve $\alpha : I \to \Omega_q$ of constant speed and (locally) minimum length between two points on Ω_q for some interval $I \subset \mathbb{R}$. It will be part of a great circle on Ω_q.

Riemannian Gradient on Ω_q

Assume that the function f is well-defined and smooth in $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ (or at least in an open neighborhood $U \supset \Omega_q$).

Riemannian Gradient on Ω_q

Assume that the function f is well-defined and smooth in $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ (or at least in an open neighborhood $U \supset \Omega_q$).

Given a geodesic curve $\alpha : (-\epsilon, \epsilon) \to \Omega_q$ with $\alpha(0) = \boldsymbol{x}$ and $\alpha'(0) = \boldsymbol{v}$, the *differential* of f at point $\boldsymbol{x} \in \Omega_q$ (or *total gradient*; *pushforward*)
$df_{\boldsymbol{x}} : T_{\boldsymbol{x}} \to T_{f(\boldsymbol{x})}(\mathbb{R}) \simeq \mathbb{R}$ is given by

$$df_x(\boldsymbol{v}) = \frac{d}{dt} f(\alpha(t))\Big|_{t=0} = \nabla f(\alpha(t))^T \alpha'(t)\Big|_{t=0} = \nabla f(\boldsymbol{x})^T \alpha'(0) = \nabla f(\boldsymbol{x})^T \boldsymbol{v} \qquad (3)$$

for any $\boldsymbol{v} \in T_{\boldsymbol{x}}$.

Riemannian Gradient on Ω_q

Assume that the function f is well-defined and smooth in $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ (or at least in an open neighborhood $U \supset \Omega_q$).

Given a geodesic curve $\alpha : (-\epsilon, \epsilon) \to \Omega_q$ with $\alpha(0) = \boldsymbol{x}$ and $\alpha'(0) = \boldsymbol{v}$, the *differential* of f at point $\boldsymbol{x} \in \Omega_q$ (or *total gradient*; *pushforward*) $df_{\boldsymbol{x}} : T_{\boldsymbol{x}} \to T_{f(\boldsymbol{x})}(\mathbb{R}) \simeq \mathbb{R}$ is given by

$$df_x(\boldsymbol{v}) = \frac{d}{dt} f(\alpha(t))\Big|_{t=0} = \nabla f(\alpha(t))^T \alpha'(t)\Big|_{t=0} = \nabla f(\boldsymbol{x})^T \alpha'(0) = \nabla f(\boldsymbol{x})^T \boldsymbol{v} \qquad (3)$$

for any $\boldsymbol{v} \in T_{\boldsymbol{x}}$.

The *Riemannian gradient* $\mathtt{grad}\, f(\boldsymbol{x}) \in T_{\boldsymbol{x}} \subset \mathbb{R}^{q+1}$ is defined by

$$\langle \mathtt{grad}\, f(\boldsymbol{x}), \boldsymbol{v} \rangle = [\mathtt{grad}\, f(\boldsymbol{x})]^T \boldsymbol{v} = \nabla f(\boldsymbol{x})^T \boldsymbol{v}. \qquad (4)$$

Riemannian Gradient on Ω_q

Assume that the function f is well-defined and smooth in $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ (or at least in an open neighborhood $U \supset \Omega_q$).

Given a geodesic curve $\alpha : (-\epsilon, \epsilon) \to \Omega_q$ with $\alpha(0) = \boldsymbol{x}$ and $\alpha'(0) = \boldsymbol{v}$, the *differential* of f at point $\boldsymbol{x} \in \Omega_q$ (or *total gradient*; *pushforward*) $df_{\boldsymbol{x}} : T_{\boldsymbol{x}} \to T_{f(\boldsymbol{x})}(\mathbb{R}) \simeq \mathbb{R}$ is given by

$$df_x(\boldsymbol{v}) = \frac{d}{dt} f(\alpha(t))\Big|_{t=0} = \nabla f(\alpha(t))^T \alpha'(t)\Big|_{t=0} = \nabla f(\boldsymbol{x})^T \alpha'(0) = \nabla f(\boldsymbol{x})^T \boldsymbol{v} \qquad (3)$$

for any $\boldsymbol{v} \in T_{\boldsymbol{x}}$.

The *Riemannian gradient* $\mathtt{grad}\, f(\boldsymbol{x}) \in T_{\boldsymbol{x}} \subset \mathbb{R}^{q+1}$ is defined by

$$\langle \mathtt{grad}\, f(\boldsymbol{x}), \boldsymbol{v} \rangle = [\mathtt{grad}\, f(\boldsymbol{x})]^T \boldsymbol{v} = \nabla f(\boldsymbol{x})^T \boldsymbol{v}. \qquad (4)$$

Thus, using the fact that $\nabla f(\boldsymbol{x}) = \mathtt{Tang}\left(\nabla f(\boldsymbol{x})\right) + \mathtt{Rad}\left(\nabla f(\boldsymbol{x})\right)$, we conclude that

$$\mathtt{grad}\, f(\boldsymbol{x}) \equiv \mathtt{Tang}\left(\nabla f(\boldsymbol{x})\right) = \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x}).$$

Riemannian Gradient on Ω_q

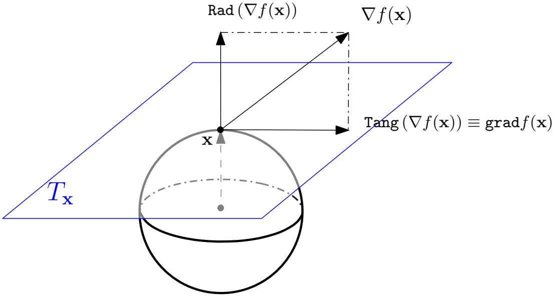
Here,

$$\texttt{Tang}\left(\nabla f(\boldsymbol{x})\right) = \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right)\nabla f(\boldsymbol{x})$$

and

$$\texttt{Rad}\left(\nabla f(\boldsymbol{x})\right) = \nabla f(\boldsymbol{x}) - \texttt{Tang}\left(\nabla f(\boldsymbol{x})\right) = \boldsymbol{x}\boldsymbol{x}^T\nabla f(\boldsymbol{x}),$$

where I_{q+1} is the identity matrix in $\mathbb{R}^{(q+1)\times(q+1)}$.



$\texttt{Rad}\left(\nabla f(\mathbf{x})\right)$
$\nabla f(\mathbf{x})$
$\texttt{Tang}\left(\nabla f(\mathbf{x})\right) \equiv \texttt{grad} f(\mathbf{x})$
$\mathbf{x}$
$T_{\mathbf{x}}$

Riemannian Hessian on Ω_q

The *Riemannian Hessian* of f at $\boldsymbol{x} \in \Omega_q$ is a symmetric bilinear map $\mathcal{H}f(\boldsymbol{x}) : T_{\boldsymbol{x}} \times T_{\boldsymbol{x}} \to \mathbb{R}$ defined as:

$$d^2 f_x(\boldsymbol{v}, \boldsymbol{u}) = \left\langle \texttt{grad} \left\langle \texttt{grad}\, f, \boldsymbol{v} \right\rangle(\boldsymbol{x}), \boldsymbol{u} \right\rangle_x = \boldsymbol{v}^T \mathcal{H}f(\boldsymbol{x}) \boldsymbol{u} \tag{5}$$

for any $\boldsymbol{u}, \boldsymbol{v} \in T_{\boldsymbol{x}}$. It satisfies the property that

$$\mathcal{H}f(\boldsymbol{x}) = (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\mathcal{H}f(\boldsymbol{x}) = \mathcal{H}f(\boldsymbol{x})(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T). \tag{6}$$

Riemannian Hessian on Ω_q

The *Riemannian Hessian* of f at $\boldsymbol{x} \in \Omega_q$ is a symmetric bilinear map $\mathcal{H}f(\boldsymbol{x}) : T_{\boldsymbol{x}} \times T_{\boldsymbol{x}} \to \mathbb{R}$ defined as:

$$d^2 f_x(\boldsymbol{v}, \boldsymbol{u}) = \langle \texttt{grad} \langle \texttt{grad} f, \boldsymbol{v} \rangle (\boldsymbol{x}), \boldsymbol{u} \rangle_x = \boldsymbol{v}^T \mathcal{H}f(\boldsymbol{x}) \boldsymbol{u} \tag{5}$$

for any $\boldsymbol{u}, \boldsymbol{v} \in T_{\boldsymbol{x}}$. It satisfies the property that

$$\mathcal{H}f(\boldsymbol{x}) = (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\mathcal{H}f(\boldsymbol{x}) = \mathcal{H}f(\boldsymbol{x})(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T). \tag{6}$$

Using the same technique as the derivation of Riemannian gradients, we can obtain the explicit form of the Riemannian Hessian of f at $\boldsymbol{x} \in \Omega_q$ as

$$\mathcal{H}f(\boldsymbol{x}) = (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T) \left[\nabla\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T \boldsymbol{x} I_{q+1} \right] (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T). \tag{7}$$

This form coincides with the usual definition of Riemannian Hessians of any smooth function f on (sub)manifolds (Absil et al., 2013).

Review of Euclidean Mean Shift Algorithm

Review of Euclidean Mean Shift Algorithm

The KDE with Euclidean data in $\mathbb{R}^d$ is given by:

$$\widehat{p}_n(\boldsymbol{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} k\left(\left\|\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right\|^2\right), \tag{8}$$

where k is a rapidly decaying kernel profile on $[0, \infty)$, e.g., $k(x) = e^{-\frac{x}{2}}$.

Review of Euclidean Mean Shift Algorithm

The KDE with Euclidean data in $\mathbb{R}^d$ is given by:

$$\widehat{p}_n(\boldsymbol{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right) = \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n} k\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|^2\right), \tag{8}$$

where k is a rapidly decaying kernel profile on $[0,\infty)$, e.g., $k(x) = e^{-\frac{x}{2}}$. The gradient of $\widehat{p}_n(\boldsymbol{x})$ has the following decomposition:

$$\begin{aligned}\nabla\widehat{p}_n(\boldsymbol{x}) &= \frac{2c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n}(\boldsymbol{x}-\boldsymbol{X}_i)\cdot k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)\\ &= \frac{2c_{k,d}}{nh^{d+2}}\left[\sum_{i=1}^{n} -k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)\right]\cdot\left[\frac{\sum_{i=1}^{n}\boldsymbol{X}_i k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}{\sum_{i=1}^{n} k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)} - \boldsymbol{x}\right].\end{aligned}$$

Review of Euclidean Mean Shift Algorithm

$$\nabla \widehat{p}_n(\boldsymbol{x}) = \underbrace{\frac{2c_{k,d}}{nh^{d+2}}\left[\sum_{i=1}^n -k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)\right]}_{\text{term 1}} \underbrace{\left[\frac{\sum_{i=1}^n \boldsymbol{X}_i k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}{\sum_{i=1}^n k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)} - \boldsymbol{x}\right]}_{\text{term 2: mean shift vector}} \tag{9}$$

The “term 1” in (9) is proportional to the density estimate at $\boldsymbol{x}$ with the “kernel” $G(\boldsymbol{x}) = -c_{g,d}k'(||\boldsymbol{x}||_2^2)$ and is generally positive.

Review of Euclidean Mean Shift Algorithm

$$\nabla \widehat{p}_n(\boldsymbol{x}) = \underbrace{\frac{2c_{k,d}}{nh^{d+2}}\left[\sum_{i=1}^n -k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)\right]}_{\text{term 1}} \underbrace{\left[\frac{\sum_{i=1}^n \boldsymbol{X}_i k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}{\sum_{i=1}^n k'\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)} - \boldsymbol{x}\right]}_{\text{term 2: mean shift vector}} \qquad (9)$$

The "term 1" in (9) is proportional to the density estimate at $\boldsymbol{x}$ with the "kernel" $G(\boldsymbol{x}) = -c_{g,d}k'(||\boldsymbol{x}||_2^2)$ and is generally positive.

Thus, the mean shift vector in (9) points toward the direction of maximum increase in $\widehat{p}_n$ and thus yields the following mean shift iteration, which is a valid mode-seeking algorithm (Carreira-Perpiñán, 2015; Chen et al., 2016):

$$\begin{aligned}\boldsymbol{x}^{(t+1)} &\leftarrow \boldsymbol{x}^{(t)} + \frac{\sum_{i=1}^n \boldsymbol{X}_i k'\left(\left|\left|\frac{\boldsymbol{x}^{(t)}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}{\sum_{i=1}^n k'\left(\left|\left|\frac{\boldsymbol{x}^{(t)}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)} - \boldsymbol{x}^{(t)} \\ &= \frac{\sum_{i=1}^n \boldsymbol{X}_i k'\left(\left|\left|\frac{\boldsymbol{x}^{(t)}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}{\sum_{i=1}^n k'\left(\left|\left|\frac{\boldsymbol{x}^{(t)}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}.\end{aligned}$$

Directional Mean Shift Algorithm

Unfortunately, the total gradient of the directional KDE $\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1-\boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right)$ does not have such a clear decomposition as $\nabla \widehat{p}_n(\boldsymbol{x})$ in (9), because

$$\nabla \widehat{f}_h(\boldsymbol{x}) = -\frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right). \qquad (10)$$

Directional Mean Shift Algorithm

Unfortunately, the total gradient of the directional KDE $\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1-\boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right)$ does not have such a clear decomposition as $\nabla \widehat{p}_n(\boldsymbol{x})$ in (9), because

$$\nabla \widehat{f}_h(\boldsymbol{x}) = -\frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right). \tag{10}$$

Fortunately, we have an alternative representation of the directional KDE as:

$$\widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right). \tag{11}$$

Directional Mean Shift Algorithm

Unfortunately, the total gradient of the directional KDE $\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$ does not have such a clear decomposition as $\nabla\widehat{p}_n(\boldsymbol{x})$ in (9), because

$$\nabla\widehat{f}_h(\boldsymbol{x}) = -\frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right). \tag{10}$$

Fortunately, we have an alternative representation of the directional KDE as:

$$\widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right). \tag{11}$$

More importantly, $\widehat{f}_h(\boldsymbol{x}) = \widetilde{f}_h(\boldsymbol{x})$ on Ω_q, since $2 - 2\boldsymbol{x}^T\boldsymbol{X}_i = ||\boldsymbol{x}-\boldsymbol{X}_i||_2^2$.

Directional Mean Shift Algorithm

The power of the expression $\widetilde{f}_h(\boldsymbol{x})$ is that its total gradient has a similar decomposition as $\nabla \widehat{p}_n(\boldsymbol{x})$ (cf. (9)):

$$\begin{aligned}\nabla \widetilde{f}_h(\boldsymbol{x}) &= \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} (\boldsymbol{x} - \boldsymbol{X}_i) \cdot L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right|\right|_2^2\right) \\ &= \underbrace{\frac{c_{h,q}(L)}{nh^2}\left[\sum_{i=1}^{n} -L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right|\right|_2^2\right)\right]}_{\text{term 1}} \cdot \underbrace{\left[\frac{\sum_{i=1}^{n} \boldsymbol{X}_i \cdot L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}{\sum_{i=1}^{n} L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)} - \boldsymbol{x}\right]}_{\text{term 2: mean shift vector}}. \end{aligned} \tag{12}$$

The "term 1" in (12) can be viewed as a proportional form of the directional KDE at $\boldsymbol{x}$ with the "kernel" $G(r) = -L'(r)$.

Directional Mean Shift Algorithm

The "term 2" in (12) is indeed the *mean shift* vector that is parallel to $\nabla \widetilde{f}_h(\boldsymbol{x})$:

$$\Xi_h(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)}{\sum_{i=1}^{n} L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)} - \boldsymbol{x}. \tag{13}$$

Directional Mean Shift Algorithm

The “term 2” in (12) is indeed the *mean shift* vector that is parallel to $\nabla \widetilde{f}_h(\boldsymbol{x})$:

$$\Xi_h(\boldsymbol{x}) = \frac{\sum\limits_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)}{\sum\limits_{i=1}^{n} L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)} - \boldsymbol{x}. \tag{13}$$

Since the total gradient $\nabla \widetilde{f}_h(\boldsymbol{x})$ becomes the Riemannian gradient of $\widetilde{f}_h(\boldsymbol{x}) = \widehat{f}_h(\boldsymbol{x})$ on Ω_q after being projected onto the tangent space $T_{\boldsymbol{x}}$, the mean shift $\Xi_h(\boldsymbol{x})$ will point in the direction of maximum increase in $\widetilde{f}_h(\boldsymbol{x}) = \widehat{f}_h(\boldsymbol{x})$ after being projected onto $T_{\boldsymbol{x}}$.

Directional Mean Shift Algorithm

The entire procedure goes as follows:

1. perform the mean shift iteration as

$$\widehat{\boldsymbol{y}}_{s+1} \leftarrow \Xi_h\left(\widehat{\boldsymbol{y}}_s\right) + \widehat{\boldsymbol{y}}_s$$

   for $s = 0, 1, ...$

Directional Mean Shift Algorithm

The entire procedure goes as follows:

1. perform the mean shift iteration as

$$\widehat{\boldsymbol{y}}_{s+1} \leftarrow \Xi_h\left(\widehat{\boldsymbol{y}}_s\right) + \widehat{\boldsymbol{y}}_s$$

   for $s = 0, 1, ...$

However, due to the manifold structure of Ω_q, translating a point $\widehat{\boldsymbol{y}}_s \in \Omega_q$ in its mean shift direction $\Xi_h\left(\widehat{\boldsymbol{y}}_s\right)$ deviates the point from Ω_q. Thus, we

2. project the translated point $\Xi_h\left(\widehat{\boldsymbol{y}}_s\right) + \widehat{\boldsymbol{y}}_s$ back to Ω_q by a simple standardization $\frac{\Xi_h(\widehat{\boldsymbol{y}}_s)+\widehat{\boldsymbol{y}}_s}{||\Xi_h(\widehat{\boldsymbol{y}}_s)+\widehat{\boldsymbol{y}}_s||_2}$.

Directional Mean Shift As a Fixed-point Iteration

Let $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\dots} \subset \Omega_q$ denote the path of successive points defined by our directional mean shift algorithm, where $\widehat{\boldsymbol{y}}_0$ is the initial point of the iteration.

Directional Mean Shift As a Fixed-point Iteration

Let $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\ldots} \subset \Omega_q$ denote the path of successive points defined by our directional mean shift algorithm, where $\widehat{\boldsymbol{y}}_0$ is the initial point of the iteration.

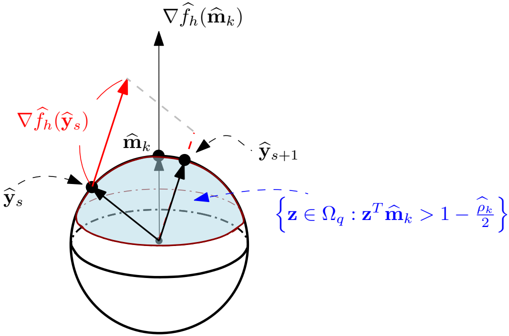
When L is non-increasing (or more specifically, $\sum_{i=1}^{n} L'\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right) < 0$), some simple algebra will show that the entire directional mean shift iteration becomes

$$\widehat{\boldsymbol{y}}_{s+1} = \frac{\Xi_h\left(\widehat{\boldsymbol{y}}_s\right) + \widehat{\boldsymbol{y}}_s}{\left|\left|\Xi_h\left(\widehat{\boldsymbol{y}}_s\right) + \widehat{\boldsymbol{y}}_s\right|\right|_2} = -\frac{\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right)}{\left|\left|\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right)\right|\right|_2}. \tag{14}$$

Directional Mean Shift In Terms of $\nabla\widehat{f}_h$

Recall that $\nabla\widehat{f}_h(\boldsymbol{x}) = -\frac{c_{h,q}(L)}{nh^2}\sum_{i=1}^{n}\boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$. Then the fixed-point equation (14) can be expressed as:

$$\widehat{\boldsymbol{y}}_{s+1} = \frac{\nabla\widehat{f}_h(\widehat{\boldsymbol{y}}_s)}{\left|\left|\nabla\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}. \tag{15}$$


$\nabla\widehat{f}_h(\widehat{\mathbf{m}}_k)$
$\nabla\widehat{f}_h(\widehat{\mathbf{y}}_s)$
$\widehat{\mathbf{m}}_k$
$\widehat{\mathbf{y}}_{s+1}$
$\widehat{\mathbf{y}}_s$
$\left\{\mathbf{z}\in\Omega_q : \mathbf{z}^T\widehat{\mathbf{m}}_k > 1-\frac{\widehat{\rho}_k}{2}\right\}$

Ascending Property of Directional Mean Shift

Let the directional density estimates along the mean shift iteration path be

$$\widehat{f}_h(\widehat{\boldsymbol{y}}_s) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right) \qquad \text{for } s = 0, 1, \ldots.$$

Ascending Property of Directional Mean Shift

Let the directional density estimates along the mean shift iteration path be

$$\widehat{f}_h(\widehat{\boldsymbol{y}}_s) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right) \qquad \text{for } s = 0, 1, \dots .$$

Theorem (Theorem 8 in Zhang and Chen (2020))

If kernel $L : [0,\infty) \to [0,\infty)$ is monotonically decreasing, differentiable, and convex with $L(0) < \infty$, then the sequence $\left\{\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right\}_{s=0,1,\dots}$ is monotonically increasing and thus converges.

The von Mises kernel $L(r) = e^{-r}$ easily satisfies the above requirements.

Proof (Sketched)

It follows from the inequality $L(x_2) - L(x_1) \geq L'(x_1) \cdot (x_2 - x_1)$.

Convergence of Directional Mean Shift

The ascending property of $\left\{\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right\}_{s=0,1,\ldots}$ under our directional mean shift algorithm is **not sufficient** to ensure the convergence of the mode estimate sequence $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\ldots}$ (Li et al., 2007; Aliyari Ghassabeh, 2013, 2015).

Convergence of Directional Mean Shift

The ascending property of $\left\{\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right\}_{s=0,1,\ldots}$ under our directional mean shift algorithm is **not sufficient** to ensure the convergence of the mode estimate sequence $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\ldots}$ (Li et al., 2007; Aliyari Ghassabeh, 2013, 2015).

To derive the convergence of $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\ldots}$, we make the following assumptions on the directional KDE $\widehat{f}_h$.

- **(C1)** The number of local modes of $\widehat{f}_h$ on Ω_q is finite, and the modes are isolated from other critical points.

Convergence of Directional Mean Shift

The ascending property of $\left\{\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right\}_{s=0,1,\ldots}$ under our directional mean shift algorithm is **not sufficient** to ensure the convergence of the mode estimate sequence $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\ldots}$ (Li et al., 2007; Aliyari Ghassabeh, 2013, 2015).

To derive the convergence of $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\ldots}$, we make the following assumptions on the directional KDE $\widehat{f}_h$.

- **(C1)** The number of local modes of $\widehat{f}_h$ on Ω_q is finite, and the modes are isolated from other critical points.
- **(C2)** Given the current values of n and $h > 0$, we assume that $\widehat{\boldsymbol{m}}_k^T \nabla \widehat{f}_h(\widehat{\boldsymbol{m}}_k) \neq 0$ for all $\widehat{\boldsymbol{m}}_k \in \widehat{\mathcal{M}}_n$, that is, $\sum_{i=1}^{n} \widehat{\boldsymbol{m}}_k^T \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{m}}_k^T \boldsymbol{X}_i}{h^2}\right) \neq 0$.

Convergence of Directional Mean Shift

The ascending property of $\left\{\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right\}_{s=0,1,\dots}$ under our directional mean shift algorithm is **not sufficient** to ensure the convergence of the mode estimate sequence $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\dots}$ (Li et al., 2007; Aliyari Ghassabeh, 2013, 2015).

To derive the convergence of $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\dots}$, we make the following assumptions on the directional KDE $\widehat{f}_h$.

- **(C1)** The number of local modes of $\widehat{f}_h$ on Ω_q is finite, and the modes are isolated from other critical points.
- **(C2)** Given the current values of n and $h > 0$, we assume that $\widehat{\boldsymbol{m}}_k^T \nabla \widehat{f}_h(\widehat{\boldsymbol{m}}_k) \neq 0$ for all $\widehat{\boldsymbol{m}}_k \in \widehat{\mathcal{M}}_n$, that is, $\sum_{i=1}^{n} \widehat{\boldsymbol{m}}_k^T \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{m}}_k^T \boldsymbol{X}_i}{h^2}\right) \neq 0$.

Indeed, $\widehat{\boldsymbol{m}}_k^T \nabla \widehat{f}_h(\widehat{\boldsymbol{m}}_k) \to \infty$ as $h \to 0$ and $nh^q \to \infty$. More generally, $\left|\left|\nabla \widehat{f}_h(\boldsymbol{x})\right|\right|_2 \to \infty$ for any fixed $\boldsymbol{x} \in \Omega_q$ as $h \to 0$ and $nh^q \to \infty$; see Lemma 10 in Zhang and Chen (2020).

Convergence of Directional Mean Shift

Theorem (Theorem 11 in Zhang and Chen (2020))

Assume (C1) and (C2) and that kernel L is monotonically increasing, continuously differentiable, convex with $L(0) < 0$*.*
Then, for each local mode $\widehat{\boldsymbol{m}}_k \in \widehat{\mathcal{M}}_n$*, there exists a* $\widehat{r}_k > 0$ *such that* $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\dots}$ *converges to* $\widehat{\boldsymbol{m}}_k$ *whenever* $||\widehat{\boldsymbol{y}}_0 - \widehat{\boldsymbol{m}}_k||_2 < \widehat{r}_k$ *and* $\widehat{\boldsymbol{y}}_0 \in \Omega_q$*.*
Moreover, under some regularity conditions (D1) and (D2') stated in the sequel, there exists a fixed constant $r^* > 0$ *such that* $\mathbb{P}(\widehat{r}_k \geq r^*) \to 1$ *as* $h \to 0$ *and* $nh^q \to \infty$*.*

Convergence of Directional Mean Shift

Theorem (Theorem 11 in Zhang and Chen (2020))

Assume (C1) and (C2) and that kernel L is monotonically increasing, continuously differentiable, convex with $L(0) < 0$.
Then, for each local mode $\widehat{\boldsymbol{m}}_k \in \widehat{\mathcal{M}}_n$, there exists a $\widehat{r}_k > 0$ such that $\{\widehat{\boldsymbol{y}}_s\}_{s=0,1,\ldots}$ converges to $\widehat{\boldsymbol{m}}_k$ whenever $||\widehat{\boldsymbol{y}}_0 - \widehat{\boldsymbol{m}}_k||_2 < \widehat{r}_k$ and $\widehat{\boldsymbol{y}}_0 \in \Omega_q$.
Moreover, under some regularity conditions (D1) and (D2′) stated in the sequel, there exists a fixed constant $r^* > 0$ such that $\mathbb{P}(\widehat{r}_k \geq r^*) \to 1$ as $h \to 0$ and $nh^q \to \infty$.

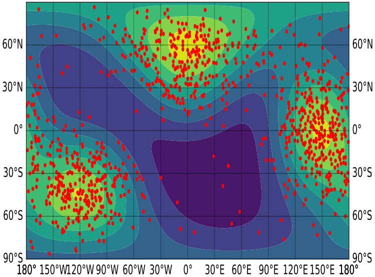
It states that when we initialize our directional mean shift algorithm sufficiently close to an estimated local mode, it will converge to this mode.

Directional Mean Shift: Simulation Study

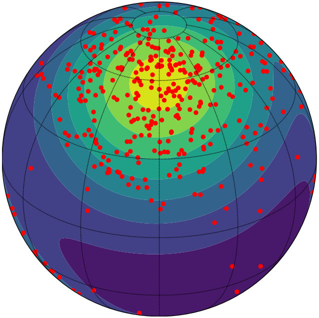
We simulate 1000 data points from the following density

$$f_3(\boldsymbol{x}) = 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_1, \nu_1) + 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_2, \nu_2) + 0.4 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_3, \nu_3)$$

with $\boldsymbol{\mu}_1 = [-120^\circ, -45^\circ]$, $\boldsymbol{\mu}_2 = [0^\circ, 60^\circ]$, $\boldsymbol{\mu}_3 = [150^\circ, 0^\circ]$, and $\nu_1 = \nu_2 = 8$, $\nu_3 = 5$.



(a) Step 0



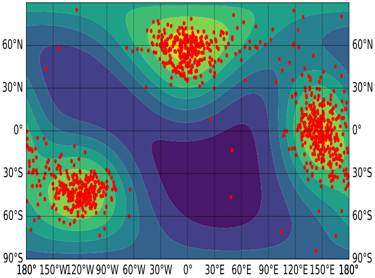
(b) Step 0

Directional Mean Shift: Simulation Study

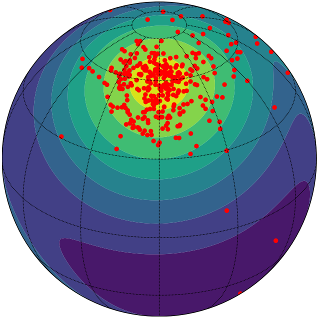
We simulate 1000 data points from the following density

$$f_3(\boldsymbol{x}) = 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_1, \nu_1) + 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_2, \nu_2) + 0.4 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_3, \nu_3)$$

with $\boldsymbol{\mu}_1 = [-120°, -45°]$, $\boldsymbol{\mu}_2 = [0°, 60°]$, $\boldsymbol{\mu}_3 = [150°, 0°]$, and $\nu_1 = \nu_2 = 8$, $\nu_3 = 5$.



(a) Step 1



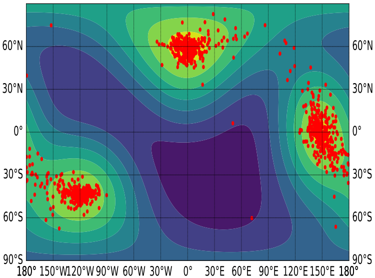
(b) Step 1

Directional Mean Shift: Simulation Study

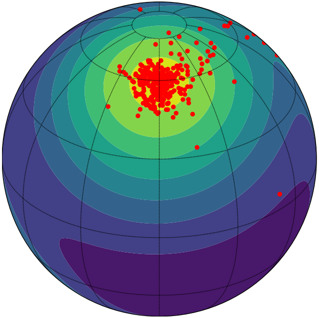
We simulate 1000 data points from the following density

$$f_3(\boldsymbol{x}) = 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_1, \nu_1) + 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_2, \nu_2) + 0.4 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_3, \nu_3)$$

with $\boldsymbol{\mu}_1 = [-120°, -45°]$, $\boldsymbol{\mu}_2 = [0°, 60°]$, $\boldsymbol{\mu}_3 = [150°, 0°]$, and $\nu_1 = \nu_2 = 8$, $\nu_3 = 5$.



(a) Step 2



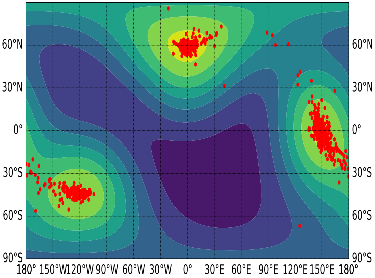
(b) Step 2

Directional Mean Shift: Simulation Study

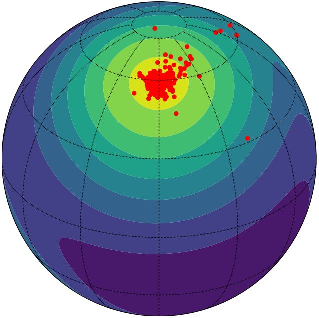
We simulate 1000 data points from the following density

$$f_3(\boldsymbol{x}) = 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_1, \nu_1) + 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_2, \nu_2) + 0.4 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_3, \nu_3)$$

with $\boldsymbol{\mu}_1 = [-120^\circ, -45^\circ]$, $\boldsymbol{\mu}_2 = [0^\circ, 60^\circ]$, $\boldsymbol{\mu}_3 = [150^\circ, 0^\circ]$, and $\nu_1 = \nu_2 = 8$, $\nu_3 = 5$.



(a) Step 3



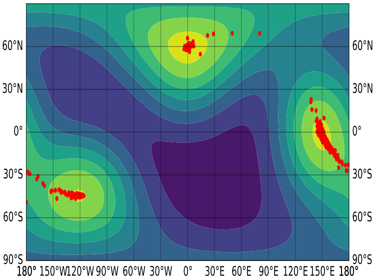
(b) Step 3

Directional Mean Shift: Simulation Study

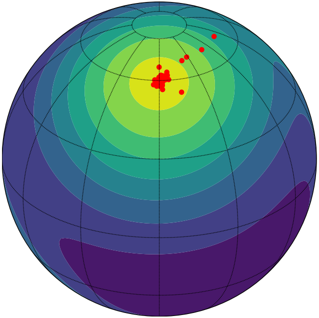
We simulate 1000 data points from the following density

$$f_3(\boldsymbol{x}) = 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_1, \nu_1) + 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_2, \nu_2) + 0.4 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_3, \nu_3)$$

with $\boldsymbol{\mu}_1 = [-120^\circ, -45^\circ]$, $\boldsymbol{\mu}_2 = [0^\circ, 60^\circ]$, $\boldsymbol{\mu}_3 = [150^\circ, 0^\circ]$, and $\nu_1 = \nu_2 = 8$, $\nu_3 = 5$.



(a) Step 5



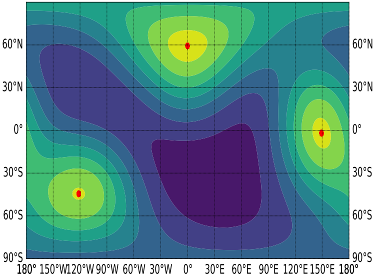
(b) Step 5

Directional Mean Shift: Simulation Study

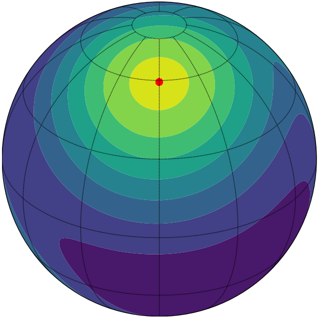
We simulate 1000 data points from the following density

$$f_3(\boldsymbol{x}) = 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_1, \nu_1) + 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_2, \nu_2) + 0.4 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_3, \nu_3)$$

with $\boldsymbol{\mu}_1 = [-120°, -45°]$, $\boldsymbol{\mu}_2 = [0°, 60°]$, $\boldsymbol{\mu}_3 = [150°, 0°]$, and $\nu_1 = \nu_2 = 8$, $\nu_3 = 5$.



(a) Step 22 (converged)



(b) Step 22 (converged)

Directional Mean Shift Clustering

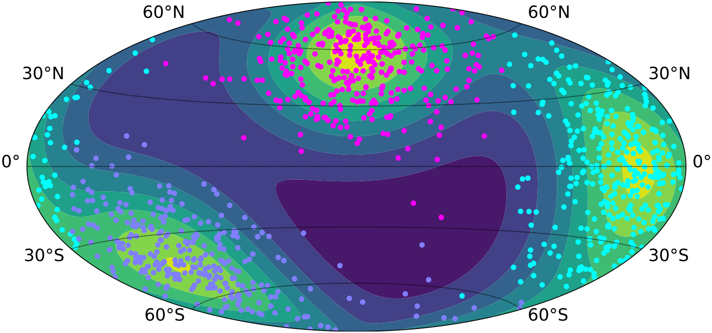


Figure: Mode clustering (Hammer projection view)

Statistical Learning Theory of Directional KDE and its Derivatives

Riemannian Gradient Estimator

Recall that the total gradient of $\widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)$ is

$$\nabla\widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} (\boldsymbol{x}-\boldsymbol{X}_i) \cdot L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right),$$

Riemannian Gradient Estimator

Recall that the total gradient of $\widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)$ is

$$\nabla \widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} (\boldsymbol{x} - \boldsymbol{X}_i) \cdot L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right),$$

while the total gradient of $\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$ is

$$\nabla \widehat{f}_h(\boldsymbol{x}) = -\frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right).$$

They are indeed different on Ω_q!

Riemannian Gradient Estimator

Recall that the total gradient of $\widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)$ is

$$\nabla \widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} (\boldsymbol{x} - \boldsymbol{X}_i) \cdot L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right),$$

while the total gradient of $\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$ is

$$\nabla \widehat{f}_h(\boldsymbol{x}) = -\frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right).$$

They are indeed different on Ω_q! However, their difference lies in the radial direction $\boldsymbol{x} \in \Omega_q$.

Riemannian Gradient Estimator

Recall that the total gradient of $\widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)$ is

$$\nabla\widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} (\boldsymbol{x}-\boldsymbol{X}_i) \cdot L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right),$$

while the total gradient of $\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$ is

$$\nabla\widehat{f}_h(\boldsymbol{x}) = -\frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right).$$

They are indeed different on Ω_q! However, their difference lies in the radial direction $\boldsymbol{x} \in \Omega_q$. Given kernel L, the Riemannian gradient estimators derived from $\nabla\widetilde{f}_h(\boldsymbol{x})$ and $\nabla\widehat{f}_h(\boldsymbol{x})$ are **the same**, i.e.,

$$\begin{aligned}\texttt{grad}\,\widehat{f}_h(\boldsymbol{x}) \equiv \texttt{Tang}\left(\nabla\widehat{f}_h(\boldsymbol{x})\right) &= \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \left(\boldsymbol{x}^T\boldsymbol{X}_i \cdot \boldsymbol{x} - \boldsymbol{X}_i\right) \cdot L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right) \\ &= \texttt{Tang}\left(\nabla\widetilde{f}_h(\boldsymbol{x})\right) \equiv \texttt{grad}\,\widetilde{f}_h(\boldsymbol{x}).\end{aligned} \tag{16}$$

Riemannian Hessian Estimator

Recall equation (7) that the Riemannian Hessian of f at $\boldsymbol{x} \in \Omega_q$ is

$$\mathcal{H}f(\boldsymbol{x}) = (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\left[\nabla\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T \boldsymbol{x} I_{q+1}\right](I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T).$$

Therefore, the Riemannian Hessian estimator of directional KDE $\widehat{f}_h$ is given by

$$\mathcal{H}\widehat{f}_h(\boldsymbol{x}) = (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\left[\nabla\nabla\widehat{f}_h(\boldsymbol{x}) - \nabla\widehat{f}_h(\boldsymbol{x})^T \boldsymbol{x} I_{q+1}\right](I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T) \qquad (17)$$

and similarly for $\mathcal{H}\widetilde{f}_h(\boldsymbol{x})$.

Riemannian Hessian Estimator

Recall equation (7) that the Riemannian Hessian of f at $\boldsymbol{x} \in \Omega_q$ is

$$\mathcal{H}f(\boldsymbol{x}) = (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\left[\nabla\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T \boldsymbol{x} I_{q+1}\right](I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T).$$

Therefore, the Riemannian Hessian estimator of directional KDE $\widehat{f}_h$ is given by

$$\mathcal{H}\widehat{f}_h(\boldsymbol{x}) = (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\left[\nabla\nabla\widehat{f}_h(\boldsymbol{x}) - \nabla\widehat{f}_h(\boldsymbol{x})^T \boldsymbol{x} I_{q+1}\right](I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T) \tag{17}$$

and similarly for $\mathcal{H}\widetilde{f}_h(\boldsymbol{x})$.

Lemma (Lemma 1 in Zhang and Chen (2020))

Assume that kernel L is twice continuously differentiable. Then,

$$\mathcal{H}\widetilde{f}_h(\boldsymbol{x}) = \mathcal{H}\widehat{f}_h(\boldsymbol{x})$$

for any point $\boldsymbol{x} \in \Omega_q$.

Assumptions

We extend the underlying directional density f from Ω_q to $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ by defining $f(\boldsymbol{x}) \equiv f\left(\frac{\boldsymbol{x}}{||\boldsymbol{x}||_2}\right)$ for all $\boldsymbol{x} \in \mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$. We also assume that the total gradient $\nabla f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, ..., \frac{\partial f(\boldsymbol{x})}{\partial x_{q+1}}\right)^T$ and total Hessian matrix $\nabla\nabla f(\boldsymbol{x}) = \left(\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}\right)_{1\leq i,j\leq q+1}$ exist, and are continuous on $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ and square integrable on Ω_q.[1] Consider the following assumptions:

[1] Note that the Riemannian gradient and Hessian are invariant under this extension.

Assumptions

We extend the underlying directional density f from Ω_q to $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ by defining $f(\boldsymbol{x}) \equiv f\left(\frac{\boldsymbol{x}}{||\boldsymbol{x}||_2}\right)$ for all $\boldsymbol{x} \in \mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$. We also assume that the total gradient $\nabla f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, ..., \frac{\partial f(\boldsymbol{x})}{\partial x_{q+1}}\right)^T$ and total Hessian matrix $\nabla\nabla f(\boldsymbol{x}) = \left(\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}\right)_{1 \leq i,j \leq q+1}$ exist, and are continuous on $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ and square integrable on Ω_q.[1] Consider the following assumptions:

- **(D1)** The extended density function f is at least three times continuously differentiable and that the derivatives are square integrable on Ω_q.
- **(D2)** $L : [0, \infty) \to [0, \infty)$ is a bounded and Riemann integrable function such that

$$0 < \int_0^\infty L^k(r) r^{\frac{q}{2}-1} dr < \infty \quad \text{for all } q \geq 1 \text{ and } k = 1, 2.$$

[1] Note that the Riemannian gradient and Hessian are invariant under this extension.

Assumptions (Continued)

- **(D2′)** Under (D2), we further assume that L is a twice continuously differentiable function on $(-\delta_L, \infty) \subset \mathbb{R}$ for some constant $\delta_L > 0$ such that

$$0 < \int_0^\infty L'(r)^k r^{\frac{q}{2}-1} dr < \infty, \quad 0 < \int_0^\infty L''(r)^k r^{\frac{q}{2}-1} dr < \infty$$

for all $q \geq 1$ and $k = 1, 2$.

Pointwise and Uniform Consistency of Directional KDE

Under conditions (D1) and (D2), the convergence rate of $\widehat{f}_h$ is (Hall et al., 1987; Zhao and Wu, 2001; García-Portugués, 2013; García-Portugués et al., 2013)

$$\widehat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x}) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^q}}\right).$$

Moreover, Bai et al. (1988) improved the result to the uniform convergence rate as:

$$\|\widehat{f}_h - f\|_\infty = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^q}}\right), \tag{18}$$

where $\|g\|_\infty = \sup_{\boldsymbol{x}\in\Omega_q} |g(\boldsymbol{x})|$.

Pointwise Consistency of Gradient and Hessian Estimators

Theorem (Theorem 2 in Zhang and Chen (2020))

Assume conditions (D1) and (D2′). For any fixed $\boldsymbol{x} \in \Omega_q$, we have

$$\mathtt{grad}\,\widehat{f}_h(\boldsymbol{x}) - \mathtt{grad}\,f(\boldsymbol{x}) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+2}}}\right)$$

as $h \to 0$ and $nh^{q+2} \to \infty$.
Under the same conditions, for any fixed $\boldsymbol{x} \in \Omega_q$, we have

$$\mathcal{H}\widehat{f}_h(\boldsymbol{x}) - \mathcal{H}f(\boldsymbol{x}) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+4}}}\right)$$

as $h \to 0$ and $nh^{q+4} \to \infty$.

Theorem (Theorem 2 in Zhang and Chen (2020))

Assume conditions (D1) and (D2'). For any fixed $\boldsymbol{x} \in \Omega_q$*, we have*

$$\mathtt{grad}\,\widehat{f}_h(\boldsymbol{x}) - \mathtt{grad}\,f(\boldsymbol{x}) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+2}}}\right)$$

as $h \to 0$ *and* $nh^{q+2} \to \infty$*.*
Under the same conditions, for any fixed $\boldsymbol{x} \in \Omega_q$*, we have*

$$\mathcal{H}\widehat{f}_h(\boldsymbol{x}) - \mathcal{H}f(\boldsymbol{x}) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+4}}}\right)$$

as $h \to 0$ *and* $nh^{q+4} \to \infty$*.*

It demonstrates that the Riemannian gradient of a directional KDE $\widehat{f}_h$ is a (pointwise) consistent estimator of the Riemannian gradient of the directional density f that generates data. A similar result holds for the Hessian.

Uniform Consistency of Gradient and Hessian Estimators

Theorem (Theorem 4 in Zhang and Chen (2020))

Assume (D1), (D2'), and the kernel condition in Giné and Guillou (2002). The uniform convergence rate of $\mathtt{grad}\,\widehat{f}_h(\boldsymbol{x})$ *on* Ω_q *is*

$$\sup_{\boldsymbol{x}\in\Omega_q}\left|\left|\mathtt{grad}\,\widehat{f}_h(\boldsymbol{x})-\mathtt{grad}\,f(\boldsymbol{x})\right|\right|_{\max}=O(h^2)+O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2}}}\right),$$

as $h\to 0$ *and* $\frac{nh^{q+2}}{|\log h|}\to\infty$.
Furthermore, the uniform convergence rate of $\mathcal{H}\widehat{f}_h(\boldsymbol{x})$ *on* Ω_q *is*

$$\sup_{\boldsymbol{x}\in\Omega_q}\left|\left|\mathcal{H}\widehat{f}_h(\boldsymbol{x})-\mathcal{H}f(\boldsymbol{x})\right|\right|_{\max}=O(h^2)+O_P\left(\sqrt{\frac{|\log h|}{nh^{q+4}}}\right),$$

as $h\to 0$ *and* $\frac{nh^{q+4}}{|\log h|}\to\infty$*, where* $||A||_{\max}$ *is the elementwise maximum norm for a matrix* $A\in\mathbb{R}^{(q+1)\times(q+1)}$.

Mode Consistency

The mean shift algorithm with directional data converges to an estimated local mode $\widehat{\boldsymbol{m}}_k$ if we initialize the algorithm with its small neighborhood.

Mode Consistency

The mean shift algorithm with directional data converges to an estimated local mode $\widehat{\boldsymbol{m}}_k$ if we initialize the algorithm with its small neighborhood.

Question: how close the collection of estimated local modes $\widehat{\mathcal{M}}_n = \left\{ \widehat{\boldsymbol{m}}_1, ..., \widehat{\boldsymbol{m}}_{\widehat{K}_n} \right\}$ is to the collection of true local modes $\mathcal{M} = \{\boldsymbol{m}_1, ..., \boldsymbol{m}_K\}$?

Mode Consistency

The mean shift algorithm with directional data converges to an estimated local mode $\widehat{\boldsymbol{m}}_k$ if we initialize the algorithm with its small neighborhood.

Question: how close the collection of estimated local modes $\widehat{\mathcal{M}}_n = \left\{\widehat{\boldsymbol{m}}_1, ..., \widehat{\boldsymbol{m}}_{\widehat{K}_n}\right\}$ is to the collection of true local modes $\mathcal{M} = \{\boldsymbol{m}_1, ..., \boldsymbol{m}_K\}$?

Given two sets $A, B \subset \Omega_q$, their Hausdorff distance is

$$\texttt{Haus}(A, B) = \inf\left\{r > 0 : A \subset B \oplus r, B \subset A \oplus r\right\},$$

where $A \oplus r = \left\{\boldsymbol{y} \in \Omega_q : \inf_{\boldsymbol{x} \in A} ||\boldsymbol{x} - \boldsymbol{y}||_2 \leq r\right\} = \left\{\boldsymbol{y} \in \Omega_q : \sup_{\boldsymbol{x} \in A} \boldsymbol{x}^T\boldsymbol{y} \geq 1 - \frac{r^2}{2}\right\}$.

Mode Consistency

Theorem (Theorem 6 in Zhang and Chen (2020))

Assume (D1), (D2′), the kernel condition in Giné and Guillou (2002), and other regularity conditions (Chen et al., 2016; Zhang and Chen, 2020). For any $0 < \delta < 1$*, when* h *is sufficiently small and* n *is sufficiently large,*

1. *there exist some constants* $A_3, B_3 > 0$ *such that*
$$\mathbb{P}\left(\widehat{K}_n \neq K\right) \leq B_3 e^{-A_3 n h^{q+4}}.$$
2. *the Hausdorff distance between the collection of local modes and its estimator satisfies*
$$\texttt{Haus}\left(\mathcal{M}, \widehat{\mathcal{M}}_n\right) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+2}}}\right),$$
as $h \to 0$ *and* $nh^{q+2} \to \infty$*.*

Computational Learning Theory of Directional Mean Shift Algorithm

Linear Convergence of Directional Mean Shift

We proved that the directional mean shift algorithm converges within a small neighborhood of $\widehat{\mathcal{M}}_n$.

Question: how fast it converges?

Linear Convergence of Directional Mean Shift

We proved that the directional mean shift algorithm converges within a small neighborhood of $\widehat{\mathcal{M}}_n$.

Question: how fast it converges?

We will show that our directional mean shift algorithm is a gradient ascent algorithm on Ω_q with an adaptive step size, which can be sufficiently small when h is small, and thus can converge linearly.

Definition (Linear Convergence)

Given a sequence $\{\boldsymbol{y}_s\}_{s=0,1,..}$ converging to $\boldsymbol{m}_k \in \mathcal{M}$, the convergence is said to be *linear* if there exists a positive constant $\Upsilon < 1$ (rate of convergence) such that $||\boldsymbol{y}_{s+1} - \boldsymbol{m}_k|| \leq \Upsilon ||\boldsymbol{y}_s - \boldsymbol{m}_k||$ when s is sufficiently large (Boyd and Vandenberghe, 2004).

Linear Convergence of Directional Mean Shift

However, Ω_q is not a conventional Euclidean space but a Riemannian manifold! Therefore, the gradient ascent update is different:

$$\boldsymbol{y}_{s+1} = \mathtt{Exp}_{\boldsymbol{y}_s}\left(\eta \cdot \mathtt{grad} f(\boldsymbol{y}_s)\right), \tag{19}$$

where $\eta > 0$ is the step size.

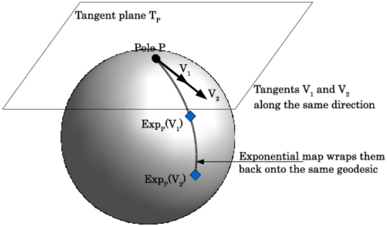
Linear Convergence of Directional Mean Shift

However, Ω_q is not a conventional Euclidean space but a Riemannian manifold! Therefore, the gradient ascent update is different:

$$\boldsymbol{y}_{s+1} = \texttt{Exp}_{\boldsymbol{y}_s}\left(\eta \cdot \texttt{grad} f(\boldsymbol{y}_s)\right), \tag{19}$$

where $\eta > 0$ is the step size.

An *exponential map* at $\boldsymbol{x} \in \Omega_q$ is a mapping $\texttt{Exp}_{\boldsymbol{x}} : T_{\boldsymbol{x}} \to \Omega_q$ such that vector $\boldsymbol{v} \in T_{\boldsymbol{x}}$ is mapped to point $\boldsymbol{y} := \texttt{Exp}_{\boldsymbol{x}}(\boldsymbol{v}) \in \Omega_q$ with $\gamma(0) = \boldsymbol{x}, \gamma(1) = \boldsymbol{y}$ and $\gamma'(0) = \boldsymbol{v}$, where $\gamma : [0,1] \to \Omega_q$ is a geodesic.


Tangent plane T_P
Pole P
V_1
V_2
Tangents V_1 and V_2 along the same direction
$Exp_P(V_1)$
Exponential map wraps them back onto the same geodesic
$Exp_P(V_2)$

Linear Convergence of Gradient Ascent on Ω_q

We prove, under some regularity conditions, that (Theorem 12 in Zhang and Chen (2020))

1. **Linear convergence of gradient ascent with** f: There exists an $r_0 > 0$ such that when the step size $\eta > 0$ is sufficiently small and the initial point $\boldsymbol{y}_0 \in \left\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 < r_0\right\}$ for some $\boldsymbol{m}_k \in \mathcal{M}$,

$$d(\boldsymbol{y}_s, \boldsymbol{m}_k) \leq \Upsilon^s \cdot d(\boldsymbol{y}_0, \boldsymbol{m}_k) \quad \text{with} \quad \Upsilon = \sqrt{1 - \frac{\eta\lambda_*}{2}},$$

where $d(\boldsymbol{p}, \boldsymbol{q}) = \left|\left|\mathtt{Exp}_{\boldsymbol{p}}^{-1}(\boldsymbol{q})\right|\right|_2$.

Linear Convergence of Gradient Ascent on Ω_q

We prove, under some regularity conditions, that (Theorem 12 in Zhang and Chen (2020))

1. **Linear convergence of gradient ascent with** f: There exists an $r_0 > 0$ such that when the step size $\eta > 0$ is sufficiently small and the initial point $\boldsymbol{y}_0 \in \left\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 < r_0\right\}$ for some $\boldsymbol{m}_k \in \mathcal{M}$,

$$d(\boldsymbol{y}_s, \boldsymbol{m}_k) \leq \Upsilon^s \cdot d(\boldsymbol{y}_0, \boldsymbol{m}_k) \quad \text{with} \quad \Upsilon = \sqrt{1 - \frac{\eta\lambda_*}{2}},$$

   where $d(\boldsymbol{p}, \boldsymbol{q}) = \left|\left|\mathtt{Exp}_{\boldsymbol{p}}^{-1}(\boldsymbol{q})\right|\right|_2$.

2. **Linear convergence of gradient ascent with** $\widehat{f}_h$: let the gradient ascent update on Ω_q be

$$\widehat{\boldsymbol{y}}_{s+1} = \mathtt{Exp}_{\boldsymbol{y}_s}\left(\eta \cdot \mathtt{grad}\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right).$$

   When the step size $\eta > 0$ is sufficiently small and the initial point $\widehat{\boldsymbol{y}}_0 \in \left\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 < r_0\right\}$ for some $\boldsymbol{m}_k \in \mathcal{M}$,

$$d\left(\widehat{\boldsymbol{y}}_s, \boldsymbol{m}_k\right) \leq \Upsilon^s \cdot d\left(\widehat{\boldsymbol{y}}_0, \boldsymbol{m}_k\right) + O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2}}}\right)$$

   with probability tending to 1, as $h \to 0$ and $\frac{nh^{q+2}}{|\log h|} \to \infty$.

Linear Convergence of Directional Mean Shift

What is the adaptive step size of our directional mean shift algorithm when viewed as a gradient ascent method on Ω_q?

Linear Convergence of Directional Mean Shift

What is the adaptive step size of our directional mean shift algorithm when viewed as a gradient ascent method on Ω_q?

Recall that the one-step iteration of our directional mean shift is

$$\widehat{\boldsymbol{y}}_{s+1} = \frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)}{\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}.$$

Then, the geodesic distance between $\widehat{\boldsymbol{y}}_{s+1}$ and $\widehat{\boldsymbol{y}}_s$ is

$$\arccos\left(\frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)^T \widehat{\boldsymbol{y}}_s}{\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}\right).$$

Linear Convergence of Directional Mean Shift

If we want $\widehat{\boldsymbol{y}}_{s+1} = \mathtt{Exp}_{\widehat{\boldsymbol{y}}_s}\left(\widehat{\eta}_s \cdot \mathtt{grad}\,\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right)$, then

$$\left|\left|\widehat{\eta}_s \cdot \mathtt{grad}\,\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2 = \arccos\left(\frac{\nabla\widehat{f}_h(\widehat{\boldsymbol{y}}_s)^T\widehat{\boldsymbol{y}}_s}{\left|\left|\nabla\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}\right).$$

Linear Convergence of Directional Mean Shift

If we want $\widehat{\boldsymbol{y}}_{s+1} = \mathtt{Exp}_{\widehat{\boldsymbol{y}}_s}\left(\widehat{\eta}_s \cdot \mathtt{grad}\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right)$, then

$$\left|\left|\widehat{\eta}_s \cdot \mathtt{grad}\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2 = \arccos\left(\frac{\nabla\widehat{f}_h(\widehat{\boldsymbol{y}}_s)^T\widehat{\boldsymbol{y}}_s}{\left|\left|\nabla\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}\right).$$

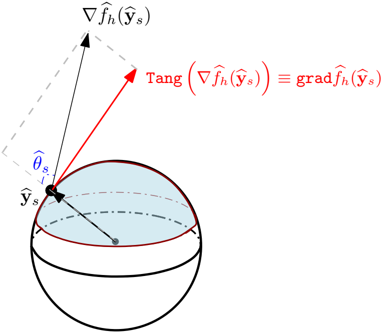
This shows that the adaptive step size is

$$\widehat{\eta}_s = \arccos\left(\frac{\nabla\widehat{f}_h(\widehat{\boldsymbol{y}}_s)^T\widehat{\boldsymbol{y}}_s}{\left|\left|\nabla\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}\right) \cdot \frac{1}{\left|\left|\mathtt{grad}\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}.$$

Linear Convergence of Directional Mean Shift

Denote the angle between $\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)$ and $\widehat{\boldsymbol{y}}_s$ by $\widehat{\theta}_s$. Then,

$$\widehat{\eta}_s = \arccos\left(\frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)^T \widehat{\boldsymbol{y}}_s}{\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}\right) \cdot \frac{1}{\left|\left|\mathtt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2} = \frac{\widehat{\theta}_s}{\left(\sin \widehat{\theta}_s\right) \cdot \left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}.$$



$\nabla \widehat{f}_h(\widehat{\mathbf{y}}_s)$
$\mathtt{Tang}\left(\nabla \widehat{f}_h(\widehat{\mathbf{y}}_s)\right) \equiv \mathtt{grad}\widehat{f}_h(\widehat{\mathbf{y}}_s)$
$\widehat{\theta}_s$
$\widehat{\mathbf{y}}_s$

Linear Convergence of Directional Mean Shift

Recall that

$$\widehat{\eta}_s = \frac{\widehat{\theta}_s}{\left(\sin \widehat{\theta}_s\right) \cdot \left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}.$$

As our directional mean shift algorithm approaches a local mode of $\widehat{f}_h$, $\widehat{\theta}_s \to 0$ and $\frac{\widehat{\theta}_s}{\sin \widehat{\theta}_s} \to 1$. Thus, $\widehat{\eta}_s$ is essentially controlled by $\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2$.

Linear Convergence of Directional Mean Shift

Recall that

$$\widehat{\eta}_s = \frac{\widehat{\theta}_s}{\left(\sin \widehat{\theta}_s\right) \cdot \left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}.$$

As our directional mean shift algorithm approaches a local mode of $\widehat{f}_h$, $\widehat{\theta}_s \to 0$ and $\frac{\widehat{\theta}_s}{\sin \widehat{\theta}_s} \to 1$. Thus, $\widehat{\eta}_s$ is essentially controlled by $\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2$.

- The larger $\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2$ at step s, the shorter the step size $\widehat{\eta}_s$.

Linear Convergence of Directional Mean Shift

Recall that

$$\widehat{\eta}_s = \frac{\widehat{\theta}_s}{\left(\sin \widehat{\theta}_s\right) \cdot \left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}.$$

As our directional mean shift algorithm approaches a local mode of $\widehat{f}_h$, $\widehat{\theta}_s \to 0$ and $\frac{\widehat{\theta}_s}{\sin \widehat{\theta}_s} \to 1$. Thus, $\widehat{\eta}_s$ is essentially controlled by $\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2$.

- The larger $\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2$ at step s, the shorter the step size $\widehat{\eta}_s$.
- Lemma 10 in Zhang and Chen (2020) shows that $\left|\left|\nabla \widehat{f}_h(\boldsymbol{x})\right|\right|_2 \to \infty$ for any $\boldsymbol{x} \in \Omega_q$ as $h \to 0$ and $nh^q \to \infty$.

Linear Convergence of Directional Mean Shift

Recall that

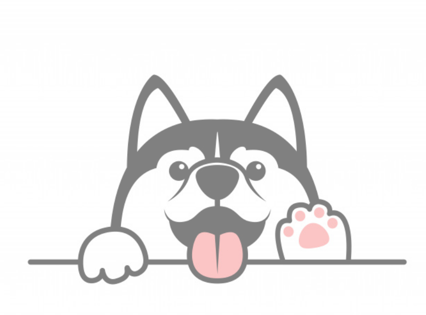
$$\widehat{\eta}_s = \frac{\widehat{\theta}_s}{\left(\sin \widehat{\theta}_s\right) \cdot \left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}.$$

As our directional mean shift algorithm approaches a local mode of $\widehat{f}_h$, $\widehat{\theta}_s \to 0$ and $\frac{\widehat{\theta}_s}{\sin \widehat{\theta}_s} \to 1$. Thus, $\widehat{\eta}_s$ is essentially controlled by $\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2$.

- The larger $\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2$ at step s, the shorter the step size $\widehat{\eta}_s$.
- Lemma 10 in Zhang and Chen (2020) shows that $\left|\left|\nabla \widehat{f}_h(\boldsymbol{x})\right|\right|_2 \to \infty$ for any $\boldsymbol{x} \in \Omega_q$ as $h \to 0$ and $nh^q \to \infty$.

Therefore, one can always select a small bandwidth parameter h such that $\widehat{\eta}_s$ lies within the feasible range for linear convergence.

Real-World Applications



Earthquakes on Earth

Earthquakes on Earth tend to occur more frequently in some regions than others.

Earthquakes on Earth

Earthquakes on Earth tend to occur more frequently in some regions than others.

We analyzed earthquakes with magnitudes of 2.5+ occurring between 2020-08-21 00:00:00 UTC and 2020-09-21 23:59:59 UTC around the world.

Earthquakes on Earth

Earthquakes on Earth tend to occur more frequently in some regions than others.

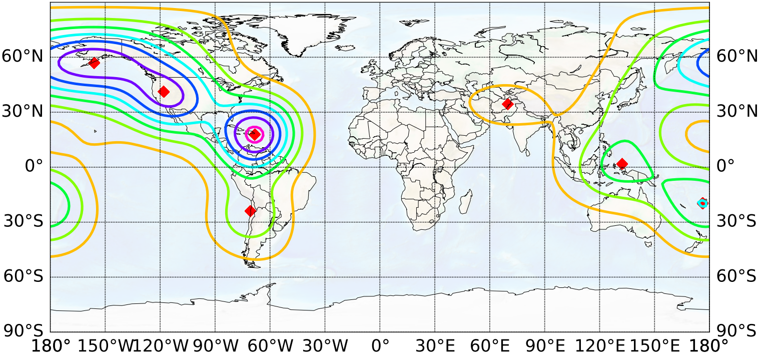
We analyzed earthquakes with magnitudes of 2.5+ occurring between 2020-08-21 00:00:00 UTC and 2020-09-21 23:59:59 UTC around the world.

The bandwidth parameter is selected via the rule of thumb in Proposition 2 in García-Portugués (2013):

$$h_{\text{ROT}} = \left[\frac{4\pi^{\frac{1}{2}} \mathcal{I}_{\frac{q-1}{2}}(\widehat{\nu})^2}{\widehat{\nu}^{\frac{q+1}{2}} \left[2q \cdot \mathcal{I}_{\frac{q+1}{2}}(2\widehat{\nu}) + (q+2)\widehat{\nu} \cdot \mathcal{I}_{\frac{q+3}{2}}(2\widehat{\nu}) \right] n} \right]^{\frac{1}{q+4}} \tag{20}$$

and the tolerance level is $\epsilon = 10^{-7}$.

Earthquakes on Earth



The earthquake modes are located near (from left to right and top to bottom) the Gulf of Alaska, the west side of the Rocky Mountain in Nevada, the Caribbean Sea, the west side of the Andes mountains in Chile, the Middle East, Indonesia, and Fiji.

Earthquakes on Earth

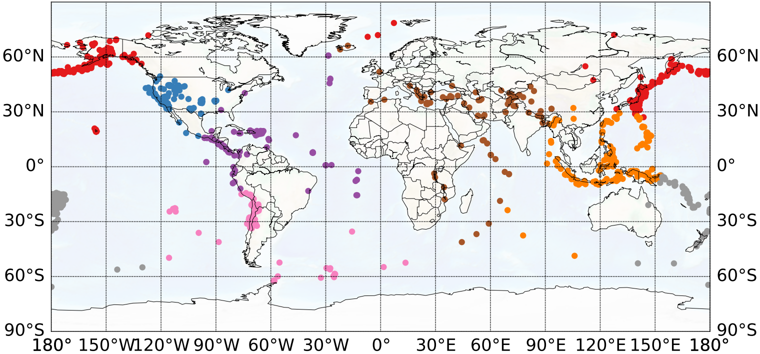
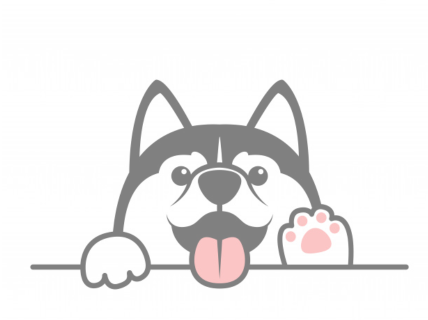


Figure: Mode clustering of Earthquakes on the world map

Conclusion and Future Directions



Conclusion and Future Directions

In this talk, we generalized the standard mean shift algorithm to directional data, and developed statistical and computational learning theory for it.

Conclusion and Future Directions

In this talk, we generalized the standard mean shift algorithm to directional data, and developed statistical and computational learning theory for it.

Possible future extensions of our work are:

1. **Bandwidth Selection**: Current studies on bandwidth selectors primarily optimize the directional KDE itself. A well-designed bandwidth selector for $\nabla \widehat{f}_h$ will be needed.

Conclusion and Future Directions

In this talk, we generalized the standard mean shift algorithm to directional data, and developed statistical and computational learning theory for it.

Possible future extensions of our work are:

1. **Bandwidth Selection**: Current studies on bandwidth selectors primarily optimize the directional KDE itself. A well-designed bandwidth selector for $\nabla \widehat{f}_h$ will be needed.
2. **Accelerated Directional Mean Shift**: Our directional mean shift algorithm would be slow on large datasets. One possible way to accelerate it is to introduce the blurring procedures (Cheng, 1995; Carreira-Perpiñán, 2006, 2008).

Conclusion and Future Directions

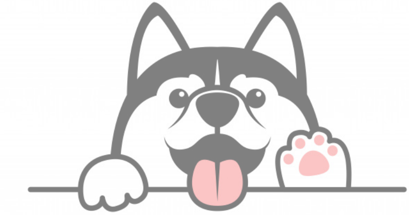
In this talk, we generalized the standard mean shift algorithm to directional data, and developed statistical and computational learning theory for it.

Possible future extensions of our work are:

1. **Bandwidth Selection**: Current studies on bandwidth selectors primarily optimize the directional KDE itself. A well-designed bandwidth selector for $\nabla \widehat{f}_h$ will be needed.
2. **Accelerated Directional Mean Shift**: Our directional mean shift algorithm would be slow on large datasets. One possible way to accelerate it is to introduce the blurring procedures (Cheng, 1995; Carreira-Perpiñán, 2006, 2008).
3. **Connections to the EM Algorithm**: The Gaussian mean shift algorithm for Euclidean data is known to be an EM algorithm (Carreira-Perpiñán, 2007). It is possible that our directional mean shift with the von Mises kernel is an EM algorithm as well (Banerjee et al., 2005).

Thank you!

More details can be found in https://arxiv.org/abs/2010.13523.
The code for our experiments is available at
https://github.com/zhangyk8/DirMS.



Reference I

P. A. Absil, R. Mahony, and J. Trumpf. An extrinsic look at the riemannian hessian. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, pages 361–368, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

Y. Aliyari Ghassabeh. On the convergence of the mean shift algorithm in the one-dimensional space. *Pattern Recognition Letters*, 34(12):1423 – 1427, 2013.

Y. Aliyari Ghassabeh. A sufficient condition for the convergence of the mean shift algorithm with gaussian kernel. *Journal of Multivariate Analysis*, 135:1 – 10, 2015.

Z. Bai, C. Rao, and L. Zhao. Kernel estimators of density function of directional data. *Journal of Multivariate Analysis*, 27(1):24 – 39, 1988.

A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, USA, 2004.

M. Á. Carreira-Perpiñán. Fast nonparametric clustering with gaussian blurring mean-shift. *Proceedings of the 23rd International Conference on Machine Learning*, 2006:153–160, 01 2006.

M. Á. Carreira-Perpiñán. Gaussian mean-shift is an em algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(5):767–776, May 2007.

M. Á. Carreira-Perpiñán. Generalised blurring mean-shift algorithms for nonparametric clustering. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

M. Á. Carreira-Perpiñán. A review of mean-shift algorithms for clustering. *arXiv preprint arXiv:1503.00687*, 2015.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. A comprehensive approach to mode clustering. *Electronic Journal of Statistics*, 10(1):210–241, 2016. URL https://doi.org/10.1214/15-EJS1102.

Reference II

Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

E. García-Portugués. Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. *Electron. J. Stat.*, 7:1655–1685, 2013.

E. García-Portugués, R. M. Crujeiras, and W. González-Manteiga. Kernel density estimation for directional-linear data. *Journal of Multivariate Analysis*, 121:152 – 175, 2013.

E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907 – 921, 2002.

P. Hall, G. S. Watson, and J. Cabrara. Kernel density estimation with spherical data. *Biometrika*, 74(4): 751–762, 12 1987. ISSN 0006-3444. URL https://doi.org/10.1093/biomet/74.4.751.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, second edition, 2012. doi: 10.1017/9781139020411.

X. Li, Z. Hu, and F. Wu. A note on the convergence of the mean shift. *Pattern Recognition*, 40(6):1756 – 1762, 2007.

K. Mardia and P. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley, 2000.

Wikipedia. von Mises distribution, 2020. URL https://en.wikipedia.org/wiki/Von_Mises_distribution. [Online; Accessed 11-November-2020].

Y. Zhang and Y.-C. Chen. Kernel smoothing, mean shift, and their learning theory with directional data. 2020. URL https://arxiv.org/abs/2010.13523.

L. Zhao and C. Wu. Central limit theorem for integrated squared error of kernel estimators of spherical density. *Sci. China Ser. A Math.*, 44(4):474–483, 2001.

Other Examples of Circular Densities

Any probability density function (pdf) $p(x)$ on $\mathbb{R}$ can be “wrapped” into a circular density as follows (e.g., page 52 in Mardia and Jupp (2000)):

$$p_w(\theta) = \sum_{k=-\infty}^{\infty} p(\theta + 2\pi k),$$

where k is an integer and $\theta \in [-\pi, \pi]$. For instance,

- wrapped normal distribution:

$$WN(\theta; \mu, \sigma) \sim \frac{1}{\sigma\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \exp\left[-\frac{(\theta - \mu + 2\pi k)^2}{2\sigma^2}\right].$$

- wrapped Cauchy distribution:

$$WC(\theta; \theta_0, \gamma) = \sum_{n=-\infty}^{\infty} \frac{\gamma}{\pi\left[\gamma^2 + (\theta + 2\pi n - \theta_0)^2\right]} = \frac{1}{2\pi} \frac{\sinh \gamma}{\cosh \gamma - \cos(\theta - \theta_0)},$$

where γ is the scale factor and θ_0 is the peak position.

Other Examples of Directional Densities

On the two-dimensional unit sphere Ω_2, the pdf of the Kent distribution is given by:

$$f_{Kent}(\boldsymbol{x}) = \frac{1}{c(\nu, \beta)} \exp\left\{\nu \boldsymbol{\gamma}_1^T \boldsymbol{x} + \beta \left[(\boldsymbol{\gamma}_2^T \boldsymbol{x})^2 - (\boldsymbol{\gamma}_3^T \boldsymbol{x})^2\right]\right\},$$

where the normalizing constant

$$c(\nu, \beta) = 2\pi \sum_{j=0}^{\infty} \frac{\Gamma\left(j + \frac{1}{2}\right)}{\Gamma(j+1)} \beta^{2j} \left(\frac{\nu}{2}\right)^{-2j-\frac{1}{2}} \mathcal{I}_{2j+\frac{1}{2}}(\nu),$$

$\boldsymbol{\gamma}_j, j = 1, 2, 3$ are orthonormal, and $\Gamma(\cdot)$ is the gamma function. In Ω_q, the Kent density is proportional to

$$f_{Kent}(\boldsymbol{x}) \propto \exp\left[\nu \boldsymbol{\gamma}_1^T \boldsymbol{x} + \sum_{j=2}^{q+1} \beta_j (\boldsymbol{\gamma}_j^T \boldsymbol{x})^2\right].$$

Detailed Derivation of Riemannian Hessian

Given a geodesic curve $\alpha : (-\epsilon, \epsilon) \to \Omega_q$ with $\alpha(0) = \boldsymbol{x}$ and $\alpha'(0) = \boldsymbol{v}$,

$$\begin{aligned}
\boldsymbol{v}^T \mathcal{H} f(\boldsymbol{x}) \boldsymbol{v} &= \frac{d^2}{dt^2} f(\alpha(t)) \Big|_{t=0} \\
&= \frac{d}{dt} \left[\nabla f(\alpha(t))^T \alpha'(t) \right] \Big|_{t=0} \\
&= \alpha'(0)^T \nabla\nabla f(\alpha(0)) \alpha'(0) + \nabla f(\alpha(0))^T \alpha''(0) \\
&= \boldsymbol{v}^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{v} - \nabla f(\boldsymbol{x})^T \boldsymbol{x} \\
&= \boldsymbol{v}^T (\nabla\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T \boldsymbol{x} I_{q+1}) \boldsymbol{v},
\end{aligned} \tag{21}$$

where we use the fact that $\alpha''(0) = -\boldsymbol{x}$.

Sketched Proof of Pointwise Consistency

Proof (Sketched)

$$\begin{aligned}
&\mathtt{Tang}\left(\nabla\widehat{f}_h(\boldsymbol{x})\right) - \mathtt{Tang}\left(\nabla f(\boldsymbol{x})\right)\\
&= \underbrace{\mathbb{E}\left[\mathtt{Tang}\left(\nabla\widehat{f}_h(\boldsymbol{x})\right)\right] - \mathtt{Tang}\left(\nabla f(\boldsymbol{x})\right)}_{\text{bias}} + \underbrace{\mathtt{Tang}\left(\nabla\widehat{f}_h(\boldsymbol{x})\right) - \mathbb{E}\left[\mathtt{Tang}\left(\nabla\widehat{f}_h(\boldsymbol{x})\right)\right]}_{\text{stochastic variation}}
\end{aligned}$$

and $\mathbb{E}\left[\mathtt{Tang}\left(\nabla\widehat{f}_h(\boldsymbol{x})\right)\right] = \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right)\mathbb{E}\left[\nabla\widehat{f}_h(\boldsymbol{x})\right]$. Then

$$\mathbb{E}\left[\nabla\widehat{f}_h(\boldsymbol{x})\right] = \frac{c_{h,q}(L)}{h^2}\int_{\Omega_q}(-\boldsymbol{y})\cdot L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y}).$$

For a variable $\boldsymbol{y}\in\Omega_q$ and a fixed point $\boldsymbol{x}\in\Omega_q$, we write $t = \boldsymbol{x}^T\boldsymbol{y}$ and

$$\boldsymbol{x} = t\boldsymbol{y} + (1-t^2)^{\frac{1}{2}}\boldsymbol{\xi},$$

where $\boldsymbol{\xi}\in\Omega_q$ is a unit vector orthogonal to $\boldsymbol{y}$. Further, an area element on Ω_q can be written as

$$\omega_q(d\boldsymbol{x}) = (1-t^2)^{\frac{q}{2}-1}dt\,\omega_{q-1}(d\boldsymbol{\xi}).$$

Sketched Proof of Pointwise Consistency

Proof (Sketched)

$$\begin{aligned}
\mathbb{E}\left[\nabla \widehat{f}_h(\boldsymbol{x})\right] &= \frac{c_{h,q}(L)}{h^2}\int_{\Omega_q}(-\boldsymbol{y})L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y}) \\
&= \frac{c_{h,q}(L)}{h^2}\int_{-1}^{1}\int_{\Omega_{q-1}}\left(-t\boldsymbol{x}-\sqrt{1-t^2}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right)L'\left(\frac{1-t}{h^2}\right) \\
&\quad \times f\left(-t\boldsymbol{x}-\sqrt{1-t^2}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right)(1-t^2)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dt \\
&= c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}(-\boldsymbol{x}-\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\cdot L'(r) \\
&\quad \times f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\cdot r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr,
\end{aligned}$$

where $\alpha_{\boldsymbol{x},\boldsymbol{\xi}} = -rh^2\boldsymbol{x} + h\sqrt{r(2-h^2r)}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}$ and $\boldsymbol{B}_{\boldsymbol{x}} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q)_{(q+1)\times q}$ is the semi-orthonormal matrix resulting from the completion of $\boldsymbol{x}$ to the orthonormal basis $\{\boldsymbol{x}, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_q\}$.

Regularity Conditions for Uniform Consistency

Let $[\tau] = (\tau_1, ..., \tau_{q+1})$ be a multi-index (i.e., $\tau_1, ..., \tau_{q+1}$ are non-negative integers and $|[\tau]| = \sum_{j=1}^{q+1} \tau_j$). Define $D^{[\tau]} = \frac{\partial^{\tau_1}}{\partial x_1^{\tau_1}} \cdots \frac{\partial^{\tau_{q+1}}}{\partial x_1^{\tau_{q+1}}}$ as the $|[\tau]|$-th order partial derivative operator. Let

$$\mathcal{K} = \left\{ \boldsymbol{u} \mapsto K\left(\frac{\boldsymbol{z}-\boldsymbol{u}}{h}\right) : \boldsymbol{u}, \boldsymbol{z} \in \Omega_q, h > 0, K(\boldsymbol{x}) = D^{[r]} L\left(\frac{1}{2}||\boldsymbol{x}||_2^2\right), |[\tau]| = 0, 1, 2 \right\}.$$

Under condition (D2′), $\mathcal{K}$ is a collection of bounded measurable functions on Ω_q. Consider the following assumption (Giné and Guillou, 2002):

- **(K1)** $\mathcal{K}$ is a bounded VC (subgraph) class of measurable functions on Ω_q, that is, there exist constants $A, \nu > 0$ such that for any $0 < \epsilon < 1$,

  $$\sup_Q N\left(\mathcal{K}, L_2(Q), \epsilon ||F||_{L_2(Q)}\right) \leq \left(\frac{A}{\epsilon}\right)^{\nu},$$

  where $N(T, d_T, \epsilon)$ is the ϵ-covering number of the pseudometric space (T, d_T), Q is any probability measure on Ω_q, and F is an envelope function of $\mathcal{K}$. The constants A and ν are usually called the VC characteristics of $\mathcal{K}$ and the norm $||F||_{L_2(Q)}$ is defined as $\left[\int_{\Omega_q} |F(\boldsymbol{x})|^2 dQ(\boldsymbol{x})\right]^{\frac{1}{2}}$.

Assumptions for Mode Consistency

Let C_3 be the upper bound for the partial derivatives of the directional density f on the compact manifold Ω_q up to the third order.

Consider the following assumptions:

- **(M1)** There exists $\lambda_* > 0$ such that

$$0 < \lambda_* \leq |\lambda_1(\boldsymbol{m}_j)|, \quad \text{for all } j = 1, ..., K,$$

  where $0 > \lambda_1(\boldsymbol{x}) \geq \cdots \geq \lambda_q(\boldsymbol{x})$ are the q smallest eigenvalues of $\mathcal{H}f(\boldsymbol{x})$.

- **(M2)** There exists $\Theta_1, \rho_* > 0$ such that

$$\left\{ \boldsymbol{x} \in \Omega_q : ||\texttt{Tang}(\nabla f(\boldsymbol{x}))||_{\max} \leq \Theta_1, \lambda_1(\boldsymbol{x}) \leq -\frac{\lambda_*}{2} < 0 \right\} \subset \mathcal{M} \oplus \rho_*,$$

  where $0 < \rho_* < \min\left\{ \sqrt{2 - 2\cos\left(\frac{3\lambda_*}{2C_3}\right)}, 2 \right\}$.

Assumptions for Mode Consistency

Condition (M1) is a weak assumption that can be satisfied by the local modes of common directional densities.
Recall that $f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}, \nu) = C_q(\nu) \cdot \exp\left(\nu \boldsymbol{\mu}^T \boldsymbol{x}\right)$. Then,

$$\nabla f_{\text{vMF}}(\boldsymbol{x}) = \nu \boldsymbol{\mu} C_q(\nu) \cdot \exp\left(\nu \boldsymbol{\mu}^T \boldsymbol{x}\right)$$

and

$$\nabla\nabla f_{\text{vMF}}(\boldsymbol{x}) = \nu^2 \boldsymbol{\mu}\boldsymbol{\mu}^T C_q(\nu) \cdot \exp\left(\nu \boldsymbol{\mu}^T \boldsymbol{x}\right),$$

which in turn indicates that at the mode $\boldsymbol{\mu} \in \Omega_q$,

$$\mathcal{H} f_{\text{vMF}}(\boldsymbol{\mu}) = -\nu C_q(\nu) \cdot e^{\nu} \left(I_{q+1} - \boldsymbol{\mu}\boldsymbol{\mu}^T\right).$$

By Brauer's theorem (Example 1.2.8 in Horn and Johnson (2012)), the eigenvalues of $\mathcal{H} f_{\text{vMF}}(\boldsymbol{\mu})$ are

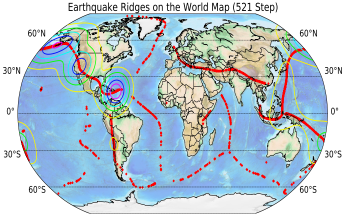
$$\left\{ 0, \underbrace{-\nu C_q(\nu) \cdot e^{\nu}, ..., -\nu C_q(\nu) \cdot e^{\nu}}_{q} \right\}.$$

Some Results on Blurring Directional Mean Shift

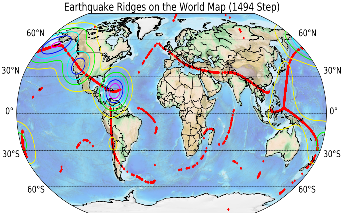
Method (Scenario)	# Est. Modes	# Steps	Avg. Err. of Est. Modes
DMS (One mode)	4.25 (1.670)	86.30 (48.774)	–
BDMS (One mode)	11.95 (2.156)	17.10 (2.700)	0.074 (0.0492)
DMS (Two modes)	2.40 (0.490)	30.55 (5.757)	–
BDMS (Two modes)	3.60 (1.114)	9.90 (1.868)	0.045 (0.0240)
DMS (Three modes)	3.00 (0.000)	28.65 (5.790)	–
BDMS (Three modes)	3.10 (0.300)	7.75 (0.698)	0.034 (0.0090)

Table: Comparisons between Directional Mean Shift (DMS) and Blurring Directional Mean Shift Algorithm (BDMS). The means and standard errors (within round brackets) are calculated with 20 repeated experiments.

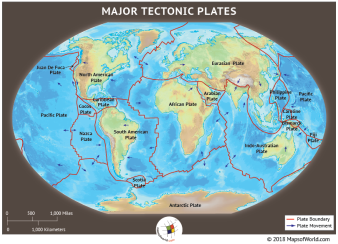
Directional Ridges on Earthquake Data



(a) Estimated directional ridges via directional SCMS



(b) Estimated directional ridges via Euclidean SCMS



(c) Real tectonic plates