Mode estimation on matrix manifolds: Convergence and robustness

Hiroaki Sasaki
Future Univ. Hakodate, Japan

Jun-ichiro Hirayama
AIST, Japan

Takafumi Kanamori
Tokyo Tech & RIKEN, Japan

Abstract

Data on matrix manifolds are ubiquitous on a wide range of research fields. The key issue is estimation of the modes (i.e., maxima) of the probability density function underlying the data. For instance, local modes (i.e., local maxima) can be used for clustering, while the global mode (i.e., the global maximum) is a robust alternative to the Fréchet mean. Previously, to estimate the modes, an iterative method has been proposed based on a Riemannian gradient estimator and empirically showed the superior performance in clustering (Ashizawa et al., 2017). However, it has not been theoretically investigated if the iterative method is able to capture the modes based on the gradient estimator. In this paper, we propose simple iterative methods for mode estimation on matrix manifolds based on the Euclidean metric. A key contribution is to perform theoretical analysis and establish sufficient conditions for the monotonic ascending and convergence of the proposed iterative methods. In addition, for the previous method, we prove the monotonic ascending property towards a mode. Thus, our work can be also regarded as compensating for the lack of theoretical analysis in the previous method. Furthermore, the robustness of the iterative methods is theoretically investigated in terms of the breakdown point. Finally, the proposed methods are experimentally demonstrated to work well in clustering and robust mode estimation on matrix manifolds.

1 Introduction

Data on Riemannian manifolds have been gathering a great deal of attentions. Probably, the simplest example is data on the unit sphere called the *directional data* in statistics (Mardia, 1972). An important group is the *matrix manifold* (Absil et al., 2008) where *Stiefel manifold*, *Grassmann manifold* and set of *symmetric positive definite (SPD) matrices* frequently appear in many practical situations. *Stiefel manifold* is the set of orthogonal matrices, and motion estimation requires orthogonal matrices as data samples by which the rigid motions can be expressed (Tuzel et al., 2005). Grassmann manifold is the set of linear subspaces, and has been used in action recognition (Slama et al., 2015). Covariance matrices, which are SPD when they have full rank, can be seen in diffusion tensor imaging (Dryden et al., 2009). More examples can be found in signal processing (Arnaudon et al., 2013), computer vision (Lui, 2012; Turaga and Srivastava, 2016) and brain science (Yger et al., 2016).

For Euclidean data, the *modes* have offered a wide-range of applications (Fukunaga and Hostetler, 1975; Comaniciu and Meer, 2002; Sager and Thisted, 1982; Chen et al., 2016), which are defined as maxima of the probability density function underlying the data. For instance, the global mode (i.e., the global maximum) can been seen as a robust alternative to the sample mean, and this robustness has been used in a number of mode-based regression methods using linear models (Lee, 1989; Yao and Li, 2014; Feng et al., 2020) as well as neural networks (Sasaki et al., 2020). On the other hand, mean shift clustering (MS) makes use of local modes (i.e., local maxima) and has a significant advantage over the standard methods that the number of clusters is automatically determined from data. Thus, it would be an important issue to estimate the modes for data on Riemannian manifolds as well.

For mode estimation, the important step is to estimate the Riemannian gradient of the probability density function. To this end, Subbarao and Meer (2006, 2009) take two steps of firstly estimating the probability density function and secondly computing its Rie-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

mannian gradient. Then, based on the computed gradient, an iterative method called the *Riemannian MS* (RMS) was proposed to estimate the modes. A notable point is that the monotonic ascending property of the iterative method towards the modes is theoretically guaranteed under certain conditions. However, the two-step approach in the Riemannian gradient estimation seems suboptimal because a good density estimator does not necessarily mean a good gradient estimator. There exist related works to RMS (Tuzel et al., 2005; Cetingul and Vidal, 2009; Caseiro et al., 2012), but these also follow a similar two-step approach in gradient estimation.

A sophisticated approach has been taken in Ashizawa et al. (2017), where a *single* step approach is adopted and a gradient model is *directly* fitted to the true Riemannian gradient. Then, based on the direct gradient estimation, an iterative method was proposed, and experimentally demonstrated to work better than RMS in clustering. However, unlike RMS, theoretical analysis of the iterative method has not been performed. In addition, this method requires to compute the exponential and logarithm maps on Riemannian manifolds, which are often computationally expensive and make the direct gradient estimation very complicated.

In this paper, we propose simple iterative methods for mode estimation on important matrix manifolds including the Stiefel, Grassmann manifolds and set of SPD matrices. We follow the direct approach in Riemannian gradient estimation as Ashizawa et al. (2017), yet employ a simple model based on the Euclidean metric. Based on our gradient models, we derive novel iterative methods for mode-seeking based on the fixed-point scheme, which has been used in Euclidean MS as well (Comaniciu and Meer, 2002, Section 2.1). In contrast with Ashizawa et al. (2017), our methods do not require to compute the exponential and logarithm maps, and thus would be substantially simpler and computationally efficient.

Furthermore, our notable contribution is to perform theoretical analysis in terms of the convergence and robustness. We establish sufficient conditions for convergence of the sequences generated by the proposed iterative methods with monotonic ascending. In addition, we prove the monotonic ascending property of the existing method (Ashizawa et al., 2017) as well. Thus, our work can be regarded as compensating for the lack of theoretical analysis in Ashizawa et al. (2017). The robustness of mode estimation on Riemannian manifolds is also investigated through the breakdown point (Huber and Ronchetti, 2009), which has not been done previously. Finally, we numerically demonstrate that the proposed methods work well in clustering and robust mode estimation on matrix manifolds.

Notations: The Frobenuous norm of a matrix $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2}$ is defined as $\|\boldsymbol{X}\|_{\mathrm{F}} := \sqrt{\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X})}$ where $^\top$ is the matrix transpose and $\mathrm{tr}(\cdot)$ denotes the trace. $\boldsymbol{O} \in \mathbb{R}^{d_1\times d_2}$ means the null matrix whose elements are all zeros, and $\boldsymbol{I}_d$ is the d by d identity matrix. For a square matrix $\boldsymbol{Z}$, $\mathrm{sym}(\boldsymbol{Z}) := \frac{1}{2}(\boldsymbol{Z}+\boldsymbol{Z}^\top)$ and $\mathrm{ddiag}(\boldsymbol{Z})$ denotes the diagonal matrix whose diagonals are those of $\boldsymbol{Z}$. The Euclidean gradient of a function $f(\boldsymbol{X})$ is denoted by $\nabla_{\boldsymbol{X}} f(\boldsymbol{X})$.

Let us denote a Riemannian manifold by $\mathbb{M}$ and the tangent space at $\boldsymbol{X} \in \mathbb{M}$ by $T_{\boldsymbol{X}}\mathbb{M}$ on which a Riemannian metric $\langle\cdot,\cdot\rangle_{\boldsymbol{X}}$ is defined . The Exponential map $\exp_{\boldsymbol{X}}(\cdot)$ at $\boldsymbol{X}$ is a mapping from $T_{\boldsymbol{X}}\mathbb{M}$ to $\mathbb{M}$, while $\log_{\boldsymbol{X}}(\cdot)$ is one from $\mathbb{M}$ to $T_{\boldsymbol{X}}\mathbb{M}$. The Riemannian gradient of a function $f(\boldsymbol{X})$ at $\boldsymbol{X} \in \mathbb{M}$ is denoted by $\mathrm{grad}(f(\boldsymbol{X})) \in T_{\boldsymbol{X}}\mathbb{M}$. When $\mathbb{M}$ is a submanifold of $\mathbb{R}^{d_1\times d_2}$, then $\mathrm{grad}(f(\boldsymbol{X})) = \boldsymbol{P}_{\boldsymbol{X}}(\nabla_{\boldsymbol{X}} f(\boldsymbol{X}))$ (Absil et al., 2008) where $\boldsymbol{P}_{\boldsymbol{X}}(\cdot)$ is the orthogonal projection onto $T_{\boldsymbol{X}}\mathbb{M}$ of an embedded submanifold or the horizontal space of a quotient manifold. The Riemannian distance $\mathrm{dist}(\boldsymbol{X},\boldsymbol{Y})$ is defined as the infimum of the length of all curves connecting $\boldsymbol{X} \in \mathbb{M}$ to $\boldsymbol{Y} \in \mathbb{M}$ (Absil et al., 2008).

2 Problem setup and related work

Problem setup: Suppose that we are given n i.i.d. data samples on a Riemannian manifold $\mathbb{M}$ drawn from the probability density function p as $\mathcal{D} := \{\boldsymbol{X}_i \in \mathbb{M}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{X})$. Our primal interest is to estimate the modes (i.e., maxima) of $p(\boldsymbol{X})$ from $\mathcal{D}$.

A simple example of $\mathbb{M}$ is the unit sphere $\Omega^{d-1} := \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\| = 1\}$ on which data is called the *directional data* in statistics (Mardia, 1972). In the application point of view, an important group is the *matrix manifold*: For instance, the *Stiefel manifold* (Absil et al., 2008) is a set of orthonormal matrices and a generalization of the unit sphere. Another example is the set of symmetric positive definite (SPD) matrices. Here, we begin with general Riemannian manifolds and then focus on particular matrix manifolds.

Related work: The *Fréchet mean* is a representative summary of data on Riemannian manifolds (Karcher, 1977), and defined as the minimizer of $\frac{1}{n}\sum_{i=1}^n \mathrm{dist}^2(\boldsymbol{X},\boldsymbol{X}_i)$ with respect to $\boldsymbol{X}$. When the Riemannian distance $\mathrm{dist}(\cdot,\cdot)$ is equal to the Euclidean distance, the Fréchet mean reduces to the sample mean. The Fréchet mean has been used on a variety of fields, but is known to be vulnerable to outliers. A robust variant is the *geometric median* (Fletcher et al., 2009), which minimizes $\frac{1}{n}\sum_{i=1}^n \mathrm{dist}(\boldsymbol{X},\boldsymbol{X}_i)$. However, as in the Euclidean median, the geometric median would also implicitly assume the underlying density is symmetric, and this might not hold when outliers concentrate on a one side of the true median because they make the underlying density highly skewed.

For Euclidean data, one of the most popular methods is the *mean shift* (MS) method (Fukunaga and Hostetler, 1975; Cheng, 1995; Comaniciu and Meer, 2002), and its Riemannian extension has been proposed (Subbarao and Meer, 2006, 2009), which we call the Riemannian mean shift (RMS). RMS starts by employing the following estimator for the data density, which is akin to *kernel density estimation* (KDE):

$$\widehat{p}(\boldsymbol{X}) := \frac{c_{h,K}}{n}\sum_{i=1}^{n} K\left(\frac{\mathrm{dist}^2(\boldsymbol{X},\boldsymbol{X}_i)}{h^2}\right), \qquad (1)$$

where $h > 0$ is a bandwidth parameter, $K(\cdot)$ is a nonnegative function and $c_{h,K}$ denotes a positive constant. Then, the Riemannian gradient is computed as

$$\mathrm{grad}(\widehat{p}(\boldsymbol{X})) = \frac{2c_{h,K}}{nh^2}\sum_{i=1}^{n}\log_{\boldsymbol{X}}(\boldsymbol{X}_i) L\left(\frac{\mathrm{dist}^2(\boldsymbol{X},\boldsymbol{X}_i)}{h^2}\right),$$

where $L(t) := -\frac{\mathrm{d}}{\mathrm{d}t}K(t)$ and we used the relation $\mathrm{grad}(\mathrm{dist}^2(\boldsymbol{X},\boldsymbol{Y})) = -2\log_{\boldsymbol{X}}(\boldsymbol{Y})$ proved in Karcher (1977). Inspired by the Euclidean mean shift method (Comaniciu and Meer, 2002), the following update rule is iteratively used to seek a mode of $\widehat{p}(\boldsymbol{X})$: Denoting the τ-th iterate by $\boldsymbol{X}(\tau)$, $\tau = 0, 1, 2, \ldots,$

$$\boldsymbol{X}(\tau+1) = \exp_{\boldsymbol{X}(\tau)}\left(\widetilde{\boldsymbol{M}}(\boldsymbol{X}(\tau))\right), \qquad (2)$$

where $\boldsymbol{X}(0)$ can be one of the samples $\boldsymbol{X}_i$, and

$$\widetilde{\boldsymbol{M}}(\boldsymbol{X}) := \frac{\sum_{i=1}^{n}\log_{\boldsymbol{X}}(\boldsymbol{X}_i) L\left(\frac{\mathrm{dist}^2(\boldsymbol{X},\boldsymbol{X}_i)}{\sigma^2}\right)}{\sum_{i=1}^{n} L\left(\frac{\mathrm{dist}^2(\boldsymbol{X},\boldsymbol{X}_i)}{\sigma^2}\right)}.$$

Eq.(2) is repeatedly applied until $\boldsymbol{X}(\tau)$ converges. Notably, Subbarao and Meer (2009) proved that the following inequality holds with the update rule (2) under some conditions: For $\tau = 0, 1, 2, \ldots,$

$$\widehat{p}(\boldsymbol{X}(\tau+1)) - \widehat{p}(\boldsymbol{X}(\tau)) \geq 0, \qquad (3)$$

which implies that (2) has the *monotonic ascending property* towards a mode of $\widehat{p}(\boldsymbol{X})$. However, the two-step approach of estimating the Riemannian gradient seems suboptimal because a good density estimator does not necessarily mean a good gradient estimator.

Ashizawa et al. (2017) adopted a more sophisticated approach for Riemannian gradient estimation, and the main idea is to *directly* fit a gradient model $\boldsymbol{g}(\boldsymbol{X})$ to the true log-density gradient $\boldsymbol{g}^*(\boldsymbol{X}) := \mathrm{grad}(\log p(\boldsymbol{X}))$ under the squared-loss as follows:

$$\begin{aligned} J(\boldsymbol{g}) &:= \int_{\mathbb{M}} \|\boldsymbol{g}(\boldsymbol{X}) - \boldsymbol{g}^*(\boldsymbol{X})\|^2_{\boldsymbol{X}} p(\boldsymbol{X})\mathrm{d}\boldsymbol{V}_{\boldsymbol{X}} - C \\ &= \int_{\mathbb{M}} \langle \boldsymbol{g}(\boldsymbol{X}), \boldsymbol{g}(\boldsymbol{X})\rangle_{\boldsymbol{X}} p(\boldsymbol{X})\mathrm{d}\boldsymbol{V}_{\boldsymbol{X}} \\ &\quad - 2\int_{\mathbb{M}} \langle \boldsymbol{g}(\boldsymbol{X}), \boldsymbol{g}^*(\boldsymbol{X})\rangle_{\boldsymbol{X}} p(\boldsymbol{X})\mathrm{d}\boldsymbol{V}_{\boldsymbol{X}}, \end{aligned} \qquad (4)$$

where $\|\cdot\|^2_{\boldsymbol{X}} := \langle\cdot,\cdot\rangle_{\boldsymbol{X}}$, $\mathrm{d}\boldsymbol{V}_{\boldsymbol{X}}$ denotes the Riemannian volume form induced by the metric $\langle\cdot,\cdot\rangle_{\boldsymbol{X}}$ (Lee, 2018) and $C := \int_{\mathbb{M}}\langle \boldsymbol{g}^*(\boldsymbol{X}), \boldsymbol{g}^*(\boldsymbol{X})\rangle_{\boldsymbol{X}} p(\boldsymbol{X})\mathrm{d}\boldsymbol{V}_{\boldsymbol{X}}$. Eq.(4) can be regarded as a Riemannian extension of the Fisher divergence, which has been used for direct density gradient estimation on the Euclidean space (Cox, 1985; Sasaki et al., 2014). When $\mathbb{M}$ is a Riemannian manifold without boundary, the following "integration by parts" (Lee, 2012, Chapter 16) is applicable to make the second term on the right-hand side of (4) to be tractable:

$$\begin{aligned} &\int_{\mathbb{M}} \langle \boldsymbol{g}(\boldsymbol{X}), \boldsymbol{g}^*(\boldsymbol{X})\rangle_{\boldsymbol{X}} p(\boldsymbol{X})\mathrm{d}\boldsymbol{V}_{\boldsymbol{X}} \\ &\qquad = -\int_{\mathbb{M}} \mathrm{div}(\boldsymbol{g}(\boldsymbol{X})) p(\boldsymbol{X})\mathrm{d}\boldsymbol{V}_{\boldsymbol{X}}, \end{aligned}$$

where div is the divergence operator on $\mathbb{M}$ (Lee, 2018). Thus, the empirical version of $J(\boldsymbol{g})$ is obtained by

$$\widehat{J}(\boldsymbol{g}) := \frac{1}{n}\sum_{i=1}^{n}\left[\langle \boldsymbol{g}(\boldsymbol{X}_i), \boldsymbol{g}(\boldsymbol{X}_i)\rangle_{\boldsymbol{X}_i} + 2\mathrm{div}(\boldsymbol{g}(\boldsymbol{X}_i))\right]. \qquad (5)$$

For $\boldsymbol{g}(\boldsymbol{X})$, the following gradient model is adopted:

$$g(\boldsymbol{X}) := \sum_{i=1}^{n} a_i \log_{\boldsymbol{X}}(\boldsymbol{X}_i)\phi\left(\frac{\mathrm{dist}^2(\boldsymbol{X},\boldsymbol{X}_i)}{\sigma^2}\right), \qquad (6)$$

where $\phi(\cdot)$ is a nonnegative function whose conditions are specified later. Then, by substituting (6) into (5), the parameter vector $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$ is estimated by minimizing $\widehat{J}(\boldsymbol{a})$. In order to seek the modes, Ashizawa et al. (2017) proposed the following iterative rule as in RMS and used it until $\boldsymbol{X}(\tau)$ converges:

$$\boldsymbol{X}(\tau+1) = \exp_{\boldsymbol{X}(\tau)}(\boldsymbol{M}(\boldsymbol{X}(\tau))), \qquad (7)$$

where

$$\boldsymbol{M}(\boldsymbol{X}) = \frac{\sum_{i=1}^{n} a_i \log_{\boldsymbol{X}}(\boldsymbol{X}_i)\phi\left(\frac{\mathrm{dist}^2(\boldsymbol{X},\boldsymbol{X}_i)}{\sigma^2}\right)}{\sum_{i=1}^{n} a_i \phi\left(\frac{\mathrm{dist}^2(\boldsymbol{X},\boldsymbol{X}_i)}{\sigma^2}\right)}.$$

Ashizawa et al. (2017) empirically demonstrated that (7) works much better than RMS in clustering on the Grassmann manifold. However, in contrast with RMS, theoretical analysis has not been performed.

Furthermore, when $\mathbb{M}$ is a matrix manifold (e.g., the set of SPD matrices), $\exp_{\boldsymbol{X}}(\cdot)$ and $\log_{\boldsymbol{X}}(\cdot)$ may require to compute the matrix exponential and logarithm (Boumal, 2020, Section 11.5), which could make the computation of $\widehat{J}(\boldsymbol{g})$ very complicated and the iteration (7) slow. Here, we follow the same direct approach in Riemannian gradient estimation as Ashizawa et al. (2017), yet employ a simpler gradient model and derive update rules for mode estimation on matrix manifolds. A key contribution is to perform theoretical analysis and establish sufficient conditions of the convergence for the proposed iterative methods with monotonic ascending. In addition, we theoretically prove the monotonic ascending property of the existing update rule (7). The robustness of the iterative methods in mode estimation is also investigated in terms of breakdown point (Huber and Ronchetti, 2009).

3 Mode seeking on matrix manifolds

This section supposes that $\mathbb{M}$ is a matrix manifold, which is constructed from $\mathbb{R}^{d_1\times d_2}$ $(d_1 \geq d_2)$ by taking the operations of embedded submanifolds and quotient manifolds (Absil et al., 2008). Thus, $\boldsymbol{X}\in\mathbb{M}$ can be regarded as an element in $\mathbb{R}^{d_1\times d_2}$ by adopting the Euclidean coordinate, yet has to satisfy a certain *condition*: For instance, if $\boldsymbol{X}$ is an element in the Stiefel manifold, $\boldsymbol{X}$ is a d_1 by d_2 *orthonormal* matrix in the sense that $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{I}_{d_2}$. Here, we introduce simple models for Riemannian gradient estimation from which we derive iterative rules to seek the modes on important matrix manifolds. We do *not* show how to estimate the models in this section, but the details are given in the supplementary material.

3.1 Iterative update rule on $\mathrm{St}(d_1, d_2)$

We first focus on the Stiefel manifold $\mathrm{St}(d_1,d_2) := \{\boldsymbol{X}\in\mathbb{R}^{d_1\times d_2} | \boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{I}_{d_2}\}$. Since it is a maximum of $\log p(\boldsymbol{X})$, from the optimality condition, a mode of $\log p(\boldsymbol{X})$ satisfies

$$\mathrm{grad}(\log p(\boldsymbol{X})) = \boldsymbol{O} \quad (\boldsymbol{X}\in\mathbb{M}). \tag{8}$$

Alternatively, the mode can be seen as a maximum of the following Lagrangian function on $\mathbb{R}^{d_1\times d_2}$ as

$$\log p(\boldsymbol{X}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}^\top(\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{I}_{d_2})) \quad (\boldsymbol{X}\in\mathbb{R}^{d_1\times d_2}),$$

where $\boldsymbol{\Lambda}$ is a d_2 by d_2 matrix of Lagrange multipliers. Then, the mode satisfies the optimality conditions given by

$$\nabla_{\boldsymbol{X}} \log p(\boldsymbol{X}) - \boldsymbol{X}\mathrm{sym}(\boldsymbol{\Lambda}) = \boldsymbol{O} \tag{9}$$

$$\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{I}_{d_2}. \tag{10}$$

Next, we substitute our Euclidean gradient model for $\nabla_{\boldsymbol{X}}\log p(\boldsymbol{X})$ in (9) and derive an update rule based on the fixed-point scheme.

For the Euclidean gradient $\nabla_{\boldsymbol{X}}\log p(\boldsymbol{X})$, we employ the following trace model:

$$\boldsymbol{g}_{\mathrm{e}}(\boldsymbol{X}) = \sum_{i=1}^n a_i\boldsymbol{X}_i\phi\left(\frac{\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X}_i)}{\sigma^2}\right), \tag{11}$$

where a_i are coefficients and the conditions of $\phi(\cdot)$ are specified in Theorem 1. Then, substituting the trace model (11) for $\nabla_{\boldsymbol{X}}\log p(\boldsymbol{X})$ in (9) yields

$$\begin{aligned}&\boldsymbol{g}_{\mathrm{e}}(\boldsymbol{X}) - \boldsymbol{X}\mathrm{sym}(\boldsymbol{\Lambda})\\ &= \sum_{i=1}^n a_i\boldsymbol{X}_i\phi\left(\frac{\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X}_i)}{\sigma^2}\right) - \boldsymbol{X}\mathrm{sym}(\boldsymbol{\Lambda}) = \boldsymbol{O}.\end{aligned}$$

Applying the fixed-point scheme to the equation above gives the following naive update rule as

$$\boldsymbol{X} \leftarrow \sum_{i=1}^n a_i\boldsymbol{X}_i\phi\left(\frac{\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X}_i)}{\sigma^2}\right)\mathrm{sym}(\boldsymbol{\Lambda})^{-1}.$$

By denoting the τ-th iterate by $\boldsymbol{X}(\tau)$ and taking $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{I}_{d_2}$ in (10) into account, the update rule is finally obtained as

$$\boldsymbol{X}(\tau+1) = \boldsymbol{Z}(\tau)(\boldsymbol{Z}(\tau)^\top\boldsymbol{Z}(\tau))^{-\frac{1}{2}} \tag{12}$$

where $\boldsymbol{Z}(\tau) := \sum_{i=1}^n a_i\boldsymbol{X}_i\phi\left(\frac{\mathrm{tr}(\boldsymbol{X}(\tau)^\top\boldsymbol{X}_i)}{\sigma^2}\right)$.

Directional data: The simplest example is the case of $d_2 = 1$, which corresponds to directional data (i.e., data on Ω_{d_1-1}). Then, from (12), the update rule for directional data $\boldsymbol{x}\in\Omega_{d_1-1}$ is obtained as

$$\boldsymbol{x}(\tau+1) = \frac{\sum_{i=1}^n a_i\boldsymbol{x}_i\phi\left(\frac{\boldsymbol{x}(\tau)^\top\boldsymbol{x}_i}{\sigma^2}\right)}{\left\|\sum_{i=1}^n a_i\boldsymbol{x}_i\phi\left(\frac{\boldsymbol{x}(\tau)^\top\boldsymbol{x}_i}{\sigma^2}\right)\right\|}, \tag{13}$$

where $\|\cdot\|$ denotes the Euclidean norm. Previously, mean shift clustering has been extended for directional data (Kobayashi and Otsu, 2010; Kafai et al., 2010; Zhang and Chen, 2020, 2021) and is called as the *directional mean shift* (DMS). Interestingly, the update rule used in DMS is a special case of (13) where $a_i = 1/n$ for $i = 1, 2, \ldots, n$. Thus, our work can be regarded as a generalization of DMS.

Oblique manifold: By essentially following the same steps as the Stiefel manifold, we can derive an update rule on the *oblique manifold* (Absil and Gallivan, 2006): $\mathrm{Ob}(d_1,d_2) := \{\boldsymbol{X}\in\mathbb{R}^{d_1\times d_2} : \mathrm{ddiag}(\boldsymbol{X}^\top\boldsymbol{X}) = \boldsymbol{I}_{d_2})\}$. The update rule is give by

$$\boldsymbol{X}(\tau+1) = \boldsymbol{Z}(\tau)(\mathrm{ddiag}(\boldsymbol{Z}(\tau)^\top\boldsymbol{Z}(\tau)))^{-\frac{1}{2}}, \tag{14}$$

whose derivation is deferred to Section A.

3.2 Iterative update rule on $\mathrm{Gr}(d_1, d_2)$

We next consider the Grassmann manifold defined by $\mathrm{Gr}(d_1, d_2) := \{\mathrm{span}(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathrm{St}(d_1, d_2)\}$, where $\mathrm{span}(\boldsymbol{X})$ denotes the linear subspace spanned by the columns of $\boldsymbol{X}$. We adopt the same Lagrangian function on $\mathbb{R}^{d_1 \times d_2}$ as the Stiefel manifold, and obtain the same optimality conditions as (9) and (10).

For the Grassmann manifold $\mathrm{Gr}(d_1, d_2)$, we employ the following Euclidean gradient model:

$$\boldsymbol{g}_{\mathrm{e}}(\boldsymbol{X}) = \sum_{i=1}^{n} a_i \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{X} \phi_i^{\mathrm{Gr}}(\boldsymbol{X}), \tag{15}$$

where $\phi_i^{\mathrm{Gr}}(\boldsymbol{X}) := \phi\left(\frac{\mathrm{tr}(\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top)}{\sigma^2}\right)$, and $\boldsymbol{X}\boldsymbol{X}^\top$ is an orthogonal projector to $\mathrm{span}(\boldsymbol{X})$ and can be also seen as a matrix representation of $\mathrm{span}(\boldsymbol{X})$. By following the same steps based on the fixed-point scheme, an update rule is derived as

$$\boldsymbol{X}(\tau+1) = \boldsymbol{Y}(\tau)\boldsymbol{X}(\tau)(\boldsymbol{X}(\tau)^\top \boldsymbol{Y}(\tau)^\top \boldsymbol{Y}(\tau)\boldsymbol{X}(\tau))^{-\frac{1}{2}}, \tag{16}$$

where $\boldsymbol{Y}(\tau) := \sum_{i=1}^{n} a_i \boldsymbol{X}_i \boldsymbol{X}_i^\top \phi_i^{\mathrm{Gr}}(\boldsymbol{X}(\tau))$.

3.3 Iterative update rule on $\mathrm{S}^+(d)$

This subsection considers the set of symmetric positive definite (SPD) matrices: $\mathrm{S}^+(d) := \{\boldsymbol{X} \in \mathbb{R}^{d \times d} \mid \boldsymbol{X} = \boldsymbol{X}^\top \succ \boldsymbol{O}\}$, where $d = d_1 = d_2$. In contrast with the Stiefel and Grassmann manifolds, the derivation of an update rule is not based on the Lagrangian function, but directly on the Riemannian gradient (8) for $\mathrm{S}^+(d)$.

Here, we employ the following model for the Euclidean gradient:

$$\boldsymbol{g}_{\mathrm{e}}(\boldsymbol{X}) = \sum_{i=1}^{n} a_i (\boldsymbol{X}_i - \boldsymbol{X}) \phi\left(\frac{\|\boldsymbol{X} - \boldsymbol{X}_i\|_{\mathrm{F}}^2}{\sigma^2}\right) \tag{17}$$

Then, with the orthogonal projector $\boldsymbol{P}_{\boldsymbol{X}}(\cdot) = \mathrm{sym}(\cdot)$ to the tangent space of $\mathrm{S}^+(d)$ (Vandereycken et al., 2009), a model of the Riemannian gradient is given by

$$\begin{aligned}\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{g}_{\mathrm{e}}(\boldsymbol{X})) &= \sum_{i=1}^{n} a_i \mathrm{sym}(\boldsymbol{X}_i - \boldsymbol{X}) \phi_i(\boldsymbol{X}) \\ &= \left[\sum_{i=1}^{n} a_i \phi_i(\boldsymbol{X})\right] \left[\frac{\sum_{i=1}^{n} a_i \boldsymbol{X}_i \phi_i(\boldsymbol{X})}{\sum_{i=1}^{n} a_i \phi_i(\boldsymbol{X})} - \boldsymbol{X}\right],\end{aligned}$$

where $\phi_i(\boldsymbol{X}) := \phi\left(\frac{\|\boldsymbol{X} - \boldsymbol{X}_i\|_{\mathrm{F}}^2}{\sigma^2}\right)$, we assumed $\sum_{i=1}^{n} a_i \phi_i(\boldsymbol{X}) \neq 0$, and used $\mathrm{sym}(\boldsymbol{X}_i - \boldsymbol{X}) = \boldsymbol{X}_i - \boldsymbol{X}$ because $\boldsymbol{X}_i$ and $\boldsymbol{X}$ are symmetric. After substituting $\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{g}_{\mathrm{e}}(\boldsymbol{X}))$ for $\mathrm{grad}(\nabla \log p(\boldsymbol{X}))$ in (8), the fixed-point scheme leads to the following update formula:

$$\boldsymbol{X}(\tau+1) = \frac{\sum_{i=1}^{n} a_i \boldsymbol{X}_i \phi_i(\boldsymbol{X}(\tau))}{\sum_{i=1}^{n} a_i \phi_i(\boldsymbol{X}(\tau))}. \tag{18}$$

Eq.(18) shows that the right-hand side is an SPD matrix if $\phi(\cdot)$ and a_i for all i are nonnegative.

4 Convergence analysis

This section theoretically investigates the monotonic ascending of the proposed and existing iterative rules, and convergence of the sequence $\{\boldsymbol{X}(\tau)\}_{\tau=0,1,\ldots}$.

4.1 Convergence with monotonic ascending on $\mathrm{St}(d_1, d_2)$, $\mathrm{Ob}(d_1, d_2)$, $\mathrm{Gr}(d_1, d_2)$ and $\mathrm{S}^+(d)$

Here, we prove the convergence of the proposed methods on $\mathrm{St}(d_1, d_2)$, $\mathrm{Ob}(d_1, d_2)$, $\mathrm{Gr}(d_1, d_2)$ and $\mathrm{S}^+(d)$. This is not a trivial task because of the following two reasons: First, each of the proposed methods is derived as a stationary point of the Lagrange function or a zero of the Riemannian gradient model. Therefore, it is unclear if the proposed methods perform even gradient ascent on these matrix manifolds. Second, since we adopt the same *direct* approach as Ashizawa et al. (2017) in the Riemannian gradient estimation, it is not so straightforward to ensure the convergence and monotonic ascending because we have no density estimate but only a gradient estimator is available.

To investigate the monotonic ascending property and convergence, in this analysis, we employ the *line integral* on Riemannian manifolds (Lee, 2012, Theorem 11.39): For the vector field $\boldsymbol{g}^*(\boldsymbol{X}) = \mathrm{grad}(\log p(\boldsymbol{X}))$ and a differentiable curve $\boldsymbol{C}(t) \in \mathbb{M}$, $t \in [0, 1]$ connecting $\boldsymbol{Y} \in \mathbb{M}$ and $\boldsymbol{X} \in \mathbb{M}$, the line integral is formulated as

$$\begin{aligned}D_{\boldsymbol{g}^*}[\boldsymbol{Y}|\boldsymbol{X}] &:= \int_0^1 \langle \dot{\boldsymbol{C}}(t), \boldsymbol{g}^*(\boldsymbol{C}(t)) \rangle_{\boldsymbol{C}(t)} \mathrm{d}t \\ &= \log p(\boldsymbol{X}) - \log p(\boldsymbol{Y}),\end{aligned} \tag{19}$$

where $\dot{\boldsymbol{C}}(t) := \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{C}(t)$. The formula (19) indicates that the Riemannian gradient enables us to evaluate the difference of the log-densities (i.e., $\log p(\boldsymbol{X}) - \log p(\boldsymbol{Y})$).

Next, we substitute our model $\boldsymbol{g}(\boldsymbol{X}) := \boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{g}_{\mathrm{e}}(\boldsymbol{X}))$ for the Riemannian gradient $\boldsymbol{g}^*(\boldsymbol{X})$ and obtain

$$D_{\boldsymbol{g}}[\boldsymbol{Y}|\boldsymbol{X}] := \int_0^1 \langle \dot{\boldsymbol{C}}(t), \boldsymbol{g}(\boldsymbol{C}(t)) \rangle_{\boldsymbol{C}(t)} \mathrm{d}t. \tag{20}$$

Then, from (19), we say that an iterative rule has the *monotonic ascending property* if

$$D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] \geq 0, \ (\tau = 0, 1, 2, \ldots).$$

The following theorem states sufficient conditions for monotonic ascending:

Theorem 1. *Suppose that the following assumptions hold for* $\mathrm{St}(d_1,d_2)$*,* $\mathrm{Ob}(d_1,d_2)$*,* $\mathrm{Gr}(d_1,d_2)$ *and* $\mathrm{S}^+(d)$*: (A1) All coefficients* a_i *are nonnegative, (A2)* $\phi(\cdot)$ *is continuous, there exists a continuous, finite, convex and monotonically increasing (or decreasing) function* $\varphi(t)$ *such that* $\phi(t) = \frac{\mathrm{d}}{\mathrm{d}t}\varphi(t)$ *(or* $\phi(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\varphi(t)$*) for* $\mathrm{St}(d_1,d_2)$*,* $\mathrm{Ob}(d_1,d_2)$ *and* $\mathrm{Gr}(d_1,d_2)$ *(or* $\mathrm{S}^+(d)$*), and* $\phi(t)$ *and* $\varphi(t)$ *on* $t \geq 0$ *are bounded for* $\mathrm{S}^+(d)$*. Regarding* $\mathrm{Gr}(d_1,d_2)$*, we further assume that (A3)* $\mathrm{tr}(\boldsymbol{X}(\tau)^\top \boldsymbol{Y}(\tau)\boldsymbol{X}(\tau)) \leq \mathrm{tr}(\{\boldsymbol{X}(\tau)^\top \boldsymbol{Y}(\tau)^\top \boldsymbol{Y}(\tau)\boldsymbol{X}(\tau)\}^{\frac{1}{2}})$ *for* $\tau = 0,1,\ldots.$ *Then, with the Euclidean metric* $\langle \boldsymbol{Z},\boldsymbol{Y}\rangle_{\boldsymbol{X}} = \mathrm{tr}(\boldsymbol{Z}^\top\boldsymbol{Y})$*,*

$$D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] \geq 0,$$

and the sequence $\{D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)]\}_{\tau=0,1,\ldots}$ *converges to zero when* (12)*,* (14)*,* (16) *and* (18) *are used as the update rules for* $\mathrm{St}(d_1,d_2)$*,* $\mathrm{Ob}(d_1,d_2)$*,* $\mathrm{Gr}(d_1,d_2)$ *and* $\mathrm{S}^+(d)$*, respectively.*

The proof is given in Section B. Theorem 1 indicates that our update rules are guaranteed to monotonically update $\boldsymbol{X}(\tau)$ towards modes of the gradient models under certain conditions. This theoretical guarantee is the significant difference to Ashizawa et al. (2017), and has not been established in the direct Riemannian gradient estimation.

Assumption (A1) can be satisfied in practice: As detailed in Section G, substituting the models in Section 3 into $\widehat{J}(\boldsymbol{g})$ leads to a quadratic form, and thus the nonnegative constraint can be easily added. Assumption (A2) holds when $\varphi(t) = \exp(t)$, which makes a function $\varphi\left(\frac{\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X}_i)}{\sigma^2}\right)$ the exponential kernel. On the other hand, if $\varphi(t) = \exp(-t)$ for $\mathrm{S}^+(d)$, it produces the Gaussian kernel in (17). Regarding Assumption (A3) for $\mathrm{Gr}(d_1,d_2)$, we conjecture that this assumption is mild because it holds when $d_2 = 1$ (See Section C for details).

Rigorously speaking, the monotonic ascending property only does not necessarily guarantee the convergence of $\{\boldsymbol{X}(\tau)\}_{\tau=0,1,\ldots}$ (Li et al., 2007). The following theorem establishes some conditions for convergence:

Theorem 2. *Assume that (B1) all coefficients* a_i *are positive, (B2) the number of zeros of the Riemannian gradient models* $\boldsymbol{g}(\boldsymbol{X})$ *is finite on* $S_0 := \{\boldsymbol{X}|D_{\boldsymbol{g}}[\boldsymbol{X}(0)|\boldsymbol{X}] \geq 0\}$*, and (B3)* $\boldsymbol{Z}(\tau)$ *in* (12) *and* (14) *and* $\boldsymbol{Y}(\tau)$ *in* (16) *have full rank for* $\tau = 0,1,\ldots.$ *Then, under Assumptions (A2,3) in Theorem 1, the sequence* $\{\boldsymbol{X}(\tau)\}_{\tau=0,1,\ldots}$ *converges to a zero of* $\boldsymbol{g}(\boldsymbol{X})$ *with monotonic ascending when* (12)*,* (14)*,* (16) *and* (18) *are used as the update rules for* $\mathrm{St}(d_1,d_2)$*,* $\mathrm{Ob}(d_1,d_2)$*,* $\mathrm{Gr}(d_1,d_2)$ *and* $\mathrm{S}^+(d)$*, respectively.*

The proof is given in Section E, and is a similar line of the proof for Theorem 2 in Li et al. (2007). Theorem 2 indicates that the sequence $\{\boldsymbol{X}(\tau)\}_{\tau=0,\ldots,}$ converges to a mode based on the Riemannian gradient models when they have no saddle points. Thus, in most of practical situations, the proposed methods are useful to estimate the modes. On the other hand, when the gradient models have flat local maxima or saddle points, the convergence remains unclear. We will endeavor to investigate this interesting point in future.

4.2 Monotonic ascending on $\mathbb{M}$

Next, we establish sufficient conditions for the monotonic ascending property of (7) proposed in Ashizawa et al. (2017).

Brief preparation: Here, we introduce some notions and tools in Riemannian manifolds used in Afsari et al. (2013) whose results are key in our proof. An upper bound of the sectional curvature and injectivity radius of $\mathbb{M}$ are denoted by κ_1 and $\mathrm{inj}(\mathbb{M})$, respectively. Then, we define a constant r_* as

$$r_* := \begin{cases} \frac{1}{2}\min\left\{\mathrm{inj}(\mathbb{M}), \frac{\pi}{\sqrt{\kappa_1}}\right\} & \kappa_1 > 0 \\ \frac{1}{2}\mathrm{inj}(\mathbb{M}) & \kappa_1 \leq 0. \end{cases}$$

The open ball centered at $\boldsymbol{X}_o \in \mathbb{M}$ and with radius r,

$$B(\boldsymbol{X}_o, r) := \{\boldsymbol{X} \in \mathbb{M} \mid \mathrm{dist}(\boldsymbol{X},\boldsymbol{X}_o) < r\},$$

is called *convex* if for any $\boldsymbol{X},\boldsymbol{Y} \in B(\boldsymbol{X}_o, r)$, there is a unique minimizing geodesic from $\boldsymbol{X}$ to $\boldsymbol{Y}$ and the geodesic entirely lies in $B(\boldsymbol{X}_o, r)$ (Lee, 2018). For any $\boldsymbol{X}_o$, $B(\boldsymbol{X}_o, r)$ is convex when $r \leq r_*$ (Petersen, 2006). By denoting a lower bound of the sectional curvatures of $\mathbb{M}$ by κ_0, the following is defined: For $r > 0$,

$$c(r,\kappa_0) := \begin{cases} 1, & \kappa_0 \geq 0 \\ r\sqrt{|\kappa_0|}\coth(r\sqrt{|\kappa_0|}), & \kappa_0 < 0 \end{cases}$$

Since $\theta\coth(\theta) \geq 1$ for $\theta \geq 0$, it follows $c(r_*,\kappa_0) \geq 1$.

Main theorem: A simple calculation shows that the update rule (7) can be expressed as

$$\boldsymbol{X}(\tau+1) = \exp_{\boldsymbol{X}(\tau)}\left(\eta(\boldsymbol{X}(\tau))\boldsymbol{g}(\boldsymbol{X}(\tau))\right),$$

where $\eta(\boldsymbol{X}) := 1/\sum_{i=1}^n a_i\phi\left(\frac{\mathrm{dist}^2(\boldsymbol{X},\boldsymbol{X}_i)}{\sigma^2}\right)$. This expression clearly indicates that the update rule (7) performs Riemannian gradient ascent at $\boldsymbol{X}(\tau)$ if $\eta(\boldsymbol{X}(\tau)) \geq 0$. However, this does not ensure that the iterative rule (7) has the monotonic ascending property as well as $\{\boldsymbol{X}(\tau)\}_{\tau=0,1,\ldots}$ converges to a mode of the gradient model because the step size $\eta(\boldsymbol{X})$ is adaptive to $\boldsymbol{X}$. Here, we establish sufficient conditions for monotonic ascending in the following theorem:

Theorem 3. *Assume that (C1) all of data samples are contained in the ball $B(\boldsymbol{X}_o, \frac{1}{3}r_*)$ (i.e., $\{\boldsymbol{X}_i\}_{i=1}^n \subset B(\boldsymbol{X}_o, \frac{1}{3}r_*)$ and $\boldsymbol{X}(0) \in B(\boldsymbol{X}_o, \frac{1}{3}r_*)$, (C2) $c(\frac{4r_*}{3}, \kappa_0) < 4$, (C3) all coefficients a_i are nonnegative and (C4) there exists a finite, convex and monotonically decreasing function $\varphi(t)$ such that $\phi(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\varphi(t)$, and $\varphi(t)$ is bounded for $t \geq 0$. Then, under the update rule (7) and the general Riemannian metric $\langle\cdot,\cdot\rangle_{\boldsymbol{X}}$,*

$$D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] \geq 0, \tag{21}$$

and the sequence $\{D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)]\}_{\tau=0,1,\ldots}$ converges to zero.

The proof is given in Section E. Compared with Theorem 1, Theorem 3 has pros and cons. First of all, Theorem 3 does not focus on particular manifolds and thus has a high generality, while Theorem 1 is intended for particular matrix manifolds. On the other hand, the generality of Theorem 3 might involve some restriction: Assumption (C1) requires data samples to exist inside a (possibly small) ball. However, Theorem 1 does not have such a restriction to data samples.

Our proof is based on a previous work (Afsari et al., 2013) whose purpose is to establish convergence conditions of the Riemannian gradient decent with a fixed step size in the Fréchet mean estimation and where a variety of convergence conditions are established. In fact, Assumptions (C1-2) come from Theorem 4.1 in Afsari et al. (2013). Thus, by adopting other convergence results in Afsari et al. (2013), we could establish different sufficient conditions with small modification in our proof.

Unlike Theorem 2, the convergence of $\{\boldsymbol{X}(\tau)\}_{\tau=0,1,}$ is not established. However, thanks to the *global convergence theorem* in Luenberger and Ye (2008), if $\{\boldsymbol{X}(\tau)\}_{\tau=0,1,}$ is contained in a compact set, there would exist a subsequence of $\{\boldsymbol{X}(\tau)\}_{\tau=0,1,}$, which converges to a zero of the Riemannian gradient model with monotonic ascending (i.e., possibly, a mode based on the gradient model).

5 Robustness in mode estimation

This section performs the breakdown point analysis to investigate the robustness of the iterative methods in mode estimation. The proof of Theorem 3 in Section E shows that (20) gives an unnnormalized density model from the Riemannian gradient model as follows: With some fixed point $\boldsymbol{X}_o$, the line integral on a curve between $\boldsymbol{X}$ and $\boldsymbol{X}_o$ is computed as

$$D_{\boldsymbol{g}}[\boldsymbol{X}|\boldsymbol{X}_o] = \frac{\sigma^2}{2}\sum_{i=1}^n a_i\varphi\left(\frac{\mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{X}_i)}{\sigma^2}\right) - C_o, \tag{22}$$

where $C_o := \frac{\sigma^2}{2}\sum_{i=1}^n a_i\varphi\left(\frac{\mathrm{dist}^2(\boldsymbol{X}_o,\boldsymbol{X}_i)}{\sigma^2}\right)$ can be seen as a constant. Eqs.(19) and (22) indicate that $\sum_{i=1}^n a_i\varphi\left(\frac{\mathrm{dist}^2(\boldsymbol{X},\boldsymbol{X}_i)}{\sigma^2}\right)$ is a model for the log-density on $\mathbb{M}$ up to the normalizing constant. Then, we can define the mode of the model as

$$\boldsymbol{M} := \underset{\boldsymbol{X}\in\mathbb{M}}{\mathrm{argmax}}\left[\sum_{i=1}^n a_i\varphi\left(\frac{\mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{X}_i)}{\sigma^2}\right)\right].$$

In the finite breakdown point analysis, we add m arbitrary data samples $\mathcal{D}' := \{\boldsymbol{X}'_i\}_{i=1}^m$ into the original samples $\mathcal{D} = \{\boldsymbol{X}_i\}_{i=1}^n$, and consider the following definition of the mode from $\mathcal{D}\cup\mathcal{D}'$:

$$\boldsymbol{M}' := \underset{\boldsymbol{X}\in\mathbb{M}}{\mathrm{argmax}}\left[\sum_{i=1}^{n+m} a_i\varphi\left(\frac{\mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{Z}_i)}{\sigma^2}\right)\right],$$

where $\boldsymbol{Z}_i \in \mathcal{D}\cup\mathcal{D}'$[1]. The *finite breakdown point* (Huber and Ronchetti, 2009) is defined as

$$\epsilon(\boldsymbol{M}, \mathcal{D}) := \min_{1\leq m\leq n}\left\{\frac{m}{n+m}\,\middle|\, \sup_{\mathcal{D}'}\mathrm{dist}(\boldsymbol{M}, \boldsymbol{M}') = \infty\right\},$$

where sup is taken over all $\mathcal{D}'$. A larger value of the breakdown point implies more robust to the contamination of $\{\boldsymbol{X}'_i\}_{i=1}^m$. The following theorem shows the finite breakdown point where $\lceil R\rceil$ (or $\lfloor R\rfloor$) denotes the smallest (or largest) integer larger (or smaller) than R:

Theorem 4. *Assume that $0 \leq a_i \leq a_{\max}$ for some constant $a_{\max} > 0$ and all $i = 1, 2, \ldots, n+m$, and $\varphi(t)$ for $t \geq 0$ is nonnegative and reaches the maximum at $t = 0$. Let $R := \sum_{i=1}^n \tilde{a}_i\varphi^*\left(\frac{\mathrm{dist}^2(\boldsymbol{M},\boldsymbol{X}_i)}{\sigma^2}\right)$ where $\tilde{a}_i := a_i/a_{\max}$ and $\varphi^*(t) := \varphi(t)/\varphi(0)$. Then, the finite breakdown point is given by*

$$\epsilon^*(\boldsymbol{M}, \mathcal{D}) = \frac{m^*}{n+m^*}, \tag{23}$$

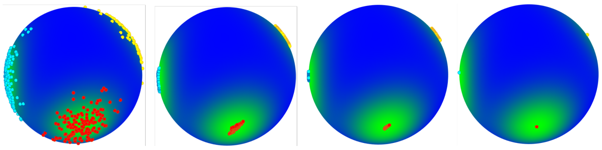
where with $\bar{A}'_m := \frac{1}{m}\sum_{i=n+1}^{n+m}\tilde{a}_i$, m^* satisfies

$$\lceil R\rceil \leq m^* \leq \left\lfloor\frac{R}{\bar{A}'_m}\right\rfloor + 1. \tag{24}$$

The proof is given in Section F. Theorem 4 shows that the break down point depends on a_i, σ and ϕ. For instance, when σ is large such that $\varphi^*\left(\frac{\mathrm{dist}^2(\boldsymbol{X},\boldsymbol{Z}_i)}{\sigma^2}\right) \approx 1$ (e.g., the Gauss kernel), then $R \approx \sum_{i=1}^n \tilde{a}_i \leq n$. From the lower-bound of m^* in (24), $R = n$ yields $\epsilon^*(\boldsymbol{M}, \mathcal{D}) \geq \frac{1}{2}$, implying our model potentially has a satisfactory robustness property.

Unlike the lower-bound, the upper-bound of m^* in (24) depends on $\bar{A}'_m$, which is the average of the coefficients

[1] $\boldsymbol{Z}_i = \boldsymbol{X}_i$ if $i \leq n$. For $i \geq n+1$, $\boldsymbol{Z}_i = \boldsymbol{X}'_i$.



(a) Iter.=0 (b) Iter.=5 (c) Iter.=10 (d) Iter.=100

Figure 1: Illustration of a mode-seeking process on Ω_{d_1-1}. Iter. means the number of iterations.

$\tilde{a}_i$ corresponding to $\boldsymbol{Z}_{i+n} = \boldsymbol{X}'_i$. This coefficient dependency to $\bar{A}'_m$ intuitively makes sense because m^* can be large when $\bar{A}'_m$ is close to zero, i.e., the coefficients $\tilde{a}_i$ for all $i = n+1, \ldots, n+m$ are approximately zeros and thus there is no influence from $\boldsymbol{X}'_i$. On the other hand, if the coefficients $\tilde{a}_i$ are large, the finite breakdown point could be small.

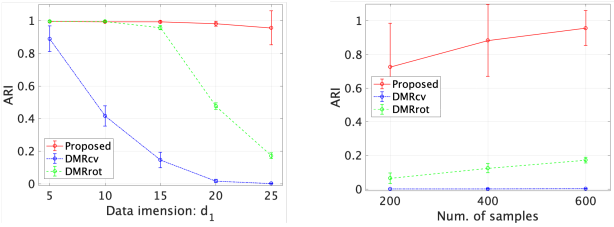
From the definition of the breakdown point, Theorem 4 is useful when $\mathbb{M}$ has an unbounded distance $\mathrm{dist}(\cdot,\cdot)$. For instance, $\mathrm{S}^+(d)$ has an unbounded distance, while the distance of $\mathrm{St}(d_1, d_2)$ should be bounded. In addition, we also note that with the almost same proof as Theorem 4, essentially the same result for the breakdown point holds for the model (17) with the Frobenuous norm, which is intended for $\mathrm{S}^+(d)$. In fact, we experimentally show the robustness of the proposed method on $\mathrm{S}^+(d)$ in Section 6.

We excluded the learning factor to the coefficients a_i in this analysis but should take it into account. However, it requires more sophisticated analysis to solve this problem. Thus, we leave this interesting but challenging problem for the future.

6 Numerical illustration

This section numerically investigates the proposed methods. As an example of $\mathrm{St}(d_1, d_2)$, mode-seeking clustering for directional data is first performed. Then, we compare our method with (7) in Ashizawa et al. (2017) on $\mathrm{Gr}(d_1, d_2)$. The robustness of the proposed method against outliers is demonstrated for $\mathrm{S}^+(d)$. Finally, our method on $\mathrm{S}^+(d)$ is applied to EEG data. All details of Riemannian gradient estimation and experimental settings are deferred to Sections G and H, respectively.

Clustering for directional data: Here, we perform mode-seeking clustering: In mode-seeking clustering, all data samples are initially regarded as the candidates for cluster centers, and updated towards local modes (Figs.1(a-d)). Then, the data samples which converged to the same mode are assigned the same cluster label. We apply the proposed iterative method (13) to mode-seeking clustering and compare it with directional mean shift (DMR) (Zhang and Chen, 2020). Since DMR is based on KDE, we used the two bandwidth selection methods: The bandwidth parameter was selected based on the *rule of thumb* for directional data (García-Portugués, 2013) (**DMRrot**) or the cross-validation based on the log-likelihood (**DMRcv**). Data was sampled from a mixture of three Von Mises densities (Fig.1(a)). The performance was measured by adjusted Rand index (ARI) (Hubert and Arabie, 1985): A larger value of ARI means a better clustering result.



(a) $n = 600$ (b) $d_1 = 25$

Figure 2: Clustering performance on directional data. Each point and error bar denote the average and standard deviation of ARI over 30 runs, respectively.

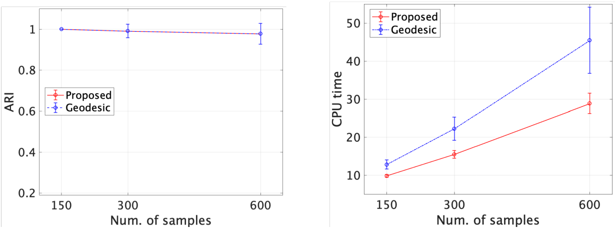


Figure 3: Comparison to an existing method (**Geodesic** (Ashizawa et al., 2017)) in terms of ARI and CPU time on $\mathrm{Gr}(d_1, d_2)$ $(d_1 = 9, d_2 = 2)$.

Fig.2(a) indicates that the proposed method works better than DMRs for higher-dimensional data: The ARI values of both DMRrot and DMRcv decrease as d_1 increases, the proposed method keeps the ARI values high among all data dimensions. Fig.2(b) shows that the proposed method produces high ARI values over various numbers of data samples.

Clustering on $\mathrm{Gr}(d_1, d_2)$**:** Next, we compare our method (16) with the existing method (7) in mode-seeking clustering on $\mathrm{Gr}(d_1, d_2)$. We followed exactly the same experimental setting in Ashizawa et al. (2017)[2]. Fig.3 indicates that the two methods show almost the same performance in terms of ARI. However,

[2] https://t-sakai-kure.github.io/software-ja.html

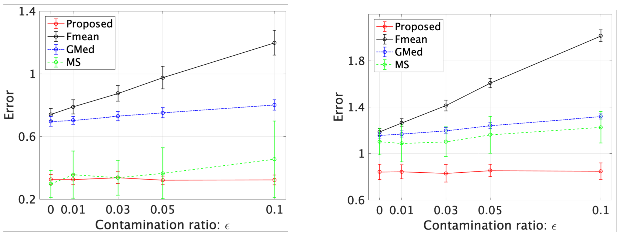


Figure 4: Robustness against outliers on $\mathrm{S}^+(d)$. The data dimensions in the left and right figures are $d = 3$ and $d = 7$, respectively.

the clear difference can be seen in the computational time: The proposed method is computationally more efficient than the existing method (Ashizawa et al., 2017). This would come from the fact that Ashizawa et al. (2017) perform the singular value decomposition to a d_1 by d_2 matrix in the exponential map $\exp_{\boldsymbol{X}}(\cdot)$ of the update rule (7) at each iterate and data sample (Section H.2). Thus, when the numbers of data samples and data dimensions are high, it would be computationally expensive.

Outlier robustness on $\mathrm{S}^+(d)$**:** Here, we investigate outlier robustness of the proposed method (18) on $\mathrm{S}^+(d)$, and compare it with the *Fréchet mean (FMean)* (Karcher, 1977), *geometric median (GMed)* (Fletcher et al., 2009) and *mean shift (MS)*. Symmetric positive definite matrices $\boldsymbol{X}_i$ were sampled from $\boldsymbol{X}_i = \boldsymbol{B}^\top \boldsymbol{B} + \mathrm{diag}(\boldsymbol{\beta}_i)$, where $\boldsymbol{B}$ is a d by d random matrix drawn from the normal density, $\mathrm{diag}(\boldsymbol{\beta}_i)$ is the diagonal matrix with the elements of $\boldsymbol{\beta}_i$ on the diagonal, and each element in $\boldsymbol{\beta}_i$ was independently sampled from a *contaminated* exponential density as $(1-\epsilon)\mu^{-1}e^{\beta/\mu} + \epsilon\mu_o^{-1}e^{(\beta+5)/\mu_o}$, where ϵ denotes the *outlier ratio*. Samples from the exponential density $\mu_o^{-1}e^{(\beta+5)/\mu_o}$ can be regarded as outliers. Here, we set $\mu = 0.5$ and $\mu_o = 0.1$. The total number of samples was $n = 500$. The performance error was defined by $\|\boldsymbol{B}^\top\boldsymbol{B} - \widehat{\boldsymbol{M}}\|_{\mathrm{F}}$, where $\widehat{\boldsymbol{M}}$ denotes an estimated mode.

Fig.4 clearly shows that the proposed method is robust against outliers. As previously mentioned, FMean is sensitive to outliers and the performance is worsened when the contamination ratio ϵ is high. The performance of GMed is not so good. A possible reason is that as in the Euclidean median, GMed would assume that the underlying density is symmetric, while the density in this experiment is highly skewed by outliers. MS is also robust against outliers, but the proposed method is clearly better than MS when the data dimension gets larger.

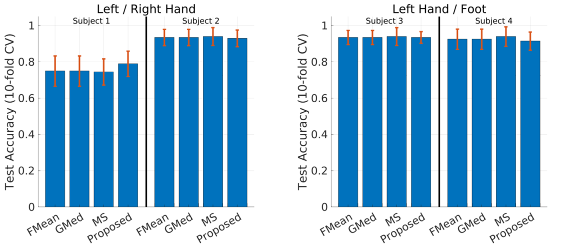


Figure 5: Cross-validated classification accuracy (mean ± SD) by different methods for tangency points.

Application to EEG data We applied our method to electroencephalograpy (EEG) data, publicly available from the BCI competition IV [3] (Dataset 1, calibration data). After basic preprocessing (see the supplementary material), we obtained 200 covariance matrices per subject, each from a single task trial of two-class motor imagery. Following Sabbagh et al. (2019), we first computed their Riemannian (geometric) projections onto a tangent space $T_{\bar{\mathbf{X}}}\mathbb{M}$ at the Fréchet mean $\bar{\mathbf{X}}$, so that they could be associated linearly with log-variances of EEG sources. Then, we applied L_2-regularized linear logistic regression and evaluated the classification accuracy of the two classes of motor imagery using a 10-fold cross-validation (CV) scheme, with a nested CV to optimize the L_2-penalty. Here, we attempted to replace the Fréchet mean as a reference point with its robust estimates considered above.

The result confirms that the proposed method performs at least comparable with the original nonrobust mean (KMean) as well GMed and MS (Fig. 5).

7 Conclusion

This paper proposed practical methods to estimate the modes on four matrix manifolds based on the Euclidean metric: Stiefel, oblique, Grassmann manifolds and the set of symmetric positive definite matrices. The key contribution is that the convergence of the proposed methods is theoretically guaranteed with monotonic ascending. In addition, we established sufficient conditions for the monotonic ascending property of an existing method (Ashizawa et al., 2017). Thus, our work can be also seen as compensating for the lack of theoretical analysis of the existing method. Furthermore, we performed the finite-breakdown point analysis to investigate the robustness of the mode estimation methods. Finally, we numerically demonstrated the usefulness of the proposed methods in clustering and robust mode estimation. In future, we will extend this approach based on the Euclidean metric to other matrix manifolds.

[3] http://www.bbci.de/competition/iv/

Acknowledgements

The authors thank the anonymous reviewers for their comments and suggestions, and also thank Tomoya Sakai, Reinmar Kobler and Motoaki Kawanabe for helpful discussions. TK was partially supported by JSPS KAKENHI Grant Number 17H00764, 19H04071, and 20H00576. JH was partially supported by JSPS KAKENHI Grant Number 21K12055.

References

P.-A. Absil and K. A. Gallivan. Joint diagonalization on the oblique manifold for independent component analysis. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 5. IEEE, 2006.

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

B. Afsari, R. Tron, and R. Vidal. On the convergence of gradient descent for finding the Riemannian center of mass. *SIAM Journal on Control and Optimization*, 51(3):2230–2260, 2013.

M. Arnaudon, F. Barbaresco, and L. Yang. Riemannian medians and means with applications to radar signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 7(4):595–604, 2013.

M. Ashizawa, H. Sasaki, T. Sakai, and M. Sugiyama. Least-squares log-density gradient clustering for Riemannian manifolds. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 537–546, 2017.

N. Boumal. *An introduction to optimization on smooth manifolds*. 2020.

N. Boumal, B. Mishra, P. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15: 1455–1459, 2014. URL `http://www.manopt.org`.

R. Caseiro, J. F. Henriques, P. Martins, and J. Batista. Semi-intrinsic mean shift on Riemannian manifolds. In *Proceedings of European conference on computer vision (ECCV)*, pages 342–355. Springer, 2012.

H. E. Cetingul and R. Vidal. Intrinsic mean shift for clustering on Stiefel and Grassmann manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1896–1902. IEEE, 2009.

Y.-C. Chen, C. Genovese, R. Tibshirani, and L. Wasserman. Nonparametric modal regression. *The Annals of Statistics*, 44(2):489–514, 2016.

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.

D. D. Cox. A penalty method for nonparametric estimation of the logarithmic derivative of a density function. *Annals of the Institute of Statistical Mathematics*, 37(1):271–288, 1985.

I. L. Dryden, A. Koloydenko, and D. Zhou. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102–1123, 2009.

Y. Feng, J. Fan, and J. A. Suykens. A statistical learning approach to modal regression. *Journal of Machine Learning Research*, 21(2):1–35, 2020.

P. T. Fletcher, S. Venkatasubramanian, and S. Joshi. The geometric median on Riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(1):S143–S152, 2009.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

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

P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley, 2009.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

M. Kafai, Y. Miao, and K. Okada. Directional mean shift and its application for topology classification of local 3D structures. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 170–177. IEEE, 2010.

T. Kanamori, T. Suzuki, and M. Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012.

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

T. Kobayashi and N. Otsu. Von Mises-Fisher mean shift for clustering on a hypersphere. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, pages 2130–2133. IEEE, 2010.

J. Lee. *Introduction to Smooth Manifolds*. Springer, 2012.

J. M. Lee. *Introduction to Riemannian manifolds*. Springer, 2018.

M.-J. Lee. Mode regression. *Journal of Econometrics*, 42(3):337–349, 1989.

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

D. G. Luenberger and Y. Ye. *Linear and Nonlinear Programming*. Springer, 2008.

Y. M. Lui. Advances in matrix manifolds for computer vision. *Image and Vision Computing*, 30(6-7):380–388, 2012.

K. V. Mardia. *Statistics of directional data*. Academic press, 1972.

P. Petersen. *Riemannian geometry*. Springer, 2006.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1657–1665, 2015.

D. Sabbagh, P. Ablin, G. Varoquaux, A. Gramfort, and D. A. Engemann. Manifold-regression to predict from MEG/EEG brain signals without source modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

T. W. Sager and R. A. Thisted. Maximum likelihood estimation of isotonic modal regression. *The Annals of Statistics*, 10(3):690–707, 1982.

H. Sasaki, A. Hyvärinen, and M. Sugiyama. Clustering via mode seeking by direct estimation of the gradient of a log-density. In *Machine Learning and Knowledge Discovery in Databases Part III- European Conference, ECML/PKDD 2014*, volume 8726, pages 19–34, 2014.

H. Sasaki, T. Sakai, and T. Kanamori. Robust modal regression with direct gradient approximation of modal regression risk. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124, pages 380–389. PMLR, 2020.

R. Slama, H. Wannous, M. Daoudi, and A. Srivastava. Accurate 3D action recognition using learning on the grassmann manifold. *Pattern Recognition*, 48 (2):556–567, 2015.

R. Subbarao and P. Meer. Nonlinear mean shift for clustering over analytic manifolds. In *Proceedings of IEEE Computer Society Conference on of Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 1168–1175, 2006.

R. Subbarao and P. Meer. Nonlinear mean shift over Riemannian manifolds. *International Journal of Computer Vision*, 84(1):1–20, 2009.

P. K. Turaga and A. Srivastava. *Riemannian computing in computer vision*. Springer, 2016.

O. Tuzel, R. Subbarao, and P. Meer. Simultaneous multiple 3D motion estimation via mode finding on lie groups. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 18–25. IEEE, 2005.

B. Vandereycken, P.-A. Absil, and S. Vandewalle. Embedded geometry of the set of symmetric positive semidefinite matrices of fixed rank. In *IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 389–392. IEEE, 2009.

W. Yao and L. Li. A new regression model: modal linear regression. *Scandinavian Journal of Statistics*, 41(3):656–671, 2014.

F. Yger, M. Berar, and F. Lotte. Riemannian approaches in brain-computer interfaces: a review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1753–1762, 2016.

Y. Zhang and Y.-C. Chen. Kernel smoothing, mean shift, and their learning theory with directional data. *arXiv preprint arXiv:2010.13523*, 2020.

Y. Zhang and Y.-C. Chen. The EM perspective of directional mean shift algorithm. *arXiv preprint arXiv:2101.10058*, 2021.

Supplementary Material:
Mode estimation on matrix manifolds: Convergence and robustness

Without loss of generality, we suppose that $\sigma = 1$ in Sections B, D, E and F.

A Derivation of (14) for the oblique manifold $\mathrm{Ob}(d_1, d_2)$

For the oblique manifold $\mathrm{Ob}(d_1, d_2)$, we formulate the Lagrangian function on $\mathbb{R}^{d_1 \times d_2}$ as

$$\log p(\boldsymbol{X}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}^\top(\mathrm{ddiag}(\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{I}_{d_2})) \quad (\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}),$$

where $\boldsymbol{\Lambda}$ is the d_2 by d_2 diagonal matrix of Lagrange multipliers. Then, the optimality conditions, which the modes satisfy, yield

$$\nabla_{\boldsymbol{X}} \log p(\boldsymbol{X}) - \boldsymbol{X}\boldsymbol{\Lambda} = \boldsymbol{O} \tag{25}$$
$$\mathrm{ddiag}(\boldsymbol{X}^\top\boldsymbol{X}) = \boldsymbol{I}_{d_2}. \tag{26}$$

Next, we substitute the trace model (11) for $\nabla_{\boldsymbol{X}} \log p(\boldsymbol{X})$ in (25) and obtain

$$\boldsymbol{g}_{\mathrm{e}}(\boldsymbol{X}) - \boldsymbol{X}\boldsymbol{\Lambda} = \sum_{i=1}^{n} a_i \boldsymbol{X}_i \phi\left(\frac{\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X}_i)}{\sigma^2}\right) - \boldsymbol{X}\boldsymbol{\Lambda} = \boldsymbol{O}.$$

Based on the fixed-point method, the following naive update rule can be derived as

$$\boldsymbol{X} \leftarrow \sum_{i=1}^{n} a_i \boldsymbol{X}_i \phi\left(\frac{\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X}_i)}{\sigma^2}\right)\boldsymbol{\Lambda}^{-1} = \boldsymbol{Z}\boldsymbol{\Lambda}^{-1},$$

where $\boldsymbol{Z} := \sum_{i=1}^{n} a_i \boldsymbol{X}_i \phi\left(\frac{\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X}_i)}{\sigma^2}\right)$. In order to satisfy (26), we compute

$$\mathrm{ddiag}(\boldsymbol{X}^\top\boldsymbol{X}) = \mathrm{ddiag}(\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{\Lambda}^{-1}) = \boldsymbol{\Lambda}^{-1}\mathrm{ddiag}(\boldsymbol{Z}^\top\boldsymbol{Z})\boldsymbol{\Lambda}^{-1},$$

where we used the fact that $\boldsymbol{\Lambda}$ is a diagonal matrix. Thus, (26) yields

$$\boldsymbol{\Lambda} = \mathrm{ddiag}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-\frac{1}{2}},$$

and we obtain the final update rule as

$$\boldsymbol{X}(\tau+1) = \boldsymbol{Z}(\tau)(\mathrm{ddiag}(\boldsymbol{Z}(\tau)^\top\boldsymbol{Z}(\tau)))^{-\frac{1}{2}},$$

where $\boldsymbol{Z}(\tau) := \sum_{i=1}^{n} a_i \boldsymbol{X}_i \phi\left(\frac{\mathrm{tr}(\boldsymbol{X}(\tau)^\top\boldsymbol{X}_i)}{\sigma^2}\right)$.

B Proof of Theorem 1

This section gives the proofs on $\mathrm{St}(d_1, d_2)$, $\mathrm{Gr}(d_1, d_2)$ and $\mathrm{S}^+(d)$. The proof of $\mathrm{Ob}(d_1, d_2)$ is omitted because it is the almost same as $\mathrm{St}(d_1, d_2)$ using the orthogonal projector onto the tangent space $T\mathrm{Ob}(d_1, d_2)$ as $\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{X}\mathrm{ddiag}(\boldsymbol{X}^\top\boldsymbol{Z})$ (Absil and Gallivan, 2006).

B.1 Proof for the Stiefel manifold $\mathrm{St}(d_1, d_2)$

Proof. Before going to the line integral (20), we show a simple relation on $\mathrm{St}(d_1, d_2)$. Since $\boldsymbol{C}(t) \in \mathrm{St}(d_1, d_2)$, it holds $\boldsymbol{C}(t)^\top \boldsymbol{C}(t) = \boldsymbol{I}_{d_2}$. By differentiating it with respect to t, we obtain

$$\dot{\boldsymbol{C}}(t)^\top \boldsymbol{C}(t) + \boldsymbol{C}(t)^\top \dot{\boldsymbol{C}}(t) = \boldsymbol{O}. \tag{27}$$

Then, we have a model for the Riemannian gradient as

$$\boldsymbol{g}(\boldsymbol{X}) := \boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{g}_{\mathrm{e}}(\boldsymbol{X})) = \sum_{i=1}^{n} a_i \boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{X}_i)\phi(\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X}_i)) = \sum_{i=1}^{n} a_i(\boldsymbol{X}_i - \boldsymbol{X}\mathrm{sym}(\boldsymbol{X}^\top \boldsymbol{X}_i))\phi(\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X}_i)), \tag{28}$$

where $\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{X}\mathrm{sym}(\boldsymbol{X}^\top \boldsymbol{Z})$ on $\mathrm{St}(d_1, d_2)$ (Absil et al., 2008). By substituting $\boldsymbol{g}(\boldsymbol{X})$ into the line integral (20) under the Euclidean metric $\langle \boldsymbol{Y}, \boldsymbol{Z} \rangle_{\boldsymbol{X}} = \mathrm{tr}(\boldsymbol{Y}^\top \boldsymbol{Z})$, we have

$$\begin{aligned} D_{\boldsymbol{g}}[\boldsymbol{Y}|\boldsymbol{X}] &= \int_0^1 \left[\sum_{i=1}^{n} a_i \mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top \boldsymbol{X}_i)\phi(\mathrm{tr}(\boldsymbol{C}(t)^\top \boldsymbol{X}_i)) - \sum_{i=1}^{n} a_i \mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top \boldsymbol{C}(t)\mathrm{sym}(\boldsymbol{C}(t)^\top \boldsymbol{X}_i))\phi(\mathrm{tr}(\boldsymbol{C}(t)^\top \boldsymbol{X}_i)) \right] \mathrm{d}t \\ &= \int_0^1 \sum_{i=1}^{n} a_i \mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top \boldsymbol{X}_i)\phi(\mathrm{tr}(\boldsymbol{C}(t)^\top \boldsymbol{X}_i))\mathrm{d}t \\ &= \sum_{i=1}^{n} a_i \left\{ \varphi(\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X}_i)) - \varphi(\mathrm{tr}(\boldsymbol{Y}^\top \boldsymbol{X}_i)) \right\}, \end{aligned} \tag{29}$$

where we derived and used the following relation:

$$\begin{aligned} 2\mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top \boldsymbol{C}(t)\mathrm{sym}(\boldsymbol{C}(t)^\top \boldsymbol{X}_i)) &= \mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top \boldsymbol{C}(t)\mathrm{sym}(\boldsymbol{C}(t)^\top \boldsymbol{X}_i)) + \mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top \boldsymbol{C}(t)\mathrm{sym}(\boldsymbol{C}(t)^\top \boldsymbol{X}_i)) \\ &= \mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top \boldsymbol{C}(t)\mathrm{sym}(\boldsymbol{C}(t)^\top \boldsymbol{X}_i)) + \mathrm{tr}(\boldsymbol{C}(t)^\top \dot{\boldsymbol{C}}(t)\mathrm{sym}(\boldsymbol{C}(t)^\top \boldsymbol{X}_i)) \\ &= \mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top \boldsymbol{C}(t)\mathrm{sym}(\boldsymbol{C}(t)^\top \boldsymbol{X}_i)) - \mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top \boldsymbol{C}(t)\mathrm{sym}(\boldsymbol{C}(t)^\top \boldsymbol{X}_i)) \\ &= 0, \end{aligned}$$

where we used the trace property as $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B}^\top)$ on the second line and applied (27) on the third line.

Next, we substitute $\boldsymbol{X}(\tau)$ and $\boldsymbol{X}(\tau+1)$ into $\boldsymbol{Y}$ and $\boldsymbol{X}$ respectively, and express the right-hand side on (29) as

$$\begin{aligned} D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] &= \sum_{i=1}^{n} a_i \varphi(\mathrm{tr}(\boldsymbol{X}(\tau+1)^\top \boldsymbol{X}_i)) - \sum_{i=1}^{n} a_i \varphi(\mathrm{tr}(\boldsymbol{X}(\tau)^\top \boldsymbol{X}_i)) \\ &\geq \sum_{i=1}^{n} a_i \phi(\mathrm{tr}(\boldsymbol{X}(\tau)^\top \boldsymbol{X}_i))\{\mathrm{tr}(\boldsymbol{X}(\tau+1)^\top \boldsymbol{X}_i) - \mathrm{tr}(\boldsymbol{X}(\tau)^\top \boldsymbol{X}_i)\}, \end{aligned} \tag{30}$$

where we applied a well-known inequality to the convex function $\varphi(\cdot)$: For a convex function $f(t)$ and $\dot{f}(t) = \frac{\mathrm{d}}{\mathrm{d}t}f(t)$,

$$f(t_x) - f(t_y) \geq \dot{f}(t_y)(t_x - t_y) \qquad (t_x, t_y \in \mathbb{R}). \tag{31}$$

Then, we employ the update rule (12) to the right-hand side of (30) and obtain

$$\begin{aligned} D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] &\geq \mathrm{tr}\left(\boldsymbol{X}(\tau+1)^\top \boldsymbol{Z}(\tau)\right) - \mathrm{tr}\left(\boldsymbol{X}(\tau)^\top \boldsymbol{Z}(\tau)\right) \\ &= \mathrm{tr}\left(\boldsymbol{X}(\tau+1)^\top \boldsymbol{X}(\tau+1)(\boldsymbol{Z}(\tau)^\top \boldsymbol{Z}(\tau))^{1/2}\right) - \mathrm{tr}\left(\boldsymbol{X}(\tau)^\top \boldsymbol{X}(\tau+1)(\boldsymbol{Z}(\tau)^\top \boldsymbol{Z}(\tau))^{1/2}\right), \end{aligned} \tag{32}$$

where we recall that $\boldsymbol{Z}(\tau) = \sum_{i=1}^{n} a_i \boldsymbol{X}_i \phi(\mathrm{tr}(\boldsymbol{X}(\tau)^\top \boldsymbol{X}_i))$. Since $\boldsymbol{X}(\tau+1)^\top \boldsymbol{X}(\tau+1) = \boldsymbol{X}(\tau)^\top \boldsymbol{X}(\tau) = \boldsymbol{I}_{d_2}$, we have

$$\mathrm{tr}\left(\boldsymbol{X}(\tau+1)^\top \boldsymbol{X}(\tau+1)(\boldsymbol{Z}(\tau)^\top \boldsymbol{Z}(\tau))^{1/2}\right) = \mathrm{tr}\left(\boldsymbol{X}(\tau)^\top \boldsymbol{X}(\tau)(\boldsymbol{Z}(\tau)^\top \boldsymbol{Z}(\tau))^{1/2}\right),$$

which further modifies the right-hand on (32) as

$$D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] \geq \frac{1}{2}\mathrm{tr}\left((\boldsymbol{X}(\tau+1)-\boldsymbol{X}(\tau))^\top(\boldsymbol{X}(\tau+1)-\boldsymbol{X}(\tau))(\boldsymbol{Z}(\tau)^\top\boldsymbol{Z}(\tau))^{1/2}\right). \tag{33}$$

Since $(\boldsymbol{X}(\tau+1)-\boldsymbol{X}(\tau))^\top(\boldsymbol{X}(\tau+1)-\boldsymbol{X}(\tau))$ and $\boldsymbol{Z}(\tau)^\top\boldsymbol{Z}(\tau)$ are positive semidefinite matrices, the right-hand side on (33) is nonnegative.

It follows from (29) and the monotonic ascending property that

$$D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] = \sum_{i=1}^n a_i\left\{\varphi(\mathrm{tr}(\boldsymbol{X}(\tau)^\top\boldsymbol{X}_i)) - \varphi(\mathrm{tr}(\boldsymbol{X}(\tau+1)^\top\boldsymbol{X}_i))\right\} \geq 0. \tag{34}$$

Function $\varphi(\mathrm{tr}(\boldsymbol{X}(\tau)^\top\boldsymbol{X}_i))$ for $\boldsymbol{X}(\tau), \boldsymbol{X}_i \in \mathrm{St}(d_1,d_2)$ is bounded because $\mathrm{St}(d_1,d_2)$ is compact. Thus, (34) means that the sequence $\{\sum_{i=1}^n a_i\varphi(\mathrm{tr}(\boldsymbol{X}^\top(\tau)\boldsymbol{X}_i))\}_{\tau=0,1,\ldots}$ is monotonically increasing and converges as τ increases, which implies $D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] \to 0$ as $\tau\to\infty$. The proof is completed. □

B.2 Proof for Grassmann manifold $\mathrm{Gr}(d_1,d_2)$

Proof. With the orthogonal projector $\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Z}$ onto $T_{\boldsymbol{X}}\mathrm{Gr}(d_1,d_2)$ (Absil et al., 2008), a model for the Riemannian model is given by

$$\boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{g}_{\mathrm{e}}(\boldsymbol{X})) = \sum_{i=1}^n a_i(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{X} - \boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{X})\phi_i^{\mathrm{Gr}}(\boldsymbol{X}), \tag{35}$$

where we recall that $\phi_i^{\mathrm{Gr}}(\boldsymbol{X}) = \phi\left(\mathrm{tr}(\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top)\right)$. Then, we compute the path integral (20) with a curve $\boldsymbol{C}(t)$ connecting $\boldsymbol{X}$ and $\boldsymbol{Y}$ as

$$\begin{aligned}
D_{\boldsymbol{g}}[\boldsymbol{Y}|\boldsymbol{X}] &= \int_0^1\sum_{i=1}^n a_i\mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{C}(t) - \dot{\boldsymbol{C}}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{C}(t))\phi_i^{\mathrm{Gr}}(\boldsymbol{C}(t))\mathrm{d}t \\
&= \sum_{i=1}^n a_i\int_0^1 \mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{C}(t))\phi_i^{\mathrm{Gr}}(\boldsymbol{C}(t))\mathrm{d}t \\
&= \sum_{i=1}^n \frac{a_i}{2}\left[\varphi_i^{\mathrm{Gr}}(\boldsymbol{X}) - \varphi_i^{\mathrm{Gr}}(\boldsymbol{Y})\right],
\end{aligned} \tag{36}$$

where $\varphi_i^{\mathrm{Gr}}(\boldsymbol{X}) := \varphi\left(\mathrm{tr}(\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top)\right)$, and we derived and used the following relation:

$$\begin{aligned}
2\mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{C}(t)) &= \mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{C}(t)) + \mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{C}(t)) \\
&= \mathrm{tr}(\dot{\boldsymbol{C}}(t)^\top\boldsymbol{C}(t)\boldsymbol{C}(t)^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{C}(t)) + \mathrm{tr}(\boldsymbol{C}(t)^\top\dot{\boldsymbol{C}}(t)\boldsymbol{C}(t)^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{C}(t)) \\
&= \mathrm{tr}((\dot{\boldsymbol{C}}(t)^\top\boldsymbol{C}(t) + \boldsymbol{C}(t)^\top\dot{\boldsymbol{C}}(t))\boldsymbol{C}(t)^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{C}(t)) \\
&= 0,
\end{aligned}$$

where we applied $\dot{\boldsymbol{C}}(t)^\top\boldsymbol{C}(t) + \boldsymbol{C}(t)^\top\dot{\boldsymbol{C}}(t) = \boldsymbol{O}$ in (27). By applying (31) for the convex function $\varphi(\cdot)$, a lower-bound of $D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)]$ is obtained as

$$\begin{aligned}
D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] &\geq \sum_{i=1}^n\frac{a_i}{2}\phi_i^{\mathrm{Gr}}(\boldsymbol{X}(\tau))\left[\mathrm{tr}(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{X}(\tau+1)\boldsymbol{X}(\tau+1)^\top) - \mathrm{tr}(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{X}(\tau)\boldsymbol{X}(\tau)^\top)\right] \\
&= \frac{1}{2}\mathrm{tr}(\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau+1)\boldsymbol{X}(\tau+1)^\top) - \frac{1}{2}\mathrm{tr}(\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau)\boldsymbol{X}(\tau)^\top) \\
&= \frac{1}{2}\mathrm{tr}(\boldsymbol{X}(\tau+1)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau+1)) - \frac{1}{2}\mathrm{tr}(\boldsymbol{X}(\tau)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau)).
\end{aligned} \tag{37}$$

Next, we need to show that the right-hand side on (37) is nonnegative. To this end, we express the right-hand side on (37) as

$$\begin{aligned}
&\frac{1}{2}\mathrm{tr}(\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau+1)\boldsymbol{X}(\tau+1)^\top)-\frac{1}{2}\mathrm{tr}(\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau)\boldsymbol{X}(\tau)^\top)\\
&=\frac{1}{2}\mathrm{tr}(\boldsymbol{X}(\tau+1)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau+1))+\frac{1}{2}\mathrm{tr}(\boldsymbol{X}(\tau)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau))-\mathrm{tr}(\boldsymbol{X}(\tau)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau))\\
&\geq\frac{1}{2}\mathrm{tr}(\boldsymbol{X}(\tau+1)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau+1))+\frac{1}{2}\mathrm{tr}(\boldsymbol{X}(\tau)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau))-\mathrm{tr}(\{\boldsymbol{X}(\tau)^\top\boldsymbol{Y}(\tau)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau)\}^{\frac{1}{2}}),
\end{aligned}\tag{38}$$

where we applied Assumption (A3) on the last line. Since $\boldsymbol{X}(\tau+1)$ is an orthogonal matrix, multiplying $\boldsymbol{X}(\tau+1)^\top$ to both sides of the update rule (16) yields

$$\boldsymbol{X}(\tau+1)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau)=(\boldsymbol{X}(\tau)^\top\boldsymbol{Y}(\tau)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau))^{\frac{1}{2}}.\tag{39}$$

Substituting (39) into the last term on the right-hand side of (38) gives

$$\begin{aligned}
&\frac{1}{2}\mathrm{tr}(\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau+1)\boldsymbol{X}(\tau+1)^\top)-\frac{1}{2}\mathrm{tr}(\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau)\boldsymbol{X}(\tau)^\top)\\
&\geq\frac{1}{2}\mathrm{tr}(\boldsymbol{X}(\tau+1)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau+1))+\frac{1}{2}\mathrm{tr}(\boldsymbol{X}(\tau)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau))-\mathrm{tr}(\boldsymbol{X}(\tau+1)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau))\\
&=\frac{1}{2}\mathrm{tr}(\boldsymbol{X}(\tau+1)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau+1))+\frac{1}{2}\mathrm{tr}(\boldsymbol{X}(\tau)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau))\\
&\qquad-\frac{1}{2}\mathrm{tr}(\boldsymbol{X}(\tau+1)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau))-\frac{1}{2}\mathrm{tr}(\boldsymbol{X}(\tau)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau+1))\\
&=\frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{X}(\tau+1)-\boldsymbol{X}(\tau))^\top\boldsymbol{Y}(\tau)(\boldsymbol{X}(\tau+1)-\boldsymbol{X}(\tau))\right\},
\end{aligned}\tag{40}$$

where we used $\mathrm{tr}(\boldsymbol{X}(\tau+1)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau))=\mathrm{tr}(\boldsymbol{X}(\tau)^\top\boldsymbol{Y}(\tau)\boldsymbol{X}(\tau+1))$ because $\boldsymbol{Y}(\tau)$ is symmetric. Since $\boldsymbol{Y}(\tau)$ is positive semidefinite, the right-hand side on (40) is nonnegative. The convergence of $\{D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)]\}_{\tau=0,1,\dots}$ can be confirmed by following the same step as $\mathrm{St}(d_1,d_2)$. Thus, the proof is completed.

□

B.3 Proof for the set of symmetric matrices $\mathrm{S}^+(d)$

Based on the norm model (17), our model for the Riemannian gradient on $\mathrm{S}^+(d)$ can be expressed as

$$\boldsymbol{g}(\boldsymbol{X})=\sum_{i=1}^n a_i\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{X}_i-\boldsymbol{X})\phi(\|\boldsymbol{X}-\boldsymbol{X}_i\|_{\mathrm{F}}^2)=\sum_{i=1}^n a_i\mathrm{sym}(\boldsymbol{X}_i-\boldsymbol{X})\phi(\|\boldsymbol{X}-\boldsymbol{X}_i\|_{\mathrm{F}}^2),\tag{41}$$

where $\boldsymbol{P}_{\boldsymbol{X}}(\cdot)=\mathrm{sym}(\cdot)$ is the orthogonal projector onto the tangent space $T_{\boldsymbol{X}}\mathrm{S}^+(d)$ (Vandereycken et al., 2009). Substituting $\boldsymbol{g}(\boldsymbol{X})$ into (20) yields

$$\begin{aligned}
D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)]&=\int_0^1\mathrm{tr}\left(\dot{\boldsymbol{C}}(t)^\top\boldsymbol{g}(\boldsymbol{C}(t))\right)\mathrm{d}t\\
&=-\sum_{i=1}^n a_i\int_0^1\mathrm{tr}\left(\dot{\boldsymbol{C}}(t)^\top\mathrm{sym}(\boldsymbol{C}(t)-\boldsymbol{X}_i)\right)\phi(\|\boldsymbol{C}(t)-\boldsymbol{X}_i\|_{\mathrm{F}}^2)\mathrm{d}t\\
&=\sum_{i=1}^n\frac{a_i}{2}\left\{\varphi(\|\boldsymbol{X}(\tau+1)-\boldsymbol{X}_i\|_{\mathrm{F}}^2)-\varphi(\|\boldsymbol{X}(\tau)-\boldsymbol{X}_i\|_{\mathrm{F}}^2)\right\},
\end{aligned}$$

where $\mathrm{sym}(\boldsymbol{C}(t) - \boldsymbol{X}_i) = \boldsymbol{C}(t) - \boldsymbol{X}_i$. Then, we apply the inequality (31) to $\varphi(\cdot)$ and have

$$\begin{aligned} D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] &\geq \sum_{i=1}^{n} \frac{a_i}{2}\phi(\|\boldsymbol{X}(\tau) - \boldsymbol{X}_i\|_{\mathrm{F}}^2)\left\{\|\boldsymbol{X}(\tau) - \boldsymbol{X}_i\|_{\mathrm{F}}^2 - \|\boldsymbol{X}(\tau+1) - \boldsymbol{X}_i\|_{\mathrm{F}}^2\right\} \\ &= \sum_{i=1}^{n} \frac{a_i}{2}\phi(\|\boldsymbol{X}(\tau) - \boldsymbol{X}_i\|_{\mathrm{F}}^2)\left\{\|\boldsymbol{X}(\tau)\|_{\mathrm{F}}^2 - 2\mathrm{tr}\left((\boldsymbol{X}(\tau) - \boldsymbol{X}(\tau+1))^\top \boldsymbol{X}_i\right) - \|\boldsymbol{X}(\tau+1)\|_{\mathrm{F}}^2\right\} \\ &= \sum_{i=1}^{n} \frac{a_i}{2}\phi(\|\boldsymbol{X}(\tau) - \boldsymbol{X}_i\|_{\mathrm{F}}^2)\left\{\|\boldsymbol{X}(\tau)\|_{\mathrm{F}}^2 - \|\boldsymbol{X}(\tau+1)\|_{\mathrm{F}}^2\right\} \\ &\qquad - \mathrm{tr}\left((\boldsymbol{X}(\tau) - \boldsymbol{X}(\tau+1))^\top \left\{\sum_{i=1}^{n} a_i \boldsymbol{X}_i \phi(\|\boldsymbol{X}(\tau) - \boldsymbol{X}_i\|_{\mathrm{F}}^2)\right\}\right). \end{aligned} \tag{42}$$

We next use the update rule (18) for the last term on the right-hand side of (42) as

$$\begin{aligned} &\mathrm{tr}\left((\boldsymbol{X}(\tau) - \boldsymbol{X}(\tau+1))^\top \left\{\sum_{i=1}^{n} a_i \boldsymbol{X}_i \phi(\|\boldsymbol{X}(\tau) - \boldsymbol{X}_i\|_{\mathrm{F}}^2)\right\}\right) \\ &\qquad = \sum_{i=1}^{n} a_i \phi(\|\boldsymbol{X}(\tau) - \boldsymbol{X}_i\|_{\mathrm{F}}^2)\mathrm{tr}\left((\boldsymbol{X}(\tau) - \boldsymbol{X}(\tau+1))^\top \boldsymbol{X}(\tau+1)\right). \end{aligned} \tag{43}$$

By substituting (43) into (42), we obtain

$$\begin{aligned} D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] &\geq \sum_{i=1}^{n} \frac{a_i}{2}\phi(\|\boldsymbol{X}(\tau) - \boldsymbol{X}_i\|_{\mathrm{F}}^2)\left\{\|\boldsymbol{X}(\tau)\|_{\mathrm{F}}^2 - 2\mathrm{tr}(\boldsymbol{X}(\tau)^\top \boldsymbol{X}(\tau+1)) + \|\boldsymbol{X}(\tau+1)\|_{\mathrm{F}}^2\right\} \\ &= \sum_{i=1}^{n} \frac{a_i}{2}\phi(\|\boldsymbol{X}(\tau) - \boldsymbol{X}_i\|_{\mathrm{F}}^2)\|\boldsymbol{X}(\tau) - \boldsymbol{X}(\tau+1)\|_{\mathrm{F}}^2. \end{aligned}$$

Since Assumptions (A1) and (A2) ensure that both a_i and $\phi(\cdot)$ are nonnegative, it holds $D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] \geq 0$. Following the same step as $\mathrm{St}(d_1, d_2)$ proves $D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] \to 0$ as $\tau \to \infty$ because $\varphi(t)$ is assumed to be bounded on $t \geq 0$. The proof is completed.

C Assumption (A3) in $d_2 = 1$

Here, we show that Assumption (A3) holds when $d_2 = 1$. To this end, we denote $\boldsymbol{X} \in \mathrm{Gr}(d_1, 1)$ by $\boldsymbol{x} \in \mathbb{R}^{d_1}$ where $\|\boldsymbol{x}\| = 1$. Then, Assumption (A3) is given by

$$\boldsymbol{x}^\top \boldsymbol{Y}(\tau)\boldsymbol{x} \leq (\boldsymbol{x}^\top \boldsymbol{Y}(\tau)^\top \boldsymbol{Y}(\tau)\boldsymbol{x})^{1/2}. \tag{44}$$

Since $\boldsymbol{Y}(\tau)$ is a symmetric positive definite matrix, the eigenvalue decomposition of $\boldsymbol{Y}(\tau)$ is given by $\boldsymbol{Y}(\tau) = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ where $\boldsymbol{\Lambda}$ denotes the diagonal matrix with nonnegative diagonals (i.e., eigenvalues) and $\boldsymbol{U}$ is a d_1 by d_1 orthogonal matrix. Inequality (44) can be equivalently expressed as

$$\sum_{i=1}^{d_1} \lambda_i (\boldsymbol{u}_i^\top \boldsymbol{x})^2 \leq \left(\sum_{i=1}^{d_1} \lambda_i^2 (\boldsymbol{u}_i^\top \boldsymbol{x})^2\right)^{1/2}, \tag{45}$$

where $\boldsymbol{u}_i$ and λ_i denote the i-th column vector in $\boldsymbol{U}$ and diagonal in $\boldsymbol{\Lambda}$, respectively. Since $\sum_{i=1}^{d_1} (\boldsymbol{u}_i^\top \boldsymbol{x})^2 = \|\boldsymbol{U}^\top \boldsymbol{x}\|^2 = \|\boldsymbol{x}\|^2 = 1$, applying Jensen's inequality assures that (45) as well as (44) hold.

D Proof of Theorem 2

Proof. Here, we follow the proof for Theorem 2 in Li et al. (2007). We only consider $\mathrm{St}(d_1, d_2)$ and $\mathrm{Gr}(d_1, d_2)$, and omit the proofs for $\mathrm{Ob}(d_1, d_2)$ and $\mathrm{S}^+(d)$ because they are essentially the same as $\mathrm{St}(d_1, d_2)$. By Assumption (B2), without loss of generality, we assume that there exist n zeros of the Riemannian gradient model $\boldsymbol{g}(\boldsymbol{X})$ such that

$$\boldsymbol{g}(\tilde{\boldsymbol{X}}_k) = \boldsymbol{0},\ k = 1, \ldots, n,$$

and define the minimum distance among $\tilde{\boldsymbol{X}}_k$ as

$$d_{\min} := \min\{\|\tilde{\boldsymbol{X}}_k - \tilde{\boldsymbol{X}}_{k'}\|_{\mathrm{F}} \mid 1 \le k \neq k' \le n\}.$$

As shown in the proof of Theorem 1, the sequence $\{\sum_{i=1}^n a_i \varphi(\mathrm{tr}(\boldsymbol{X}^\top(\tau)\boldsymbol{X}_i))\}_{\tau=0,1,\ldots}$ converges as τ increases. From (33) and (34), we have

$$\begin{aligned}&\sum_{i=1}^n a_i \left\{\varphi(\mathrm{tr}(\boldsymbol{X}(\tau+1)^\top \boldsymbol{X}_i)) - \varphi(\mathrm{tr}(\boldsymbol{X}(\tau)^\top \boldsymbol{X}_i))\right\}\\ &\ge \frac{1}{2}\mathrm{tr}\left((\boldsymbol{X}(\tau+1)-\boldsymbol{X}(\tau))^\top(\boldsymbol{X}(\tau+1)-\boldsymbol{X}(\tau))(\boldsymbol{Z}(\tau)^\top\boldsymbol{Z}(\tau))^{1/2}\right)\\ &\ge \frac{1}{2}\lambda_{\min}\|\boldsymbol{X}(\tau+1)-\boldsymbol{X}(\tau)\|_{\mathrm{F}}^2,\end{aligned}$$

where $\lambda_{\min}$ is the minimum eigenvalue of $(\boldsymbol{Z}(\tau)^\top\boldsymbol{Z}(\tau))^{1/2}$ and positive from Assumption (B3). The convergence of $\{\sum_{i=1}^n a_i \varphi(\mathrm{tr}(\boldsymbol{X}^\top(\tau)\boldsymbol{X}_i))\}_{\tau=0,1,\ldots}$ implies that $\boldsymbol{X}(\tau+1)-\boldsymbol{X}(\tau) \to 0$ as $\tau \to \infty$. On the other hand, by applying the update rule (12), the Riemannian gradient model for $\mathrm{St}(d_1, d_2)$ can be expressed as

$$\begin{aligned}\boldsymbol{g}(\boldsymbol{X}(\tau)) &= \sum_{i=1}^n a_i(\boldsymbol{X}_i - \boldsymbol{X}(\tau)\mathrm{sym}(\boldsymbol{X}(\tau)^\top\boldsymbol{X}_i))\phi(\mathrm{tr}(\boldsymbol{X}(\tau)^\top\boldsymbol{X}_i))\\ &= \boldsymbol{X}(\tau+1)(\boldsymbol{Z}(\tau)^\top\boldsymbol{Z}(\tau))^{\frac{1}{2}} - \boldsymbol{X}(\tau)\mathrm{sym}(\boldsymbol{X}(\tau)^\top\boldsymbol{X}(\tau+1)(\boldsymbol{Z}(\tau)^\top\boldsymbol{Z}(\tau))^{\frac{1}{2}}).\end{aligned}$$

Thus, when $\boldsymbol{X}(\tau+1)-\boldsymbol{X}(\tau) \to 0$, $\boldsymbol{g}(\boldsymbol{X}(\tau)) \to 0$ as $\tau \to \infty$ because $\boldsymbol{X}(\tau)^\top\boldsymbol{X}(\tau+1) \to \boldsymbol{I}_{d_2}$ and $\mathrm{sym}((\boldsymbol{Z}(\tau)^\top\boldsymbol{Z}(\tau))^{\frac{1}{2}}) = (\boldsymbol{Z}(\tau)^\top\boldsymbol{Z}(\tau))^{\frac{1}{2}}$. From the convergence of $\boldsymbol{X}(\tau+1)-\boldsymbol{X}(\tau)$ and $\boldsymbol{g}(\boldsymbol{X}(\tau))$, for all $0 < \epsilon < d_{\min}/3$, there exists some $T_\epsilon > 0$ such that

$$\|\boldsymbol{X}(\tau+1)-\boldsymbol{X}(\tau)\|_{\mathrm{F}} < \epsilon, \quad \tau \ge T_\epsilon, \tag{46}$$

$$\|\boldsymbol{g}(\boldsymbol{X}(\tau))\|_{\mathrm{F}} < \epsilon, \quad \tau \ge T_\epsilon. \tag{47}$$

Next, we show that $\boldsymbol{X}(\tau)$ exists in any neighborhoods of the zeros of $\boldsymbol{g}(\boldsymbol{X})$ when $\tau \ge T_\epsilon$. To this end, let us define a ball with the center $\tilde{\boldsymbol{X}}_k$ as

$$S_{\epsilon,k} := \left\{\boldsymbol{X} | \|\boldsymbol{X} - \tilde{\boldsymbol{X}}_k\|_{\mathrm{F}} < \epsilon,\ \boldsymbol{X} \in S_0\right\}.$$

Since ϕ is continuous, $\boldsymbol{g}(\boldsymbol{X}) \neq 0$ on $V_0 := S_0 \setminus \bigcup_{k=1}^n S_{\epsilon,k}$ where $\setminus$ denotes the set difference and we recall $S_0 := \{\boldsymbol{X} | D[\boldsymbol{X}(0)|\boldsymbol{X}] \ge 0\}$. Thus, there exists a constant $c_\epsilon > 0$ such that

$$\|\boldsymbol{g}(\boldsymbol{X})\|_{\mathrm{F}} > c_\epsilon \text{ for all } \boldsymbol{X} \in V_0. \tag{48}$$

From (47) and (48), we have

$$\{\boldsymbol{X}(\tau)\}_{\tau \ge T_\epsilon} \subset \bigcup_{k=1}^n S_{\epsilon,k}. \tag{49}$$

Finally, we confirm that $\boldsymbol{X}(\tau)$ converges to a zero of $\boldsymbol{g}(\boldsymbol{X})$. Let $\boldsymbol{X}'$ and $\boldsymbol{X}''$ be two points in distinct balls such that $\boldsymbol{X}' \in S_{\epsilon,k_1}$ and $\boldsymbol{X}'' \in S_{\epsilon,k_2}$ for some $1 \le k_1 \neq k_2 \le n$. Then,

$$\|\boldsymbol{X}' - \boldsymbol{X}''\|_{\mathrm{F}} \ge \|\tilde{\boldsymbol{X}}_{k_1} - \tilde{\boldsymbol{X}}_{k_2}\|_{\mathrm{F}} - \|\boldsymbol{X}' - \tilde{\boldsymbol{X}}_{k_1}\|_{\mathrm{F}} - \|\boldsymbol{X}'' - \tilde{\boldsymbol{X}}_{k_2}\|_{\mathrm{F}} > d_{\min} - \epsilon - \epsilon > \epsilon, \tag{50}$$

where we used $d_{\min} > 3\epsilon$. Thus, (46) and (50) mean that the sequence $\{\boldsymbol{X}(\tau)\}_{\tau \ge T_\epsilon}$ has to exist in a single ball $S_{\epsilon,k}$ for some k, which implies $\|\boldsymbol{X}(\tau) - \tilde{\boldsymbol{X}}_k\|_{\mathrm{F}} < \epsilon$. Therefore, $\{\boldsymbol{X}(\tau)\}_{\tau=0,1,\ldots}$ converges to a zero of $\boldsymbol{g}(\boldsymbol{X})$ with monotonic ascending.

Regarding $\mathrm{Gr}(d_1, d_2)$, the difference from $\mathrm{St}(d_1, d_2)$ is a way to prove $\boldsymbol{X}(\tau) - \boldsymbol{X}(\tau) \to 0$ as $\tau \to \infty$. As shown in the proof of Theorem 1, $\{\sum_{i=1}^n a_i \varphi_i^{\mathrm{Gr}}(\boldsymbol{X}(\tau))\}_{\tau=0,1,\dots}$ converges as τ increases. Then, from (37), we have

$$\begin{aligned}\sum_{i=1}^n a_i \left[\varphi_i^{\mathrm{Gr}}(\boldsymbol{X}(\tau+1)) - \varphi_i^{\mathrm{Gr}}(\boldsymbol{X}(\tau))\right] &\geq \mathrm{tr}((\boldsymbol{X}(\tau+1) - \boldsymbol{X}(\tau))^\top \boldsymbol{Y}(\tau)(\boldsymbol{X}(\tau+1) - \boldsymbol{X}(\tau))) \\ &\geq \lambda_{\min}^{\mathrm{y}} \|\boldsymbol{X}(\tau+1) - \boldsymbol{X}(\tau)\|_{\mathrm{F}}^2, \end{aligned} \tag{51}$$

where $\lambda_{\min}^{\mathrm{y}} 0$ denotes the minimum eigenvalue of $\boldsymbol{Y}(\tau)$ and is strictly positive by Assumption (B3). From (51), the convergence of $\{\sum_{i=1}^n a_i \varphi_i^{\mathrm{Gr}}(\boldsymbol{X}(\tau))\}_{\tau=0,1,\dots}$ implies $\boldsymbol{X}(\tau+1) - \boldsymbol{X}(\tau)) \to 0$ as $\tau \to \infty$. Then, by following the same steps as $\mathrm{St}(d_1, d_2)$, the proof is completed.

□

E Proof of Theorem 3

E.1 Useful lemma

Let us define the objective function used in the Fréchet mean estimation as

$$f(\boldsymbol{X}) = \frac{1}{2}\sum_{i=1}^n w_i \mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{X}_i),$$

where w_i denotes a weight such that $0 \leq w_i \leq 1$ and $\sum_{i=1}^n w_i = 1$. The objective function $f(\boldsymbol{X})$ is minimized by applying a Riemannian gradient descent method whose iterative rule is given by

$$\boldsymbol{X}(\tau+1) = \exp_{\boldsymbol{X}(\tau)}(-t\mathrm{grad}(f(\boldsymbol{X}(\tau)))),$$

where $t > 0$ is a step size parameter. The following lemma provides sufficient conditions that $f(\boldsymbol{X})$ is monotonically decreased as τ increases:

Lemma 1 (Theorem 4.1 in Afsari et al. (2013))**.** *Assume that* $\{\boldsymbol{X}_i\}_{i=1}^n \subset B(\boldsymbol{X}_o, \frac{1}{3}r_*)$ *and* $\boldsymbol{X}(0) \in B(\boldsymbol{X}_o, \frac{1}{3}r_*)$. *Then, for all* $t \in \left(0, \frac{2}{c_{\kappa_0}(4r_*/3)}\right)$,

$$f(\boldsymbol{X}(\tau)) - f(\boldsymbol{X}(\tau+1)) \geq 0.$$

E.2 Main proof

Proof. Since $\mathrm{grad}(\mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{X}_i)) = -2\log_{\boldsymbol{X}}(\boldsymbol{X}_i)$ (Karcher, 1977), we express $\boldsymbol{g}(\boldsymbol{X})$ as

$$\boldsymbol{g}(\boldsymbol{X}) = \sum_{i=1}^n a_i \log_{\boldsymbol{X}}(\boldsymbol{X}_i)\phi\left(\mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{X}_i)\right) = \frac{1}{2}\sum_{i=1}^n a_i \mathrm{grad}\left(\varphi\left(\mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{X}_i)\right)\right),$$

where we used $\phi(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\varphi(t)$. Substituting the model $\boldsymbol{g}(\boldsymbol{X})$ into the path integral (20) yields

$$\begin{aligned} D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] &= \sum_{i=1}^n \frac{a_i}{2}\int_0^1 \langle \dot{\boldsymbol{C}}(t), \mathrm{grad}\left(\varphi\left(\mathrm{dist}^2(\boldsymbol{C}(t), \boldsymbol{X}_i)\right)\right)\rangle_{\boldsymbol{C}(t)}\mathrm{d}t \\ &= \sum_{i=1}^n \frac{a_i}{2}\left[\varphi\left(\mathrm{dist}^2(\boldsymbol{X}(\tau+1), \boldsymbol{X}_i)\right) - \varphi\left(\mathrm{dist}^2(\boldsymbol{X}(\tau), \boldsymbol{X}_i)\right)\right] \\ &\geq \sum_{i=1}^n \frac{a_i}{2}\phi\left(\mathrm{dist}^2(\boldsymbol{X}(\tau), \boldsymbol{X}_i)\right)\left[\mathrm{dist}^2(\boldsymbol{X}(\tau), \boldsymbol{X}_i) - \mathrm{dist}^2(\boldsymbol{X}(\tau+1), \boldsymbol{X}_i)\right], \end{aligned}$$

where we applied the inequality (31) to $\varphi(\cdot)$ on the second line. By defining $w_i(\tau) := a_i\phi\left(\mathrm{dist}^2(\boldsymbol{X}(\tau), \boldsymbol{X}_i)\right)$, we express the right-hand side as

$$D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] \geq \left(\frac{1}{2}\sum_{i=1}^n w_i(\tau)\right)\{f_\tau(\boldsymbol{X}(\tau)) - f_\tau(\boldsymbol{X}(\tau+1))\}, \tag{52}$$

where with $\tilde{w}_i(\tau) := \frac{w_i(\tau)}{\sum_{k=1}^n w_k(\tau)}$,

$$f_\tau(\boldsymbol{X}) := \sum_{i=1}^n \tilde{w}_i(\tau)\mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{X}_i).$$

We note that $0 \leq \tilde{w}_i(\tau) \leq 1$ and $\sum_{i=1}^n \tilde{w}_i(\tau) = 1$ by Assumptions (C3-4).

Next, we show that the right-hand side on (52) is nonnegative based on Lemma 1. The update rule (7) is equivalently expressed as the exponential map of the Riemannian gradient of $f_\tau(\boldsymbol{X})$ at $\boldsymbol{X} = \boldsymbol{X}(\tau)$ as follows:

$$\begin{aligned}\boldsymbol{X}(\tau+1) &= \exp_{\boldsymbol{X}(\tau)}\left(\frac{\sum_{i=1}^n a_i \log_{\boldsymbol{X}}(\boldsymbol{X}_i)\phi\left(\mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{X}_i)\right)}{\sum_{i=1}^n a_i \phi\left(\mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{X}_i)\right)}\right)\\ &= \exp_{\boldsymbol{X}(\tau)}\left(\sum_{i=1}^n \tilde{w}_i(\tau)\log_{\boldsymbol{X}}(\boldsymbol{X}_i)\right) = \exp_{\boldsymbol{X}(\tau)}\left(-\frac{1}{2}\mathrm{grad}(f_\tau(\boldsymbol{X}(\tau))\right). \qquad (53)\end{aligned}$$

By Assumptions (C1-2), Lemma 1 ensures that for each τ,

$$f_\tau(\boldsymbol{X}(\tau)) - f_\tau(\boldsymbol{X}(\tau+1)) \geq 0, \qquad (54)$$

under the update rule (53). Thus, combining (54) with (52), it is established $D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)] \geq 0$. As done in $\mathrm{S}^+(d)$, the convergence of $\{D_{\boldsymbol{g}}[\boldsymbol{X}(\tau)|\boldsymbol{X}(\tau+1)]\}_{\tau=0,1,\dots}$ can be also confirmed. Thus, the proof is completed. □

F Proof of Theorem 4

Proof. We follow the proof of Theorem 2.5 in Yao and Li (2014). We start by the following simple fact:

$$\mathrm{dist}(\boldsymbol{M}, \boldsymbol{M}') \leq \mathrm{dist}(\boldsymbol{M}, \boldsymbol{X}_i) + \mathrm{dist}(\boldsymbol{M}', \boldsymbol{X}_i).$$

This indicates that $\mathrm{dist}(\boldsymbol{M}, \boldsymbol{M}')$ is bounded if $\mathrm{dist}(\boldsymbol{M}', \boldsymbol{X}_i)$ is bounded for all $i = 1, \dots, n$. Let us express $\tilde{a}_i := a_i/a_{\max}$ and $\varphi^*(t) = \varphi(t)/\varphi(0)$. Thus, we have $\tilde{a}_i \leq 1$ and $\varphi^*(t) \leq 1$, and note that the following holds:

$$\boldsymbol{M}' := \underset{\boldsymbol{X}\in\mathbb{M}}{\mathrm{argmax}}\left[\sum_{i=1}^{n+m} \tilde{a}_i \varphi^*\left(\mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{Z}_i)\right)\right].$$

Here, without loss of generality, we assume that

$$\boldsymbol{Z}_i = \begin{cases} \boldsymbol{X}_i & i = 1, \dots, n \\ \boldsymbol{X}'_i & i = n+1, \dots, n+m. \end{cases}$$

Since the coefficients a_i are assumed to be nonnegative, we first obtain

$$\begin{aligned}\sum_{i=1}^{m+n} \tilde{a}_i \varphi^*\left(\mathrm{dist}^2(\boldsymbol{M}, \boldsymbol{Z}_i)\right) &= \sum_{i=1}^{n} \tilde{a}_i \varphi^*\left(\mathrm{dist}^2(\boldsymbol{M}, \boldsymbol{Z}_i)\right) + \sum_{i=n+1}^{m+n} \tilde{a}_i \varphi^*\left(\mathrm{dist}^2(\boldsymbol{M}, \boldsymbol{Z}_i)\right)\\ &\geq \sum_{i=1}^{n} \tilde{a}_i \varphi^*\left(\mathrm{dist}^2(\boldsymbol{M}, \boldsymbol{X}_i)\right) = R. \qquad (55)\end{aligned}$$

Next, we prove that $\mathrm{dist}(\boldsymbol{M}', \boldsymbol{X}_i)$ is bounded if $m < R$, which implies that there exist some $\xi > 0$ and $C > 0$ such that $A_n\xi + m < R$ with $A_n := \sum_{i=1}^n \tilde{a}_i$ and $\varphi^*(t) \leq \xi$ for $|t| \geq C$, respectively. For $\boldsymbol{Y} \in \mathbb{M}$ satisfying $\mathrm{dist}(\boldsymbol{Y}, \boldsymbol{X}_i) \geq C$ for $i = 1, \dots, n$,

$$\begin{aligned}\sum_{i=1}^{m+n} \tilde{a}_i \varphi^*\left(\mathrm{dist}^2(\boldsymbol{Y}, \boldsymbol{Z}_i)\right) &\leq \xi\sum_{i=1}^{n} \tilde{a}_i + \sum_{i=n+1}^{m+n} \tilde{a}_i \varphi^*\left(\mathrm{dist}^2(\boldsymbol{Y}, \boldsymbol{Z}_i)\right)\\ &\leq \xi\sum_{i=1}^{n} \tilde{a}_i + \sum_{i=n+1}^{m+n} \tilde{a}_i\\ &\leq A_n\xi + m, \qquad (56)\end{aligned}$$

where $\tilde{a}_i \leq 1$ and $\varphi^* \left(\mathrm{dist}^2(\boldsymbol{Y}, \boldsymbol{X}_i)\right) \leq 1$ for all i. Combining (55) with (56) under $m + A_n\xi < R$ yields

$$\sum_{i=1}^{m+n} \tilde{a}_i \varphi^* \left(\mathrm{dist}^2(\boldsymbol{Y}, \boldsymbol{X}_i)\right) < \sum_{i=1}^{m+n} \tilde{a}_i \varphi^* \left(\mathrm{dist}^2(\boldsymbol{M}, \boldsymbol{X}_i)\right). \tag{57}$$

Substituting $\boldsymbol{Y} = \boldsymbol{M'}$ in (57) contradicts to the definition of $\boldsymbol{M'}$. Thus, if $m < R$, it must be $\mathrm{dist}(\boldsymbol{M'}, \boldsymbol{X}_i) < C$ for some $\boldsymbol{X}_i$, which means that $\mathrm{dist}(\boldsymbol{M'}, \boldsymbol{M})$ is bounded.

Next, we assume that $m > R/\bar{A}'_m$ which implies there exist some $\xi > 0$ and C such that $m\bar{A}'_m > R + m\bar{A}'_m\xi$ and $\varphi^*(t) \leq \xi$ for $|t| \geq C$, respectively. For $\boldsymbol{Y}$ satisfying $\mathrm{dist}^2(\boldsymbol{Y}, \boldsymbol{Z}_i) \geq C$ for all $i = n+1, \ldots, n+m$,

$$\begin{aligned}\sum_{i=1}^{m+n} \tilde{a}_i \varphi^* \left(\mathrm{dist}^2(\boldsymbol{Y}, \boldsymbol{Z}_i)\right) &= \sum_{i=1}^{n} \tilde{a}_i \varphi^* \left(\mathrm{dist}^2(\boldsymbol{Y}, \boldsymbol{X}_i)\right) + \sum_{i=n+1}^{n+m} \tilde{a}_i \varphi^* \left(\mathrm{dist}^2(\boldsymbol{Y}, \boldsymbol{Z}_i)\right) \\ &\leq \sum_{i=1}^{n} \tilde{a}_i \varphi^* \left(\mathrm{dist}^2(\boldsymbol{M}, \boldsymbol{X}_i)\right) + \sum_{i=n+1}^{n+m} \tilde{a}_i \xi \\ &= R + mA'_m\xi. \end{aligned} \tag{58}$$

By assuming that $\boldsymbol{Z}_{n+i}$ (i.e., $\boldsymbol{X}'_i$) are same for all $i = 1, \ldots, m$ and there exists $\boldsymbol{M}^*$ such that $\mathrm{dist}(\boldsymbol{M}^*, \boldsymbol{X}'_i) = 0$, we obtain

$$\begin{aligned}\sum_{i=1}^{m+n} \tilde{a}_i \varphi^* \left(\mathrm{dist}^2(\boldsymbol{M}^*, \boldsymbol{Z}_i)\right) &= \sum_{i=1}^{n} \tilde{a}_i \varphi^* \left(\mathrm{dist}^2(\boldsymbol{M}^*, \boldsymbol{X}_i)\right) + \sum_{i=n+1}^{n+m} \tilde{a}_i \varphi^* \left(0\right) \\ &\geq mA'_m \\ &> R + mA'_m\xi \end{aligned} \tag{59}$$

Eqs (58) and (59) yield

$$\sum_{i=1}^{m+n} \tilde{a}_i \varphi^* \left(\mathrm{dist}^2(\boldsymbol{M}^*, \boldsymbol{Z}_i)\right) > \sum_{i=1}^{m+n} \tilde{a}_i \varphi^* \left(\mathrm{dist}^2(\boldsymbol{Y}, \boldsymbol{Z}_i)\right). \tag{60}$$

When $\boldsymbol{Y} = \boldsymbol{M'}$, (60) contradicts to the definition of $\boldsymbol{M'}$. Thus, $\mathrm{dist}^2(\boldsymbol{M'}, \boldsymbol{Z}_i) < C$ for some $i \in \{n+1, \ldots, m+n\}$. The triangle inequality,

$$\mathrm{dist}(\boldsymbol{O}, \boldsymbol{Z}_i) < \mathrm{dist}(\boldsymbol{O}, \boldsymbol{M'}) + \mathrm{dist}(\boldsymbol{M'}, \boldsymbol{Z}_i),$$

implies that $\mathrm{dist}(\boldsymbol{O}, \boldsymbol{M'})$ diverges if $\mathrm{dist}(\boldsymbol{O}, \boldsymbol{Z}_i) \to \infty$. This means that $\mathrm{dist}(\boldsymbol{M}, \boldsymbol{M'}) \to \infty$ if $m > \frac{R}{A'_m}$.

Finally, we confirm that there exits some m^* satisfying (24) from the results that $\mathrm{dist}(\boldsymbol{M}, \boldsymbol{M'})$ is bounded if $m < R$ and $\mathrm{dist}(\boldsymbol{M}, \boldsymbol{M'})$ diverges if $m > \frac{R}{A'_m}$, □

G Details of the Riemannian gradient estimator

This section gives the details of our Riemannian gradient estimator on the Grassmann manifold $\mathrm{Gr}(d_1, d_2)$, SPD matrices $\mathrm{S}^+(d)$ and unit sphere Ω_{d_1-1}. We follow the direct approach in Ashizawa et al. (2017) and employ the empirical version of the Fisher divergence as

$$\widehat{J}(\boldsymbol{g}) := \frac{1}{n} \sum_{k=1}^{n} \left[\langle \boldsymbol{g}(\boldsymbol{X}_k), \boldsymbol{g}(\boldsymbol{X}_k)\rangle_{\boldsymbol{X}_k} + 2\mathrm{div}(\boldsymbol{g}(\boldsymbol{X}_k))\right].$$

G.1 Gradient estimator on $\mathrm{Gr}(d_1, d_2)$ and $\mathrm{S}^+(d)$

Here, $\boldsymbol{X}$ is a d_1 by d_2 matrix for $\mathrm{Gr}(d_1, d_2)$, while it is a square matrix for $\mathrm{S}^+(d)$ in $d = d_1 = d_2$. When the Euclidean metric is employed, it can be written as

$$\langle \boldsymbol{g}(\boldsymbol{X}), \boldsymbol{g}(\boldsymbol{X})\rangle_{\boldsymbol{X}} = \mathrm{tr}(\boldsymbol{g}(\boldsymbol{X})^\top \boldsymbol{g}(\boldsymbol{X})), \qquad \mathrm{div}(\boldsymbol{g}(\boldsymbol{X})) = \sum_{l=1}^{d_1} \sum_{m=1}^{d_2} \partial_{lm}[\boldsymbol{g}(\boldsymbol{X})]_{lm},$$

where ∂_{lm} denotes the partial derivative with respect to the (l,m)-th element in $\boldsymbol{X}$. Substituting these into the empirical divergence $\widehat{J}(\boldsymbol{g})$ yields

$$\widehat{J}(\boldsymbol{g}) = \frac{1}{n}\sum_{k=1}^{n}\left[\mathrm{tr}(\boldsymbol{g}(\boldsymbol{X}_k)^\top \boldsymbol{g}(\boldsymbol{X}_k)) + 2\sum_{l=1}^{d_1}\sum_{m=1}^{d_2}\partial_{lm}[\boldsymbol{g}(\boldsymbol{X}_k)]_{lm}\right].$$

When (35) and (41)[4] are adapted for $\boldsymbol{g}(\boldsymbol{X})$ in $\mathrm{Gr}(d_1,d_2)$ and $\mathrm{S}^+(d)$ respectively, $\widehat{J}(\boldsymbol{g})$ simply takes a quadratic form of the coefficient vector $\boldsymbol{a} = (a_1, a_2, \dots, a_n)^\top$ as

$$\widehat{J}(\boldsymbol{a}) = \boldsymbol{a}^\top \boldsymbol{H}\boldsymbol{a} + 2\boldsymbol{a}^\top \boldsymbol{h}. \tag{61}$$

Regarding $\mathrm{Gr}(d_1,d_2)$, $\boldsymbol{H}$ and $\boldsymbol{h}$ are given by

$$[\boldsymbol{H}]_{ij} = \frac{1}{n}\sum_{k=1}^{n}\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{X}_k)^\top \boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{X}_j\boldsymbol{X}_j^\top\boldsymbol{X}_k))\phi_i^{\mathrm{Gr}}(\boldsymbol{X}_k)\,\phi_j^{\mathrm{Gr}}(\boldsymbol{X}_k)$$

$$[\boldsymbol{h}]_i = \frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{d_1}\sum_{m=1}^{d_2}\partial_{lm}\left\{[\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{X}_k))]_{lm}\phi_i^{\mathrm{Gr}}(\boldsymbol{X}_k)\right\}.$$

where $\phi_i^{\mathrm{Gr}}(\boldsymbol{X}) = \phi\left(\frac{\mathrm{tr}(\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top)}{\sigma^2}\right)$, and we recall that $\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Z}) = \boldsymbol{Z} - \boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Z}$. On the other hand, for $\mathrm{S}^+(d)$,

$$[\boldsymbol{H}]_{ij} = \frac{1}{n}\sum_{k=1}^{n}\mathrm{tr}((\boldsymbol{X}_i - \boldsymbol{X}_k)^\top(\boldsymbol{X}_j - \boldsymbol{X}_k))\phi\left(\frac{\|\boldsymbol{X}_k - \boldsymbol{X}_i\|_{\mathrm{F}}^2}{\sigma^2}\right)\phi\left(\frac{\|\boldsymbol{X}_k - \boldsymbol{X}_j\|_{\mathrm{F}}^2}{\sigma^2}\right)$$

$$[\boldsymbol{h}]_i = \frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{d}\sum_{m=1}^{d}\left\{\frac{[\boldsymbol{X}_i - \boldsymbol{X}_k]_{lm}^2}{\sigma^2} - 1\right\}\phi'\left(\frac{\|\boldsymbol{X}_k - \boldsymbol{X}_j\|_{\mathrm{F}}^2}{\sigma^2}\right),$$

where $\phi'(t) := \frac{\mathrm{d}}{\mathrm{d}t}\phi(t)$. After applying the ℓ_2-regularization and adding the nonnegative constraint on $\boldsymbol{a}$, the optimal coefficient vector $\widehat{\boldsymbol{a}}$ is determined by solving a quadratic optimization problem as follows:

$$\min_{\boldsymbol{a}}\left[\boldsymbol{a}^\top(\boldsymbol{H} + \lambda\boldsymbol{I}_n)\boldsymbol{a} + 2\boldsymbol{a}^\top\boldsymbol{h}\right] \quad \text{s.t.} \quad \boldsymbol{a} \geq \boldsymbol{0}, \tag{62}$$

where $\boldsymbol{I}_n$ denotes the n by n identity matrix and $\lambda > 0$ is the regularization parameter. Finally, by substituting the optimal vector $\widehat{\boldsymbol{a}}$ into the model $\boldsymbol{g}(\boldsymbol{X})$, we obtain our estimator for the Riemannian gradient.

G.2 Gradient estimator on Ω_{d_1-1}

Here, we focus on the unit sphere Ω_{d_1-1} which is a special case of $d_2 = 1$ in $\mathrm{St}(d_1,d_2)$. Thus, we express data points and samples as vectors $\boldsymbol{x} \in \mathbb{R}^{d_1}$ and $\boldsymbol{x}_i \in \mathbb{R}^{d_1}$, respectively. With the Euclidean metric $\langle \boldsymbol{y}, \boldsymbol{z}\rangle_{\boldsymbol{x}} = \boldsymbol{y}^\top\boldsymbol{z}$ and our model for the Riemannian gradient,

$$\boldsymbol{g}(\boldsymbol{x}) = \sum_{i=1}^{n} a_i(\boldsymbol{x}_i - \boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{x}_i)\phi\left(\frac{\boldsymbol{x}^\top\boldsymbol{x}_i}{\sigma^2}\right), \tag{63}$$

we reach $\widehat{J}(\boldsymbol{a})$ with the exactly same quadratic form as (61). Note that (63) is a special case of the model (28) for $\mathrm{St}(d_1,d_2)$ in $d_2 = 1$. The (i,j)-th and i-th elements in $\boldsymbol{H}$ and $\boldsymbol{h}$ are given by

$$[\boldsymbol{H}]_{ij} := \frac{1}{n}\sum_{k=1}^{n}\left\{\boldsymbol{x}_i^\top\boldsymbol{x}_j - (\boldsymbol{x}_k^\top\boldsymbol{x}_i)(\boldsymbol{x}_k^\top\boldsymbol{x}_j)\right\}\phi\left(\frac{\boldsymbol{x}_k^\top\boldsymbol{x}_i}{\sigma^2}\right)\phi\left(\frac{\boldsymbol{x}_k^\top\boldsymbol{x}_j}{\sigma^2}\right)$$

$$[\boldsymbol{h}]_i := \frac{1}{n}\sum_{k=1}^{n}\left[\frac{1 - (\boldsymbol{x}_k^\top\boldsymbol{x}_i)^2}{\sigma^2}\phi'\left(\frac{\boldsymbol{x}_k^\top\boldsymbol{x}_i}{\sigma^2}\right) - (d_1 + 1)(\boldsymbol{x}_k^\top\boldsymbol{x}_i)\phi\left(\frac{\boldsymbol{x}_k^\top\boldsymbol{x}_i}{\sigma^2}\right)\right],$$

respectively. By apply the ℓ_2-regularization and adding the nonnegative constraint on $\boldsymbol{a}$, the optimal coefficient vector $\widehat{\boldsymbol{a}}$ is determined by solving the same form of the problem as (62). Finally, the substitution of the optimal vector $\widehat{\boldsymbol{a}}$ into the model $\boldsymbol{g}(\boldsymbol{X})$ yields our estimator for the Riemannian gradient on Ω_{d_1-1}.

[4] Note that it is supposed to be $\sigma = 1$ in (35) and (41) just for notational simplicity.

H Experimental details

This section describes the experimental details in Section 6. For the proposed methods, in all of experiments, we fixed the regularization parameter as $\lambda = n^{-0.9}$ as suggested in Kanamori et al. (2012), while the width parameter σ was determined by the five-hold cross validation with respect to the empirical Fisher divergence $\widehat{J}(\boldsymbol{g})$. In addition, to decrease the computational cost, as in the Nyström approximation (Rudi et al., 2015), we used a subset of data samples as the center points in $\phi(\cdot)$ by randomly choosing B samples from n data samples where $B = \min(100, n)$.

H.1 Clustering for directional data Ω_{d_1-1}

The following methods were applied to directional data for clustering:

- **Proposed method**: We employed the direct Riemannian gradient estimator described in Section G.2 where $\phi\left(\frac{\mathrm{tr}(\boldsymbol{x}^\top \boldsymbol{x}_i)}{\sigma^2}\right) = \exp\left(\frac{-1+\mathrm{tr}(\boldsymbol{x}^\top \boldsymbol{x}_i)}{\sigma^2}\right)$. After the gradient model was estimated, we substituted the estimated coefficients $\widehat{a}_i$ and applied the update rule (13) to all data samples until they converge.

- **DMRrot**: Directional mean shift (DMR) employs the following kernel density estimation:

$$\widehat{p}_{\mathrm{KDE}}(\boldsymbol{x}) = \frac{1}{nZ_h} \sum_{i=1}^{n} L\left(\frac{1 - \boldsymbol{x}^\top \boldsymbol{x}_i}{h^2}\right), \tag{64}$$

  where Z_h denotes the normalizing constant and L is the von Mises kernel given by $L\left(\frac{1-\boldsymbol{x}^\top \boldsymbol{x}_i}{h^2}\right) \propto \exp\left(\frac{\boldsymbol{x}^\top \boldsymbol{x}_i}{h^2}\right)$. Based on $\widehat{p}_{\mathrm{KDE}}(\boldsymbol{x})$, the following update rule was proposed in Zhang and Chen (2020):

$$\boldsymbol{x}(\tau + 1) = -\frac{\sum_{i=1}^{n} \boldsymbol{x}_i L'\left(\frac{1-\boldsymbol{x}(\tau)^\top \boldsymbol{x}_i}{h^2}\right)}{\left\|\sum_{i=1}^{n} \boldsymbol{x}_i L\left(\frac{1-\boldsymbol{x}(\tau)^\top \boldsymbol{x}_i}{h^2}\right)\right\|}, \tag{65}$$

  where $L'(t) := \frac{\mathrm{d}}{\mathrm{d}t} L(t)$. Following Zhang and Chen (2020), the bandwidth parameter was selected based on the *rule of thumb* for directional data (García-Portugués, 2013).

- **DMRcv**: DMRcv used the same update rule as (65), but the bandwidth parameter h was cross-validated based on the log-likelihood of $\widehat{p}_{\mathrm{KDE}}(\boldsymbol{x})$.

Data was sampled based on Von Mises densities as follows[5]: The first angle of data samples was drawn from a mixture of three Von Mises densities (Fig.1(a)), while the other angles are independently from an identical Von Mises density. The performance was measured by adjusted Rand index (ARI) (Hubert and Arabie, 1985): ARI is less than or equal to one, and a larger value of ARI means a better clustering result.

H.2 Clustering on $\mathrm{Gr}(d_1, d_2)$

For clustering on the Grassmann manifold $\mathrm{Gr}(d_1, d_2)$, we followed the experimental setting in Ashizawa et al. (2017)[6]. Each data sample $\boldsymbol{X}_i$ was generated as

$$\boldsymbol{X}_i = \left(\begin{array}{cc|c} \cos\tau_i & -\sin\tau_i & \boldsymbol{O}_{2,d_1-2} \\ \sin\tau_i & \cos\tau_i & \\ \hline \multicolumn{2}{c|}{\boldsymbol{O}_{d_1-2,2}} & \boldsymbol{I}_{d_1-2} \end{array}\right) \boldsymbol{S} \begin{pmatrix} \cos\eta_i & -\sin\eta_i \\ \sin\eta_i & \cos\eta_i \end{pmatrix}, \tag{66}$$

[5] The MATLAB code for generating random numbers from a Von Mises density was available at `https://jp.mathworks.com/matlabcentral/fileexchange/37241-vmrand-fmu-fkappa-varargin`.

[6] The MATLAB code is available at `https://t-sakai-kure.github.io/software-ja.html`.

where $\boldsymbol{O}_{d,d'}$ denotes the $d \times d'$ null matrix, $\boldsymbol{S}$ is the d_1 by d_2 orthonormal matrix obtained by applying the singular value decomposition to a random matrix, and η_i and τ_i are samples as follows:

$$\eta_i \sim N\left(0, (\pi/2)^2\right), \quad \tau_i \sim \begin{cases} N\left(0, \frac{\pi^2}{15^2}\right) & \text{for } i = 1, \dots, \frac{n}{3}, \\ N\left(\frac{2\pi}{3}, \frac{\pi^2}{15^2}\right) & \text{for } i = \frac{n}{3}+1, \dots, \frac{2n}{3}, \\ N\left(\frac{4\pi}{3}, \frac{\pi^2}{15^2}\right) & \text{for } i = \frac{2n}{3}+1, \dots, n, \end{cases} \tag{67}$$

where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2. Thus, in the data samples, there exist three clusters according to τ in (67).

We applied the following two methods in clustering to the data samples on $\mathrm{Gr}(d_1, d_2)$:

- **Proposed method**: In our gradient model (35), we employed

$$\phi_i^{\mathrm{Gr}}(\boldsymbol{X}) = \exp\left(-\frac{d_2 - \mathrm{tr}(\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top)}{2\sigma^2}\right). \tag{68}$$

  Then, the Riemannian gradient model was estimated as in Section G.1. The estimated coefficients $\widehat{a}_i$ were substituted in our update rule (16), and the data samples were updated toward the modes on $\mathrm{Gr}(d_1, d_2)$.

- **Geodesic** (Ashizawa et al., 2017): In the Riemannian gradient model (6), the logarithm map was given by $\log_{\boldsymbol{X}}(\boldsymbol{Z}) = (I_{d_1} - \boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{X}$ and the same function as (68) was employed for $\phi(\cdot)$ because the geodesic distance is given by

$$\mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{Y}) = d_2 - \mathrm{tr}(\boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{Y}).$$

  The estimation procedure for the gradient model is essentially the same as described in Section G.1. As in the proposed method, the regularization parameter was fixed at $\lambda = n^{-0.9}$, while the width parameter σ was determined by cross validation with respect to the empirical Fisher divergence $\widehat{J}(\boldsymbol{g})$. Regarding the update rule (7), the exponential map was computed as follows:

$$\exp_{\boldsymbol{X}} \boldsymbol{Z} = (\boldsymbol{X}\boldsymbol{V}\cos\boldsymbol{\Sigma} + \boldsymbol{U}\sin\boldsymbol{\Sigma})\boldsymbol{V}^\top,$$

  where $\boldsymbol{U}, \boldsymbol{\Sigma}$, and $\boldsymbol{V}$ come from the singular value decomposition of $\boldsymbol{Z} \in \mathbb{R}^{d_1 \times d_2}$, i.e., $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, and "cos" and "sin" are applied element-wisely to the diagonal of $\boldsymbol{\Sigma}$.

H.3 Outlier robustness on $\mathrm{S}^+(d)$

We generated data samples of symmetric positive definite matrices from

$$\boldsymbol{X}_i = \boldsymbol{B}^\top \boldsymbol{B} + \mathrm{diag}(\boldsymbol{\beta}_i),$$

where $\boldsymbol{B}$ is a d by d random matrix drawn from the normal density, $\mathrm{diag}(\boldsymbol{\beta}_i)$ is the diagonal matrix with the elements of $\boldsymbol{\beta}_i$ on the diagonal, and $\boldsymbol{\beta}_i$ is a d-dimensional vector and sampled from a *contaminated* exponential density as

$$(1-\epsilon)\mu^{-1}e^{\beta/\mu} + \epsilon\mu_o^{-1}e^{(\beta+5)/\mu_o}.$$

The samples from $\mu_o^{-1}e^{(\beta+5)/\mu_o}$ can be regarded as outliers and ϵ denotes the *outlier ratio*. Here, we set $\mu = 0.5$ and $\mu_o = 0.1$. The total number of samples was $n = 500$.

In order to investigate the outlier robustness, we applied the the following methods to the generated data samples:

- **Proposed method**: In the gradient model (41), we employed

$$\phi_i\left(\frac{\|\boldsymbol{X} - \boldsymbol{X}_i\|_{\mathrm{F}}^2}{\sigma^2}\right) = \exp\left(-\frac{\|\boldsymbol{X} - \boldsymbol{X}_i\|_{\mathrm{F}}^2}{2\sigma^2}\right).$$

  After estimating the coefficients a_i in the Riemannian gradient model as described in Section G.1, the estimated ones used in our update rule (18), and the data samples were updated toward the modes.

- **Mean shift (MS)**: Based on the Euclidean gradient of the kernel density estimation,

  $$\widehat{p}_{\mathrm{KDE}}(\boldsymbol{X}) = \frac{1}{nZ_h} \exp\left(-\frac{\|\boldsymbol{X} - \boldsymbol{X}_i\|_{\mathrm{F}}^2}{2h^2}\right),$$

  where h and Z_h denote the width parameter and normalizing constant respectively, the following update rule was derived as in Section 3.3:

  $$\boldsymbol{X}(\tau+1) = \frac{\sum_{i=1}^n \boldsymbol{X}_i \exp\left(-\frac{\|\boldsymbol{X}(\tau)-\boldsymbol{X}_i\|_{\mathrm{F}}^2}{2h^2}\right)}{\sum_{i=1}^n \exp\left(-\frac{\|\boldsymbol{X}(\tau)-\boldsymbol{X}_i\|_{\mathrm{F}}^2}{2h^2}\right)},$$

  where the bandwidth parameter h was determined by the five-hold cross validation based on the log-likelihood of $\widehat{p}_{\mathrm{KDE}}$.

- **Karcher mean (KMean)**[7] (Karcher, 1977): Karcher mean was estimated by minimizing $\frac{1}{n}\sum_{i=1}^n \mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{X}_i)$.

- **Geometric median (GMed)**[7] (Fletcher et al., 2009): Geometric median was estimated by minimizing $\frac{1}{n}\sum_{i=1}^n \mathrm{dist}(\boldsymbol{X}, \boldsymbol{X}_i)$.

Regarding the proposed method and MS, the estimate from the geometric median was used for $\boldsymbol{X}(0)$ (i.e., the initial point). The performance was measured by

$$\|\boldsymbol{B}^\top \boldsymbol{B} - \widehat{\boldsymbol{M}}\|_{\mathrm{F}},$$

where $\widehat{\boldsymbol{M}}$ denotes an estimated mode.

H.4 Application to EEG data

EEG data was measured from four human subjects performing a cued motor imagery task. A total of 200 task trials per subject were provided, each labeled with two classes of motor imagery, either left/right-hand movements (two subjects) or left-hand/foot movements (other two subjects). The data was given in 59 channels, downsampled at 100 Hz, and further band-pass filtered (8-30Hz) and then re-referenced to the common average. Finally, we computed sample covariance matrices for every trial, using 1-3s after the cued onset.

Since the data samples lie on $\mathrm{S}^+(d)$, in addition to the proposed method, we applied KMean, GMed and MS which were used as described in Section H.3.

[7] We used the *Maopt* package (Boumal et al., 2014).