Diffusion means in geometric spaces

Eltzner, Benjamin; Hansen, Pernille E.H.; Huckemann, Stephan F.; Sommer, Stefan

Published in: Bernoulli

DOI: 10.3150/22-BEJ1578

Publication date: 2023

Document version Publisher's PDF, also known as Version of record

Document license: Unspecified

We introduce a location statistic for distributions on non-linear geometric spaces, the diffusion mean, serving as an extension and an alternative to the Fréchet mean. The diffusion mean arises as the generalization of Gaussian maximum likelihood analysis to non-linear spaces by maximizing the likelihood of a Brownian motion. The diffusion mean depends on a time parameter $t$, which admits the interpretation of the allowed variance of the diffusion. The diffusion $t$-mean of a distribution $X$ at time $t$, given the end-point distribution $X$, is the most likely origin of a Brownian motion at time $t$, given the end-point distribution $X$. We give a detailed description of the asymptotic behavior of the diffusion estimator and provide sufficient conditions for the diffusion estimator to be strongly consistent. Particularly, we present a smeary central limit theorem for diffusion means and we show that joint estimation of the mean and diffusion variance rules out smeariness in all directions simultaneously in general situations. Furthermore, we investigate properties of the diffusion mean for distributions on the sphere $S^m$. Experimentally, we consider simulated data and data from magnetic pole reversals, all indicating similar or improved convergence rate compared to the Fréchet mean. Here, we additionally estimate $t$ and consider its effects on smeariness and uniqueness of the diffusion mean for distributions on the sphere.

Keywords: Diffusion mean; generalized Fréchet mean; geometric statistics; maximum likelihood estimation; spherical statistics

1. Introduction

In many systems of interest, measured quantities cannot be easily represented in a Euclidean space. Important examples include directional data on a circle or sphere, cf. Mardia and Jupp (2000) and landmark shapes, cf. Dryden and Mardia (2016), Small (1996). For a statistical treatment of such data, it is necessary to generalize statistical concepts from Euclidean spaces to more general spaces.

The mean, which is an important location statistic, was generalized to metric spaces by Fréchet (1948) whose generalization has had a lasting impact on the statistics literature. For this reason, in statistics, this mean is called the Fréchet mean while in the field of geometry it is referred to as the barycenter. It is defined for random variables $X$ on metric spaces $(Q,d)$ as minimizers of the Fréchet function, i.e. the set of points $E(X) = \arg\min_{y \in Q} \mathbb{E}[d(X,y)^2]$. A common example is the intrinsic mean, also called the Riemann center of mass for a smooth manifold $Q$ with geodesic distance $d$, leading to many important results regarding uniqueness Afsari (2011), Groissier (2005), Karcher (1977), Kendall (1990), strong consistency of estimators Bhattacharya and Patrangenaru (2003), Evans and Jaffe (2020), Huckemann (2011a), Huckemann and Eltzner (2018), Schötz (2022), Ziezold (1977) and central limit theorems (CLTs) Bhattacharya and Patrangenaru (2005), Eltzner and Huckemann (2019), Huckemann (2011b), Renlund (2011).

While we will briefly touch on other types of Fréchet means below, to keep language simple, in this entire article, the standalone term Fréchet mean always refers to the intrinsic mean.
While the Fréchet mean reduces to the usual mean vector on Euclidean spaces, it is by far not the only generalization that satisfies this property. A widely used alternative mean concept exists on embedded submanifolds $M \subset \mathbb{R}^k$. There one can define an extrinsic mean as the shortest distance orthogonal projection in the ambient space of the mean in $\mathbb{R}^k$ to $M$. This mean uses not the squared geodesic distance but the squared chordal distance, which is simply the distance in $\mathbb{R}^k$. Bhattacharya and Patrangenaru (2003, 2005), Hendriks and Landsman (1996, 1998), Hotz (2013). We discuss the relation of intrinsic and extrinsic means to the proposed diffusion means below. Also relying on embeddings, residual, Ziezold and Procrustean means have been defined Dryden and Mardia (2016), Gower (1975), Huckemann (2012), Le and Kume (2000), Mardia and Jupp (2000), Ziezold (1994). On Euclidean spaces, embedded in a higher dimensional Euclidean space, all of these reduce to the usual mean vector: the expected vector value.

The mean value concept was extended further by Ziezold (1977) to quasi-metric spaces and later the generalized Fréchet means Huckemann (2011a,b) were introduced for continuous functions $\rho : Q \times P \rightarrow [0, \infty)$ on topological spaces $Q$ and $P$, with $Q$ being the data space and $P$ the descriptor space. Generalized Fréchet means can be understood, e.g. as geodesics or nested lower dimensional subspaces Huckemann and Eltzner (2018). They are defined as minimizing a squared loss function $\rho$ linking the data space to the descriptor space obtaining a meaningful mean set $E^{(\rho)}(X) = \arg\min_{p \in P} \mathbb{E}[\rho(X,p)^2]$.

In this paper we introduce diffusion means on geometric spaces and give a theoretical account for the definition on connected smooth manifolds. Diffusion means have several desirable properties. Like the Fréchet mean, they reduce to the usual mean vector on Euclidean spaces and are defined purely intrinsically, without reference to an embedding. In fact, they do not even require a Riemannian structure, defining isotropic Brownian motion on a manifold equipped with a reference measure suffices for the definition of diffusion means. Furthermore, they are closely tied to a measure of variance, which can be simultaneously estimated. We show that diffusion means for optimal diffusion variance, unlike the intrinsic Fréchet mean, only exhibit smeariness as described by Eltzner and Huckemann (2019) in extreme corner cases and otherwise satisfy a standard CLT with rate $n^{-1/2}$. In consequence, diffusion means estimated simultaneously with the diffusion variance are more reliable than sample Fréchet means.

The diffusion mean value can be realized as a generalized Fréchet mean for $P = Q$ that gives a location statistic for random variables on $Q$. The definition is motivated by the idea of generalizing the maximum likelihood estimate (MLE) of the location parameter of the multivariate normal distribution by maximizing the likelihood of a Brownian motion instead of minimizing squared distance. The Euclidean MLE of the location parameter $\mu$ of $\mathcal{N}(\mu, \Sigma)$ is independent of $\Sigma$ and coincides with the mean vector, and the family of normal distributions are generated by Brownian motions with transition densities being the heat kernel. There is no single canonical generalization of the normal distribution to manifolds, especially for anisotropic $\Sigma$. However, the isotropic case $\Sigma = t \cdot \text{id}$ can be generalized whenever Brownian motions and heat kernels can be defined. Thus, we base the maximum likelihood estimation on the family heat kernels $p(x, \mu, t)$, because they are the transition densities of Brownian motions on manifolds. Similar to the Euclidean case, the unknown parameters $\mu$ can be considered as the mean value and $t$ as the variance parameter. For a fixed $t > 0$, we choose the heat kernel $p$ as our loss function and define the diffusion $t$-mean set as the minimizers

$$E_t(X) = \arg\min_{y \in Q} \mathbb{E}[-\ln p(X,y,t)].$$

The definition is not limited to spaces with a smooth structure and we mention how the definition can be extended to random variables on graphs by only using the existence of random walks. However, in contrast to Euclidean space, the estimated value of $\mu$ in general depends on $t$ and uniqueness is not always guaranteed.
In Euclidean space, the variance $t$ does not influence the estimated mean. While diffusion means on manifolds typically depend on $t$, this does not mean that diffusion means for larger $t$ are less dependent on the data. Instead, with increasing $t$ the diffusion process and thus the diffusion mean is influenced by the manifold’s geometry on an increasingly larger scale.

The logarithm of the heat kernel is naturally connected to the squared geodesic distance by the limit $\lim_{t \to 0} -2t \ln p(x, y, t) = \text{dist}(x, y)^2$ Hsu (2002), which allows us to consider diffusion means as an extension of Fréchet means. We can interpret $t$ as the amount of variance connected to the diffusion mean, where the Fréchet mean represents the limit of vanishing diffusion variance. Throughout the paper we consider the similarities and differences between diffusion means and the Fréchet mean. To exemplify this, we compare in Figure 1 the estimated variance of the Fréchet and diffusion mean on different bimodal Brownian normal distributions.

By sampling from a Brownian motion on $\mathcal{S}^2$ with variance $\sigma^2 > 0$ starting at the north pole $\mu$ and south pole $-\mu$ respectively, we obtain samples from two Brownian normal distributions. Combining the two, we get a bimodal Brownian normal distribution by sampling from the northern distribution with probability $1 - \alpha$ and from the southern with probability $\alpha$. The plots in Figure 1 show the result of estimating the Fréchet mean and the diffusion mean for $t = 0.4, 0.6, 1, 2, 4$ when setting $\sigma = 0.3$ and $\alpha = 0.2$. The Fréchet sample mean is unique at $\mu$, but the estimator appears to converge to the population mean more slowly than with the standard CLT rate $1/\sqrt{n}$.

This slower rate has been termed smearing by Huckemann and Eltzner (2018) and among others, it lowers the power of asymptotic tests. We see a clear improvement in the convergence rate for the diffusion means: Their curves approach a horizontal above the usual CLT rate. This phenomenon has been coined finite sample smeariness (FSS) by Hundrieser, Eltzner and Huckemann (2020), which, among others, still negatively affects powers asymptotic test. Clearly, the increase of FSS of diffusion means wanes with increasing $t$. 

Figure 1. Panel (b) compares the convergence rate of the diffusion means for $t = 0.4, 0.6, 1, 2, 4$ and the Fréchet means on the Brownian normal distribution from Panel (a). With increasing $t$, the blue curve (Fréchet mean) follows 2-smeariness, whereas the curves for the diffusion means approach the usual CLT rate.
Figure 2. The Fréchet mean of the two pole distribution on $S^m$, detailed in Example 6, displayed in (a) for $m = 2$ does not exist for any $\alpha > 0$ as the Riemann center of mass is a latitude ring. For $t > \delta(m)$, the diffusion $t$-mean is $\mu$ whenever $\alpha$ is less than the upper bound $\Lambda_m(t)$ displayed in (b), converging to $\frac{1}{2}$ for all $m \geq 2$.

Contrary to the Fréchet mean, but similar to the extrinsic mean, a unique diffusion mean admits a positive probability mass on the cut locus. In Hotz and Huckemann (2015), it was shown that for distributions on the circle $S^1$, a unique Fréchet mean implies no point mass on the cut locus of the mean, and this result has later been extended further to distributions on complete and connected manifolds Le and Barden (2014). On the sphere $S^m$ of dimension $m \geq 2$, this means that a point $\mu$ cannot be the Fréchet mean of a distribution with nonzero probability mass in the point $-\mu$.

We show that the diffusion means are compatible with positive mass on the cut locus by considering the spherical distribution $P(X = \mu) = 1 - \alpha$ and $P(X = -\mu) = \alpha$ in Figure 2 (a), which we will refer to as the two pole distribution. For $t > 1.1$, we show that the diffusion $t$-mean is $\mu$ whenever $\alpha \leq \Lambda_m(t)$ for a bound $\Lambda_m(t)$ satisfying $\lim_{t \to \infty} \Lambda_m(t) \to \frac{1}{2}$ for all $m \geq 2$.

From a practical viewpoint, a response to the properties such as smeariness of the Fréchet mean is often to use the extrinsic mean. However, the extrinsic mean is per definition dependent on the choice of embedding to which there is often no canonical choice. Compared to this, the diffusion mean being an intrinsic construction is not dependent on an arbitrary choice of embedding while simultaneously often exhibiting less smeariness compared to the Fréchet mean. For the circle and the sphere, we show in Section 3 that the diffusion mean for $t \to \infty$ converges to the extrinsic mean in the standard isometric embeddings. Thus, in these spaces, the diffusion means give an intrinsic interpretation of extrinsic means.

While the Fréchet mean is the center in the sense of minimizing squared distances, the diffusion mean, per its definition, should be interpreted as the most likely explanation of the data under the statistical model of Brownian motion. Since Brownian motions arise in the limit of random walks assuming only i.i.d. steps, this is arguably a very natural model of the dispersion between mean and observations with very weak assumptions on the probabilistic structure. In a sense much weaker than the Fréchet mean which measures dispersion between mean and observations by the energy of smooth geodesics. Both concepts generalize the Euclidean expected value to more general spaces, conserving in one case a metric property (Fréchet mean) and in the other case a probabilistic property (diffusion mean). Using a likelihood criterion for defining means on manifolds has previously been explored in Said and Manton (2012) on compact Lie groups, in Sommer et al. (2017) on landmark manifolds and
general Riemannian manifolds Sommer and Svane (2017) Sommer (2015), and also in probabilistic principal component analysis Zhang and Fletcher (2013) Nye et al. (2017). The contributions of this paper are

1. A theoretical account of diffusion means on geodesically and stochastically complete connected Riemannian manifolds with bounded sectional curvature, and we explain how diffusion means can be defined on Lie groups lacking a bi-invariant Riemannian metric.
2. An exploration of the diffusion mean estimator, in particular the asymptotic behavior in terms of consistency and convergence rate, and an investigation of the effect of joint estimation of mean and diffusion variance on the convergence rate, where the latter essentially rules out smeariness.
3. Proving that diffusion means admit positive mass on the cut locus.
4. Showing that for zero diffusion, diffusion means reduce to Fréchet means.
5. Several examples of improved convergence rate for bimodal distributions on $S^n$.

In Section 2 we cover some basic background knowledge on differential geometry and geometric statistics to set the scene for the upcoming sections. A formal account of the diffusion means can be found in Section 3, covering both geodesically and stochastically complete connected Riemannian manifolds and other possible extensions, which are motivated by the possible application to computational anatomy and network theory, respectively. Furthermore, we prove that the diffusion mean admits positive mass on the cut locus contrary to the Fréchet mean.

Diffusion means, as MLEs, directly allow for the development of Bayes methods. Empirical Bayes methods based on shrinkage for Fréchet means in nonpositive curvature contexts studied by McCormack and Hoff (2022), Yang, Doss and Vemuri (2020) could thus be extended.

An estimator of the diffusion mean is introduced in Section 4. We consider the estimator’s asymptotic behavior and provide sufficient conditions for it to be strongly consistent in the sense of Ziezold (ZC) and Bhattacharya and Patrangenaru (BPC). We also provide sufficient conditions for the generalized CLT to hold for the diffusion estimator, thus giving conditions for $k$-smeariness of the estimator. We consider various bimodal distributions on $S^m$ and calculate the diffusion $t$-means and estimate the asymptotic convergence rate for various $t > 0$. Furthermore, we give precise asymptotic results for the convergence of diffusion means to Fréchet means as $t \to 0$. The results and examples of the paper point to the importance of $t$ and we conclude the section by discussing the estimation of $t$ and the joint estimation of $\mu$ and $t$ parameters. This includes two examples of estimating $t$, one indicating that we are able to estimate the true variance of a Brownian normal distribution. We show that joint estimation of mean and diffusion variance rules out smeariness in all directions in general situations. We illustrate this by reducing finite sample smeariness in geomagnetism data. Finally, we show for the geomagnetism data as well as for wind direction data on the circle that, in consequence, the power of hypothesis tests increases.

2. Background

2.1. Differential geometry

In this section we give a brief introduction to some of the essential notation and definitions in differential geometry. A Riemannian $m$-manifold $(\mathcal{M}, g)$ is a smooth $m$-dimensional manifold $\mathcal{M}$ equipped with a Riemannian metric $g$, i.e. a collection of inner products $g_y(u, v) = \langle u, v \rangle_y$ on each tangent space $T_y\mathcal{M}$, such that $y \mapsto \langle u(y), v(y) \rangle_y$ is smooth for all sections $u, v$ of $T\mathcal{M}$. Unless the dimension $m$ and metric $g$ are important, we omit these to simplify notation. The inner product induces norms on each tangent space, thus allowing for the definition of curve length $L(\gamma) = \int_0^1 \| \frac{d}{dt} \gamma(t) \|_{\gamma(t)} dt$ of a curve...
$\gamma : [0, 1] \to M$. A curve is called a geodesic if it is locally length minimizing everywhere. By assuming that $M$ is complete, there exists at least one length minimizing geodesic between every pair of points $x, y \in M$ and the Riemannian metric induces a distance $\text{dist}(x, y) = \min_{\gamma} L(\gamma)$ where the minimum is taken over all geodesics joining $x$ and $y$.

For each tangent space we define the exponential map $\exp : T_y M \to M$ as the function mapping a tangent vector $v \in T_y M$ to the point reached in unit time by moving along the unique geodesic starting in $y$ with tangent vector $v$. The exponential map is a diffeomorphism in a star-shaped neighborhood of zero and we denote the maximal domain on which it is diffeomorphic as $\mathcal{D}(y) \subset T_y M$. The image $\exp(\mathcal{D}(y)) \subset M$ covers all of $M$ except for the cut locus $C(y)$, which are all points $x$ such that a minimal geodesic from $y$ to $x$ is not longer minimal beyond $x$. Since the exponential map is a diffeomorphism, it has an inverse $\log : M \setminus C(y) \to T_p M$ which again is a diffeomorphism, and it can therefore define a local coordinate system around $y$ whenever $T_y M$ is equipped with a basis.

The notion of Brownian motion extends to Riemannian manifolds, where it can be characterized in numerous ways. One way is through the minimal heat kernel, as it is the transition density of a Brownian motion. The heat kernel is the fundamental solution to the heat equation $\frac{\partial}{\partial t} u(x, y, t) = \frac{1}{2} \Delta u(x, y, t)$ where $\Delta$ is the Laplace Beltrami operator on $M$. The heat equation is not guaranteed to have a unique solution or any solutions at all. In order to ensure existence, we only consider stochastically complete manifolds as defined in the following definition. See Hsu (2002) for more detail.

**Definition 2.1.** A Riemannian manifold $M$ is said to be stochastically complete if there exists a minimal solution $p$ of the heat equation $\frac{\partial}{\partial t} u(x, y, t) = \frac{1}{2} \Delta u(x, y, t)$ satisfying $\int_M p(x, y, t)dy = 1$ for all $x \in M$ and $t > 0$. We say that $p$ is the minimal heat kernel of $M$.

Some of the more important properties of the minimal heat kernel, that we will be using throughout the paper, include:

1. $p$ is strictly positive and $p \in C^\infty(M \times M \times \mathbb{R}^+)$.  
2. Symmetry of the first two variables: $p(x, y, t) = p(y, x, t)$.  
3. $p$ is the transition density function of Brownian motion on $M$.

Being the transition density of a Brownian motion, the heat kernel serves as a density of a generalized normal distribution on which we base our maximum likelihood method in the following section.

### 2.2. Geometric statistics

The field of (Euclidean) statistics widely relies on the assumption that the data can be embedded into a vector space. The assumption is both crucial to definitions such as the mean value, but also to the analysis of asymptotic behavior of estimators. Geometric statistics aims at generalizing the successful tools and results of Euclidean statistics to nonlinear spaces such as manifolds. Due to the different structure of nonlinear spaces, it is not always easy to determine unique or canonical generalizations of concepts from Euclidean statistics, and even basic concepts such as mean value and variance have lead to different generalizations.

One of the most common generalization of the mean to manifolds is the Fréchet mean (set) $M = \text{argmin}_{\gamma \in M} \mathbb{E}[d(X, \gamma)^2]$. Results concerning the existence and uniqueness Karcher (1977)Kendall (1990) pioneered the field by connecting the effect of curvature to the two properties. The Fréchet mean is estimated by the estimator $M_n = \text{argmin}_{\gamma \in M} \sum_i d(X_i, \gamma)^2$. Under certain assumptions, this is a strongly consistent estimator in the sense of both Ziezold (ZC) and Bhattacharya and Patrangenaru
Diffusion means in geometric spaces

(BPC) Ziezold (1977) Bhattacharya and Patrangenaru (2003). An estimator $M_n$ of a set $M$ is said to be a strongly consistent estimator in the sense of Ziezold (ZC) if for almost all $\omega$

$$\cap_{n=1}^{\infty} \overline{\cup_{k=n}^{\infty} M_k(\omega)} \subset M.$$

It is said to be a strongly consistent estimator in the sense of Bhattacharya and Patrangenaru (BPC) if $M \neq \emptyset$ and for every $\varepsilon > 0$ and almost all $\omega$ there is a number $n = n(\varepsilon, \omega) > 0$ such that

$$\cup_{k=n}^{\infty} M_k(\omega) \subset B(M, \varepsilon),$$

where $B(M, \varepsilon)$ is the union of balls of radius $\varepsilon$ around all points in $M$. In Section 4 we show that the diffusion mean estimator is strongly consistent in both senses under the same conditions as the Fréchet mean estimator. We also show that under corresponding assumptions, diffusion means converge for $t \to 0$ to the Fréchet means in the sense of Ziezold and Bhattacharya-Patrangenaru. The former is also called outer limit, the latter one-sided Hausdorff limit, cf. Evans and Jaffe (2020), Schötz (2022).

The first generalization of the CLT to general manifolds was achieved by Bhattacharya and Patrangenaru (BP-CLT) Bhattacharya and Patrangenaru (2005). The BP-CLT establishes sufficient conditions for the Fréchet estimator to share the asymptotic behavior of the Euclidean average. However, the Fréchet mean estimator does not always exhibit this asymptotic behavior. It was first discovered for certain distributions on the circle Hotz and Huckemann (2015), that the estimator converged at a rate $n^{-\alpha}$ for $\alpha < 1/2$ to a non-normal distribution. In order to theoretically account for this behavior, the definition of smeariness for estimators on manifolds and a generalized CLT was formulated by Eltzner and Huckemann (2019). The theorem does not only generalize to means of general loss functions $\mathbb{E}[\rho(y, X)]$, but also allows for general convergence rates $n^{-\alpha}$ and different limiting distribution.

A sequence of random variables $y_n \xrightarrow{p} y$ on a Riemannian manifold is said to be $k$-smeary if the sequence of random real valued vectors $\frac{1}{n^{n-k+1}} \log y_n$ converges to a non-deterministic random vector in $\mathbb{R}^m$, see Eltzner and Huckemann (2019) for more detail. When $k = 0$, the convergence rate coincides with the usual rate $\sqrt{n}$ of the Euclidean average.

The cornerstone in statistics underlying the CLT is the Gaussian normal distribution, due to its frequent occurrence in real data, but also due to its appearance in mean value estimation. One generalization of the normal distribution is the Normal Law Pennec (2006), which utilizes the exponential map to map a normal distribution on $\mathbb{R}^m$ to the manifold. The Brownian normal distribution uses the natural connection between the Brownian motion and normal distribution in Euclidean spaces. The Brownian normal probability distribution coincides with the end point distribution of a Brownian motion, thus we can generate samples from the generalized normal distribution by sampling from a Brownian motion.

3. Diffusion means

Throughout this paper $M$ denotes a geodesically and stochastically complete and connected Riemannian manifold with bounded sectional curvature. This ensures existence of a minimal heat kernel $p$ that is bounded.

In this section we define the diffusion mean of random variables on $M$ and we describe how the diffusion means connect to both the Fréchet mean and the generalized Fréchet means. We provide a registry of manifolds where a closed form for the heat kernel have been obtained. This includes the hyperspheres and the hyperbolic spaces. In Subsection 3.1 we focus on distributions on the sphere and calculate the diffusion means for the two pole distribution in Figure 2a. We conclude the section by mentioning other types of spaces for which the diffusion means can be defined in Subsection 3.2.
The definition of diffusion means is inspired by the Euclidean maximum likelihood estimation and the mean is one of the parameters of a normal density which can be estimated from data. In the general case of a manifold data space, we consider the generalized normal distribution and therefore define diffusion means as the most likely origin points of a Brownian motion hitting the observed data.

For the random variable $X$ on $\mathcal{M}$, with underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we define the log-likelihood function $L_t : \mathcal{M} \to \mathbb{R}$ to be

$$L_t(y) = \mathbb{E}[-\ln(p(X,y,t))] \quad (1)$$

for each $t > 0$ where $p$ denotes the minimal heat kernel from Definition 2.1. This gives rise to the diffusion means as the global minima of the log-likelihood function, i.e. the points maximizing the log-likelihood of a Brownian motion. Notice that we negate the logarithmic heat kernel in the log-likelihood function, in order to characterize diffusion means as minima instead of maxima.

**Definition 3.1.** With the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $X$ be a random variable on $\mathcal{M}$ and fix $t > 0$. The diffusion $t$-mean set $E_t(X)$ of $X$ is the set of global minima of the log-likelihood function $L_t$, i.e.

$$E_t(X) = \arg\min_{y \in \mathcal{M}} \mathbb{E}[-\ln(p(X,y,t))].$$

If $E_t(X)$ contains a single point $\mu_t$, we say that $\mu_t$ is the diffusion $t$-mean of $X$.

Notice, that for each $t > 0$ there is an associated mean set $E_t$, leading to a parametric family of mean sets. In contrast, the Fréchet mean does not have an additional parameter and one gets only one mean set. The diffusion means provide a natural extension of the Fréchet mean, due to the convergence of the diffusion means set to the set of classical Fréchet means in the sense of Ziezold and in the sense of Bhattacharya and Patrangenaru, for which we present sufficient conditions later in Subsection 4.1.

The time parameter $t$ can be interpreted as a *variance* parameter. This is because a Brownian motion is more likely to travel far on the manifold for large $t$, whereas it is more likely to be centralized around its origin for small $t$. This means that for small $t$, the diffusion $t$-mean is sensitive to probability mass far from the mean, which is also the case for the Fréchet mean. For larger $t$ however, this sensitivity is diminished, which we exemplify later in Example 6.

**Remark 1.** In the language of generalized Fréchet means from Huckemann (2011a), the definitions above correspond to choosing $p(x,y)^2 = -\ln(p(x,y,t)\beta_t^{-1})$ for a fixed $t > 0$ and constant $\beta_t > 0$. For this to make sense however, we must ensure existence of a $\beta_t > 0$ such that $\ln(p(x,y,t)\beta_t^{-1}) \leq 0$. If for all $t > 0$, the heat kernel $p(x,y,t)$ is bounded, then letting $\beta_t$ be a bound on $p(x,y,t)$ for each $t > 0$ solves this issue since

$$\arg\min_{y \in \mathcal{M}} \mathbb{E}[-\ln(p(X,y,t)\beta_t^{-1})] = \arg\min_{y \in \mathcal{M}} \mathbb{E}[-\ln(p(X,y,t))].$$

The heat kernel is bounded on compact manifolds, but not necessarily on non-compact complete manifolds. It follows from Cheng, Li and Yau (1981, Theorem 4), that bounded sectional curvature suffices for $\beta_t > 0$ to exist for all $t > 0$, which is why we include this restriction on $\mathcal{M}$. This condition is not limiting, since the existence of such a constant is also necessary for a diffusion $t$-mean to exist.

Instead of fixing $t > 0$, we can also minimize the log-likelihood function $(y,t) \mapsto L_t(y)$ as a function of both $t$ and $y$, thus obtaining the optimal diffusion parameters

$$D(X) = \arg\min_{(y,t) \in \mathcal{M} \times \mathbb{R}_+} \mathbb{E}[-\ln p(X,y,t)]. \quad (2)$$
This paper mostly regards the properties and asymptotic behavior of estimating the diffusion mean by itself, but we do include an example of estimating the parameters simultaneously in Section 5.1.

The closed form of the heat kernel has not been obtained for general manifolds. In the following five examples, we take a look at some of the spaces for which a closed form has been obtained.

**Example 1.** The heat kernel $p$ on the Euclidean space $\mathbb{R}^m$ is given by the function

$$p(x, y, t) = \frac{1}{(2\pi t)^{m/2}} e^{-|x-y|^2/2t}$$

for $x, y \in \mathbb{R}^m$ and $t > 0$. The diffusion $t$-means of a random variable $X$ do not depend on $t$ and coincide with the expected value $\mathbb{E}[X]$ since

$$\arg\min_{y \in \mathbb{R}^m} L_t(y) = \arg\min_{y \in \mathbb{R}^m} \int_{\mathbb{R}^m} -\frac{2}{m} \ln(2\pi t) + \left(\frac{|x-y|^2}{2t}\right) d\mathbb{P}_X(x)$$

$$= \arg\min_{y \in \mathbb{R}^m} \int_{\mathbb{R}^m} |x-y|^2 d\mathbb{P}_X(x) = \arg\min_{y \in \mathbb{R}^m} \mathbb{E}[(X-y)^2].$$

Thus, $\mu_t = \mathbb{E}[X]$ for all $t > 0$.

**Example 2.** The heat kernel on the circle $S^1$ is given by the wrapped Gaussian

$$p(x, y, t) = \frac{1}{\sqrt{2\pi t}} \left( \sum_{k \in \mathbb{Z}} \exp \left( \frac{-(x-y+2\pi k)^2}{2t} \right) \right)$$

for $x, y \in \mathbb{R}/2\pi\mathbb{Z} \cong S^1$ and $t > 0$, and the log-likelihood function for the random variable $X : \Omega \to S^1$ becomes

$$L_t(y) = -\int_{S^1} \ln \left( \frac{1}{\sqrt{2\pi t}} \left( \sum_{k \in \mathbb{Z}} \exp \left( \frac{-(x-y+2\pi k)^2}{2t} \right) \right) \right) d\mathbb{P}_X(x)$$

$$= \ln(\sqrt{2\pi t}) + \int_{S^1} \left[ \frac{(x-y)^2}{2t} - \ln \left( \sum_{k \in \mathbb{Z}} \exp \left( \frac{-2\pi k^2}{2t} \right) \exp \left( \frac{-4\pi k(x-y)}{2t} \right) \right) \right] d\mathbb{P}_X(x).$$

Notably, even on this simple space, the $t$-dependence cannot be split off into a separate summand and $\mu_t$ is therefore explicitly dependent on $t$.

**Example 3.** The heat kernel on the spheres $S^m$ for $m \geq 2$ can be expressed as the uniformly and absolutely convergent series Zhao and Song (2018, Theorem 1)

$$p(x, y, t) = \sum_{l=0}^{\infty} e^{-l(l+m-1)/2} \frac{2l + m - 1}{m - 1} \frac{1}{A_s^m} C_l^{(m-1)/2}(\langle x, y \rangle_{\mathbb{R}^{m+1}})$$

for $x, y \in S^m$ and $t > 0$, where $C_l^m$ are the Gegenbauer polynomials and $A_s^m = \frac{2\pi^{(m+1)/2}}{\Gamma((m+1)/2)}$ the surface area of $S^m$. For $m = 2$, the Gegenbauer polynomials $C_l^{1/2}$ coincide with the Legendre polynomials $P_l^0$ and the heat kernel on $S^2$ is

$$p(x, y, t) = \sum_{l=0}^{\infty} e^{-l(l+1)/4} \frac{2l + 1}{4\pi} P_l^0(\langle x, y \rangle_{\mathbb{R}^3}).$$
Again, \( \mu_t \) is explicitly dependent on \( t \) on these spaces.

**Example 4.** The heat kernel on the hyperbolic space \( \mathbb{H}^m \) can be expressed by the following formulas Grigor’yan and Noguchi (1998) for \( m \geq 1 \). For odd \( m = 2k + 1 \), the heat kernel is given by

\[
p(\mathbf{x}, \mathbf{y}, t) = \frac{(-1)^k}{2^k \pi^k} \int_0^\infty \frac{\partial}{\partial \rho} \left( \frac{1}{\sinh \rho} \frac{1}{\partial \rho} \right)^k e^{-\frac{k^2 \rho_t}{2} - \frac{\rho^2}{8}} \prod \sum_{n=0}^\infty H_n(\mathbf{x}, \mathbf{y}) t^n
\]

where \( \rho = \text{dist}_{\mathbb{H}^m}(\mathbf{x}, \mathbf{y}) \) and for even \( m = 2k + 2 \), it is given by

\[
p(\mathbf{x}, \mathbf{y}, t) = \frac{(-1)^k}{2^{k+1} \pi^{k+\frac{3}{2}}} \rho^{-\frac{3}{2}} e^{-\frac{2k^2 t}{8}} \int_0^\infty \frac{\partial^k}{\partial \rho^k} \left( \frac{1}{\sinh \rho} \frac{1}{\partial \rho} \right)^k \int_0^\infty \frac{\rho^2 e^{-\frac{s^2}{8}}}{(\cosh s - \cosh \rho)\frac{3}{2}} ds.
\]

Again, \( \mu_t \) is explicitly dependent on \( t \) on these spaces.

**Example 5.** The fundamental solution to the heat equation on a Lie group \( G \) of dimension \( m \) is Arede (1991)

\[
p(\mathbf{x}, e, t) = (2\pi t)^{-m/2} \prod_{\alpha \in \Sigma^+} \frac{i\alpha(H)}{2 \sin(i\alpha(H)/2)} \exp \left( \frac{\|H\|^2}{2t} + \frac{\|\rho\|^2 t}{2} \right) \cdot E_x(x \tau > t),
\]

where \( e \) is the neutral element, \( x = \exp(Ad(g)H) \in G - C(e) \) for some \( g \in G \), \( \Sigma^+ \) is the set of positive roots, \( \rho = \sum_{\alpha \in \Sigma^+} \alpha \), and \( \tau \) is the hitting time of \( C(e) \) by \( (x_s)_{0 \leq s \leq t} \). Lie groups are relevant data spaces for many application, e.g. in the modeling of joint movement for robotics, prosthesis development and medicine.

The symmetry in \( x \) and \( y \), which exists by construction, is clearly visible in the examples above. Moreover, when \( t > 0 \) and \( \mathbf{x} \in M \) where \( M = \mathbb{R}^n, S^n \) or \( \mathbb{H}^n \), the heat kernel only depends on the geodesic distance Alonso-Orán et al. (2019).

**Remark 2.** For spaces where a closed form has not been obtained, we can turn to various estimates. We include some of the most important below.

1. For complete Riemannian manifolds of dimension \( m \), we have the asymptotic expansion Hsu (2002, Theorem 5.1.1)

\[
p(\mathbf{x}, \mathbf{y}, t) \sim \left( \frac{1}{2\pi t} \right)^{\frac{m}{2}} e^{-\frac{d(x,y)^2}{8t}} \sum_{n=0}^\infty H_n(\mathbf{x}, \mathbf{y}) t^n
\]

on compact subset with \( x \notin C(y) \). Here \( H_n \) are smooth functions satisfying a recursion formula Chavel (1984) with \( H_0(x,y) = \sqrt{J(\exp_x)(\exp_y^{-1}(y))} \) and \( J \) denoting the Jacobian.

2. Assuming also non-negative Ricci curvature, the heat kernel is bounded Li and Yau (1986, Corollary 3.1)

\[
p(\mathbf{x}, \mathbf{y}, t) \leq \frac{C(\varepsilon)}{Vol(x, \sqrt{t})} \exp \left( -\frac{\text{dist}(x,y)^2}{(2 + \varepsilon) t} \right)
\]

for \( 0 < \varepsilon < 1 \) and constant \( C(\varepsilon) \), where \( Vol(x, \sqrt{t}) \) denotes the volume of the ball around \( x \) of radius \( \sqrt{t} \).
3. Using bridge sampling, the heat kernel can be estimated by the expectation over a guided process Delyon and Hu (2006), Papaspiliopoulos and Roberts (2012). An example of this is the estimated heat kernel on high-dimensional landmark manifolds Sommer et al. (2017). This approach can be carried even further to allow direct sampling of an approximation of the diffusion mean, thus avoiding nested optimization as is in general needed for the Fréchet mean Jensen and Sommer (2022).

It is natural to compare the computational cost of estimating the Fréchet mean and the diffusion mean. The Fréchet mean is by no means simple to compute in spaces without closed form solutions of geodesic distances. In such spaces, the computation in general involves iterative minimization of the Fréchet variance which in turn requires evaluating the Riemannian logarithm at all data points. When these logarithms have no closed form, this results in a nested optimization problem which can be very expensive to solve numerically. The use of bridges where the heat kernel is approximated via Monte Carlo sampling of the Brownian motion conditioned on hitting the data points is quite efficient, even on high-dimensional manifolds. The computational expense is not directly comparable to computing the Riemannian logarithm (it depends on factors including the numerical bridge simulation scheme and number of simulations per bridge). However, as a practical example, the diffusion mean including the variance parameter $t$ can be found via bridge sampling on $S^2$ (not using the heat kernel expansion) for 100 data points in roughly 15 s. on a laptop. In addition to this, the diffusion mean can be approximated by direct sampling, see Jensen and Sommer (2022). This is in general much faster than computing the Fréchet mean, at the expense of additional uncertainty in the mean estimates.\footnote{Example code for both iterative optimization using bridge sampling and direct sampling can be found at https://bitbucket.org/stefansommer/jaxgeometry/src/main/diffusion_mean.ipynb.}

### 3.1. Diffusion means on spheres

We here take a closer look at the diffusion means on the hyperspheres $S^m$ for $m \geq 1$.

#### 3.1.1. Relation to Fréchet mean and extrinsic mean

We prove that the diffusion mean interpolates between the Fréchet mean and the extrinsic mean on spheres as the diffusion time parameter $t$ varies between 0 and $\infty$. The result is illustrated for simulated data in Figure 3. The auxiliary lemmas A.1, A.2, and A.3 can be found in Supplement A.1 (Eltzner et al., 2023).

**Theorem 3.2.** Consider a random variable on $S^m$ with $m \geq 1$. Then the diffusion mean set converges to the extrinsic mean set for $t \to \infty$.

**Proof.** For $S^1$, we use the form for the wrapped normal density shown in Lemma A.1 and set $s := e^{-t/2}$, so the limit $t \to \infty$ is equal to the limit $s \to 0$. Then, consider the function

$$
\rho(x, y, s) := -\frac{1}{s} \log(2\pi p(x, y, t(s))) \\
= -\frac{1}{s} \sum_{k=1}^{\infty} \left( \log \left(1 - s^{2k} \right) + \log \left(1 + 2s^{2k-1} \cos(x - y) + s^{4k-2} \right) \right).
$$

In Lemma A.2 we show that the derivative $\frac{d}{ds} \rho(x, y, s)$ exists and is uniformly bounded for any $x, y \in [-\pi, \pi)$ and $s \in [0, 1/2]$.
For $S^m$ with $m > 1$, let $s := e^{-\frac{m}{2}t}$ and equivalently $t(s) := -\frac{2}{m} \log(s)$. Then, consider the function

$$
\rho(x, y, s) := -1/s \log \left( A_S^m p(x, y, t(s)) \right)
$$

$$
= -1/s \log \left( \sum_{k=0}^{\infty} e^{-k(k+m-1)t(s)/2} \frac{2k + m - 1}{m - 1} c_k^{(m-1)/2} (\langle x, y \rangle_{\mathbb{R}^{m+1}}) \right)
$$

$$
= -1/s \log \left( \sum_{k=0}^{\infty} s^k s^{-k(m-1)/2(m-1)} e^{-k(k+m-1)t(s)/2} \frac{2k + m - 1}{m - 1} c_k^{(m-1)/2} (\langle x, y \rangle_{\mathbb{R}^{m+1}}) \right). \tag{7}
$$

In Lemma A.3 we show that the derivative $\frac{d}{ds} \rho(x, y, s)$ exists and is uniformly bounded for any $x, y \in S^m$ and $s \in [0, 1/2]$. In both cases, we have for any fixed $s$

$$
\text{argmin}_{z \in S^m} L_t(s)(z) = \text{argmin}_{z \in S^m} \mathbb{E}[-\log(p(X, z, t(s)))] = \text{argmin}_{z \in S^m} \mathbb{E}[\rho(x, z, s)].
$$

Furthermore, from the existence and uniform boundedness of the derivative in $s$, we derive a uniform continuity property as follows

$$
\forall \varepsilon > 0 \exists \delta > 0 \forall x, y \in S^m \forall s \in [0, \delta] : |\rho(x, y, s) - \rho(x, y, 0)| < \varepsilon,
$$

because $S^m$ is compact. Now, we can derive

$$
|\mathbb{E}[\rho(X, z, s)] - \mathbb{E}[\rho(X, z, 0)]| \leq \mathbb{E}[|\rho(X, z, s) - \rho(X, z, 0)|] < \mathbb{E}[\varepsilon] = \varepsilon
$$

and thus

$$
y \in \text{argmin}_{z \in S^m} \mathbb{E}[\rho(X, z, \delta)] \Leftrightarrow \mathbb{E}[\rho(X, y, 0)] < \mathbb{E}[\rho(X, y, \delta)] + \varepsilon = \inf_{z \in S^m} \mathbb{E}[\rho(X, z, \delta)] + \varepsilon
$$
Diffusion means describe centers of diffusion processes, which suggests interpreting the intrinsic and extrinsic means as centers of diffusion in the limits of vanishing and diverging diffusion variance \( t \), respectively. If one estimates the diffusion means simultaneously with the variance, this will yield a diffusion mean close to the intrinsic mean for very concentrated data and close to the extrinsic mean for very spread out data. Based on this, if the choice is only between intrinsic and extrinsic mean, one may heuristically suggest to prefer intrinsic means as location statistics for concentrated data and extrinsic means for very spread out data.

In Huckemann (2012) it was found that for rather concentrated data, the \textit{residual mean} is close to a geodesic passing through the extrinsic and intrinsic mean but on the opposite side of the extrinsic mean from the intrinsic mean. Our findings raise the question for future research, whether diffusion means also connect to residual means.

\[
< \inf_{z \in \mathcal{S}^m} \mathbb{E}[\rho(X,z,0)] + 2\varepsilon. 
\]

It is thus clear that in the limit \( \varepsilon \to 0 \), the set \( \argmin_{z \in \mathcal{S}^m} \mathbb{E}[\rho(X,z,\delta)] \) must converge to the set \( \argmin_{z \in \mathcal{S}^m} \mathbb{E}[\rho(X,z,0)] \) in the Hausdorff topology.

This ensures that

\[
\lim_{s \to 0} \argmin_{z \in \mathcal{S}^m} \mathbb{E}[\rho(X,z,s)] = \argmin_{z \in \mathcal{S}^m} \mathbb{E}[\rho(X,z,s)].
\]

For the circle, we recall Equation (6) and use the expansion \( \log(1 + x) = x + O(x^2) \)

\[
\rho(x,y,s) = -\frac{1}{s} \sum_{k=1}^{\infty} \left( -s^{2k} + O(s^{4k}) + 2s^{2k-1} \cos(x-y) + O(s^{4k-2}) \right) = -2 \cos(x-y) + O(s).
\]

Thus \( \rho(x,y,0) = -2 \cos(x-y) \). Embedding \( \mathcal{S}^1 \) in \( \mathbb{R}^2 \) via setting \( W = (\cos X, \sin X), u = (\cos y, \sin y) \), we have that

\[
\argmin_{z \in \mathcal{S}^1} \mathbb{E}[\rho(X,z,0)] = \argmax_{u \in \mathbb{R}^2, \|u\|=1} \mathbb{E}[\langle W,u \rangle].
\]

Constrained maximization yields the maximizer \( \frac{\mathbb{E}[W]}{\|\mathbb{E}[W]\|} \) in case of \( \mathbb{E}[W] \neq 0 \) and all of \( \mathcal{S}^1 \) else. In both cases this is the extrinsic mean, as asserted.

For the sphere, to see that the diffusion mean approaches the extrinsic mean as \( t \to \infty \), we highlight the lowest orders in \( s \) in the series expansion (7) and use \( \log(1 + x) = x + O(x^2) \) as well as the Gegenbauer polynomial \( C_1^{(m-1)/2} = (m-1)x \)

\[
\rho(x,y,s) = -\frac{1}{s} \log \left( 1 + s(m+1)\langle x,y \rangle_{\mathbb{R}^{m+1}} + O(s^2) \right) = -(m+1)\langle x,y \rangle_{\mathbb{R}^{m+1}} + O(s),
\]

so that

\[
\argmin_{z \in \mathcal{S}^m} \mathbb{E}[\rho(X,z,0)] = \argmax_{z \in \mathbb{R}^{m+1}, \|z\|=1} \mathbb{E}[\langle X,z \rangle].
\]

Constrained maximization yields again the maximizer \( \frac{\mathbb{E}[X]}{\|\mathbb{E}[X]\|} \) in case of \( \mathbb{E}[X] \neq 0 \) and all of \( \mathcal{S}^m \) else. In both cases this is the extrinsic mean, as asserted. \( \square \)
3.1.2. Two pole distribution

We now turn to studying diffusion means of the two pole distribution in Figure 2a, particularly because the behavior of the Fréchet mean has been studied for this distribution. We begin by introducing some useful notation and the choice of coordinates that will be used throughout the paper.

We choose non-standard polar coordinates \( \theta_1, \ldots, \theta_m \in [-\pi/2, \pi/2] \) and \( \phi \in [-\pi, \pi) \) of \( S^m \)

\[
x = \begin{pmatrix} - \left( \prod_{j=1}^{n-2} \cos \theta_j \right) \cos \phi \\ - \left( \prod_{j=1}^{n-2} \cos \theta_j \right) \sin \phi \\ \vdots \\ - \cos \theta_1 \cos \theta_2 \sin \theta_3 \\ - \cos \theta_1 \sin \theta_2 \\ \sin \theta_1 \\ \end{pmatrix},
\]

such that the north pole \( \mu = (0, 1, 0, \ldots, 0) \) has polar coordinates \( (0, \ldots, 0, -\pi/2) \). Using these coordinates we may, due to rotation symmetry around the north pole, assume w.l.o.g.

that any fixed point \( y \in S^m \) has coordinates \( (0, \ldots, 0, -\pi/2 + \delta) \) for some \( \delta \in [0, \pi] \). We denote these coordinates by \( y_\delta \).

**Remark 3.** It follows from Alonso-Orán et al. (2019, Theorem 1.3), that the map \( \delta \in [0, \pi] \mapsto p(\mu, y_\delta, t) \) is decreasing for each fixed \( t > 0 \), since the geodesic distance is \( \text{dist}(\mu, y_\delta) = \arccos(\mu, y_\delta) = \delta \) for \( \delta \in [0, \pi] \).

To simplify the notation of the heat kernel on \( S^m \) as written in Equation (4) and the logarithmic heat kernel, we define the functions \( h_{t,m}, \ell_{t,m} : [-1, 1] \to \mathbb{R} \) for each \( t > 0 \) and integer \( m \geq 2 \) by

\[
h_{t,m}(x) = \sum_{l=0}^{\infty} e^{-l(l+m-1)\frac{1}{2} t} \frac{2l + m - 1}{m - 1} A_m^{(m-1)/2}(x), \quad \ell_{t,m}(x) = \ln h_{t,m}(x).
\]

Then \( p(x, y, t) = h_{t,m}(\langle x, y \rangle_{\mathbb{R}^{m+1}}) \) and the log-likelihood function \( L_t \) of a random variable \( X \) on \( S^m \) with density \( p_X(x) \) can written in terms of \( \ell_{t,m} \) as

\[ L_t(y) = \int_{S^m} -\ell_{t,m}(\langle x, y \rangle_{\mathbb{R}^{m+1}}) p_X(x) dx. \]

The following two lemmas uncover some useful properties of \( h_{t,m} \) and \( \ell_{t,m} \). The lemmas are proved in Supplement A.2.

**Lemma 3.3.** The derivative \( \frac{d}{dx} \ell_{t,m}(x) \) is positive for all \( m \geq 2 \) and \( t > 0 \). For all integers \( m \geq 2 \), define

\[
\delta(m) = \begin{cases} \frac{2}{m+1} \ln ((m+1)+4(m+3)), & \text{if } m = 2, 3, \\ \ln (\frac{32(m+3)}{9(m+1)}), & \text{if } m \geq 4. \end{cases}
\]

The second derivative \( \frac{d^2}{dx^2} \ell_{t,m}(x) \) is negative for all \( m \geq 2 \) and \( t > \delta(m) \).
In this section we take a look at how the diffusion to random variables on spaces where it is possible to take identically distributed and independent steps. The assumption that the data space is a well-behaved smooth manifold is convenient for the existence of a heat kernel, but the assumption is not necessary for the construction. Diffusion means do not depend on the smooth structure of a manifold or even the existence of a metric. The definition can be extended to random variables on spaces where it is possible to take identically distributed and independent steps. In this section we take a look at how the diffusion $t$-means can be defined for random variables on more general spaces.

**Lemma 3.4.** For $m \geq 2$ and $n = 0, 1$, the map $\frac{dh_t}{dx^n} h_{t,m}(x)$ is positive for all $t > 0$. For $n \geq 2$ the map is positive for $t \geq 1$.

Example 6. Fix $m \geq 2$ and $t > \delta(m)$, and consider the random variable $X$ on $S^m$ distributed according to the two pole distribution $P(X \in -\mu) = \alpha$ and $P(X \in \mu) = 1 - \alpha$ displayed in Figure 2a. The log-likelihood function of $X$ at $y_\delta$ can be expressed as a map of $\delta \in [0, \pi]$, $L_{t,m}(y_\delta) = -(1 - \alpha)\ell_{t,m}(\cos \delta) - \alpha \ell_{t,m}(-\cos \delta) =: \tilde{L}_{t,m}(\delta)$, since $\langle y_\delta, \mu \rangle_{\mathbb{R}^m} = \cos \delta$. Figure 4 provides a plot of the graph for $t = 3$ and various values of $\alpha$. The first derivative with respect to $\delta$ is

$$\tilde{L}_{t,m}'(\delta) = \sin \delta \left( (1 - \alpha)\ell_{t,m}'(\cos \delta) - \alpha \ell_{t,m}'(-\cos \delta) \right).$$

Clearly $\tilde{L}_{t,m}'(0) = \tilde{L}_{t,m}(\pi) = 0$. For $t > \delta(m)$ as defined in Equation (9), it follows by Lemma 3.3 that $\tilde{L}_{t,m}'(\delta)$ is strictly positive on $(0, \pi)$ whenever

$$\alpha \leq \Lambda_m(t) := \frac{h_{t,m}(-1)\ell_{t,m}'(1)}{h_{t,m}'(-1)h_{t,m}(1) + h_{t,m}(-1)\ell_{t,m}'(1)}.$$ 

Thus, there is a unique minimum of $L_{t,m}$ at $\mu$ for $\alpha \leq \Lambda_m(t)$, which means that $\mu$ is the unique diffusion $t$-mean when the weight is sufficiently small. The functions $\Lambda_m$ can be seen in Figure 2b for $m = 2, 3, 4, 5$. Furthermore, we have $\lim_{m \to \infty} \Lambda_m(t) = \frac{1}{2}$ for all $m \geq 2$, which implies that for any $\alpha < \frac{1}{2}$ there exists a $t_\alpha > 0$ such that $\mu$ is the unique diffusion $t_\alpha$-mean of the distribution with this particular $\alpha$. This means that we can pick the parameter $t$ to allow for almost arbitrary mass on the opposite pole of the mean. Finally, let us consider what happens for $\alpha = \frac{1}{2}$, where $\tilde{L}_{t,m}$ simplifies to

$$\tilde{L}_{t,m}'(\theta) = \frac{1}{2} \sin \theta \left( \ell_{t,m}'(\cos \theta) - \ell_{t,m}'(-\cos \theta) \right).$$

Clearly $\tilde{L}_{t,m}'(0) = \tilde{L}_{t,m}'(\pi/2) = \tilde{L}_{t,m}(\pi) = 0$. Still assuming $t > \delta(m)$, it follows directly from Lemma 3.3 that $\ell_{t,m}'(\cos \theta) - \ell_{t,m}'(-\cos \theta)$ is negative for $\theta \in (0, \pi/2)$ and positive for $\theta \in (\pi/2, 0)$. This implies that $\delta = \pi/2$ is the only minimum of $\tilde{L}_{t,m}$ and that the diffusion $t$-mean set for $t > \delta(m)$ is the set of all of the points on the equator between $\mu$ and $-\mu$.

### 3.2. Diffusion means on Lie groups and graphs

The assumption that the data space is a well-behaved smooth manifold is convenient for the existence of a heat kernel, but the assumption is not necessary for the construction. Diffusion means do not depend on the smooth structure of a manifold or even the existence of a metric. The definition can be extended to random variables on spaces where it is possible to take identically distributed and independent steps. In this section we take a look at how the diffusion $t$-means can be defined for random variables on more general spaces.
Figure 4. The log-likelihood function for the two pole distribution on $S^2$ with $t = 3$ and weight $\alpha = \Lambda_2(t) + b$ on the south pole for various $b$.

1. Defining mean sets on non-Riemannian affine connection manifolds is of great interest, due to the direct application in fields such as computational anatomy. There exists a symmetric bi-invariant connection on all Lie groups, which allows for an affine connection structure of the Lie group which may not be metric, meaning that the Riemannian barycenters are not defined. Due to the affine connection structure, it is still possible to define a mean value, for example through exponential barycenters Pennec (2015, 2018). As diffusion means only depend on the likelihood resulting from a Brownian motion, this structure also allows for random walks to exist, therefore also for the possibility of defining diffusion means. A specific example of such a non-compact Lie group is the special Euclidean group $SE(n)$ (rigid transformations in $\mathbb{R}^n$) of which there exists no bi-invariant Riemannian metric Pennec and Arsigny (2013). In contrast, the bi-invariant connection defines parallel transport and horizontal vector fields on the frame bundle $FSE(n)$ of the group. A Brownian motion can be defined by stochastic development using the horizontal fields, and a choice of measure on the group implies a smooth solution to the corresponding heat equation on a sub-bundle of $FSE(n)$. This in turn defines a version of heat flow on $SE(n)$ and thereby a diffusion mean.

2. Motivated by its possible application in network theory, the definition of diffusion means can also be extend to graphs. Let $G = (V, E)$ be a graph of $n = |V|$ vertices and $m = |E|$ edges and let $d_i$ denote the number of edges connected to vertex $v_i$ and $d_{ij}$ denote the number of edges between $v_i$ and $v_j$. Define $P$ to be the matrix with entries $P_{ij} = \frac{d_{ij}}{d_i}$. This is exactly the transition density of a random walk on $G$ for $t = 1$. Allowing for an arbitrary number of steps $t \in \mathbb{N}$, the transition density is $P^t$. Thus, starting in $v_i$, the probability of hitting $v_j$ in time $t$ with a random walk is $(P^t)_{ij}$.

The distribution of a stochastic variable $X$ on the graph $G$ is given by the probabilities $\Pr(X = v_j)$ for $j = 1, \ldots, n$. To define diffusion $t$-mean sets, we want to find the $i$ maximizing the likelihood function $L_t(i) = \sum_{j=1}^n \Pr(X = v_j)(P^t)_{ij} = e_iP^tp^X$ where $p^X \in \mathbb{R}^n$ has entries $p^X_j = \Pr(X = v_j)$. Thus for each $t \in \mathbb{N}$, we define the diffusion $t$-means of $X$ to be $E_t = \arg\min_{1 \leq i \leq n} e_iP^tp^X$.

4. Asymptotic analysis

In this section we define the diffusion estimator and investigate its asymptotic behavior. For manifolds of bounded sectional curvature, we present sufficient conditions for strong consistency in the sense of Ziezold and in the sense of Bhattacharya and Patrangenaru as defined in Subsection 2.2. The conditions
Diffusion means in geometric spaces

rely on the interpretation of diffusion means as generalized Fréchet means, and are no more restrictive than in the case of the estimator

$$M_N = \arg\min_{y \in M} \frac{1}{N} \sum_{i=1}^{N} \text{dist}(y, x_i)^2$$

of the Fréchet mean set. For compact manifolds these conditions always hold. Next, we present sufficient conditions for the estimator to satisfy the general CLT of Eltzner and Huckemann (2019) and consider the consequence on two spherical distributions.

**Definition 4.1.** Let $X$ be a random variable on $\mathcal{M}$. Let $p$ be the minimal heat kernel on $\mathcal{M}$. For samples $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$ we define the sample log-likelihood function $L_{t,n} : \mathcal{M} \rightarrow \mathbb{R}$,

$$L_{t,n}(y) = -\frac{1}{n} \ln \left( \prod_{i=1}^{n} p(X_i, y, t) \right) = -\frac{1}{n} \sum_{i=1}^{n} \ln p(X_i, y, t)$$

for every $n \in \mathbb{N}$ and the sample diffusion $t$-mean sets of $X_1, \ldots, X_n$ are the sets

$$E_{t,n} = \arg\min_{y \in \mathcal{M}} L_{t,n}(y).$$

Due to continuity of $p$ in the second component, every $E_{t,n}$ is a closed set. Considering the collection $\{E_{t,n}\}$ as random closed sets in the sense of Molchanov (2005), they give rise to an $\mathcal{M}$-estimator of the diffusion $t$-mean set $E_t$, which we call the **diffusion estimator**. The following lemmas present sufficient conditions for the diffusion estimator to be strongly consistent.

**Lemma 4.2.** Let $X : \Omega \rightarrow \mathcal{M}$ be a random variable on $\mathcal{M}$ and fix $t > 0$. The diffusion estimator $E_{t,n}$ of $E_t(X)$ is strongly consistent in the sense of Ziezold (ZC) if either of the two following conditions hold:

1. $X$ has compact support.
2. The map $y \mapsto \ln(p(x, y, t))$ is uniformly continuous and $\mathbb{E}[-\ln p(x, X, t)] < \infty$ for all $x \in \mathcal{M}$.

**Proof.** This follows immediately from Huckemann (2011a, Theorem A.3) by Remark 1. \hfill $\square$

**Lemma 4.3.** Let $X : \Omega \rightarrow \mathcal{M}$ be a random variable on $\mathcal{M}$ and fix $t > 0$. The diffusion estimator $E_{t,n}$ is strongly consistent in the sense of Bhattacharya and Patrangenaru (BPC) if $E_t \neq \emptyset$ and the following conditions hold:

1. $E_{t,n}$ satisfies (ZC)
2. Every closed bounded subset of $\bigcup_{n=1}^{\infty} E_{t,n}$ is compact (The Heine-Borel property)
3. There exist $y_0 \in \mathcal{M}$ and $C > 0$ such that $P(-\ln p(X, y_0, t) < C) > 0$ and for every sequence $y_n \in \mathcal{M}$ with $d(y_0, y_n) \rightarrow \infty$ there is a sequence $R_{t,n} \rightarrow \infty$ with $-\ln p(x, y_n, t) > R_{t,n}$ for all $x \in \mathcal{M}$ with $-\ln p(x, y_0, t) < C$.

**Proof.** This follows directly from Huckemann (2011a, Theorem A.4) by Remark 1. \hfill $\square$

As for general results on manifolds, the situation simplifies when the manifold is compact. This is also the case for the previous two lemmas. In fact, the conditions are all satisfied in this case and can be dropped completely.
Corollary 4.4. Fix \( t > 0 \) and let \( X \) be a random variable on \( \mathcal{M} \) and assume that \( \mathcal{M} \) is compact. Then, the diffusion estimator \( E_{t,n} \) satisfies (ZC) and (BPC).

Proof. Fix \( t > 0 \). The estimator satisfies (ZC) because \( p_t(x,y) = -\ln(p(x,y,t)) \beta_t \) is smooth on a compact domain, hence also uniformly continuous and bounded. For (BPC), the three assumptions 1.-3. of Lemma 4.3 follow directly from \( \mathcal{M} \) being compact. Also, this ensures \( E_t \neq \emptyset \).

We will now move on from the topic of strong consistency to a generalized CLT for the diffusion estimator. Before we state the CLT for the diffusion estimator, we introduce some notation. Fix \( t > 0 \) and let \( X \) be a random variable with unique diffusion \( t \)-mean \( \mu_t \in \mathcal{M} \). For a measurable selection \( \mu_{t,n} \in E_{t,n} \), define the sequence \( x_{t,n} \in T_{\mu_t} \mathcal{M} \) by

\[
x_{t,n} = \begin{cases} 
\log_{\mu_t}(\mu_{t,n}) & \text{if } \mu_{t,n} \notin C(\mu_t) \\
0 & \text{else}
\end{cases}
\]  

(11)

Define also \( \tilde{L}_t : D(\mu_t) \to \mathbb{R} \) by \( \tilde{L}_t(x) = L_t(\exp_{\mu_t}(x)) \), which can be extended continuously to all of \( T_{\mu_t} \mathcal{M} \). Also, for each \( y \in \mathcal{M} \) we define the map \( q_t(x,y) = -\ln p(\exp_{\mu_t}(x), y, t) \).

Theorem 4.5 (General CLT). Fix \( t > 0 \) and let \( X \) be a random variable on the manifold \( \mathcal{M} \) of dimension \( m \). Assume that \( \mu_t \) is the unique diffusion \( t \)-mean and that for every measurable selection \( \mu_{t,n} \) of \( E_{t,n} \) we have \( \mu_{t,n} \overset{P}{\to} \mu_t \), and define \( x_{t,n} \) as in Equation (11). Furthermore, assume that there exists \( 2 \leq r \in \mathbb{R} \), a rotation matrix \( R \in SO(m) \) and non-zero vectors \( T_1, \ldots, T_m \) such that

\[
\tilde{L}(x) = \tilde{L}_t(0) + \sum_{i=1}^{m} T_i ||(Rx)_i||^r + o(||x||^r).
\]  

(12)

Then for every measurable selection \( \mu_{t,n} \in E_{t,n} \) it holds that

\[
\sqrt{n}(V_1, \ldots, V_n)^T \overset{D}{\to} N(0, \frac{1}{r^2} T^{-1} \Cov[\grad_x q_t(x, X)|_{x=0}] T^{-1}),
\]  

(13)

where \( V_i = (Rx_n)_i ||(Rx_n)_i||^{r-2} \) and \( T = \text{diag}(T_1, \ldots, T_m) \).

Proof. The result follows immediately from Eltzner and Huckemann (2019, Theorem 11) if all assumptions are met. Given the two assumptions above, the only remaining non-trivial condition is the almost surely locally Lipschitz and differentiability at the mean. Both properties hold since \( p \) is smooth and thereby locally Lipschitz.

This Theorem is used as the basis for the definition of smeariness. Here, we introduce the term for compact manifolds.

Definition 4.6. Fix \( t > 0 \) and let \( X \) be a random variable on the compact manifold \( \mathcal{M} \). If \( E_t = \{ \mu_t \} \) and the Taylor expansion of Equation (12) exists and thus the limiting distribution as in Equation (13) holds for every measurable selection \( \mu_{t,n} \), then the diffusion estimator of \( X \) is called \( k \)-th order smeary or simply \( k \)-smeary with \( k = r - 2 \).

As mentioned, several cases of smeariness have been discovered for the Fréchet mean. In the following examples we investigate the existence of smeariness for the diffusion means and also the influence...
of the time parameter $t$. We consider the order of smeariness as introduced in Definition 4.6 on the two pole distribution in Example 6 and on a bimodal Brownian normal distribution in Example 8. In Section 5.1, we consider the convergence rate of the diffusion estimator on a real data set. Lastly, in Supplement B, we examine a previously studied distribution Eltzner and Huckemann (2019).

Example 7. Let $X : \Omega \to S^m$ have the two pole distribution from Example 6 and fix $t > \delta(m)$. As proved earlier, the point $\mu$ is the unique diffusion $t$-mean whenever $\alpha \leq \Lambda(t)$. In order to determine the asymptotic rate of the diffusion estimator we calculate the second derivative of $\tilde{L}_{t,m}$ at $\delta = 0$. The second derivative

$$
\tilde{L}''_{t,m}(\delta) = \cos \delta \left( -\alpha \ell'_t, m(-\cos \delta) + (1 - \alpha) \ell''_t, m(\cos \delta) \right) + \sin \delta^2 \left( -\alpha \ell''_t, m(-\cos \delta) - (1 - \alpha) \ell'_t, m(\cos \delta) \right)
$$

vanishes at $\delta = 0$ exactly for $\alpha = \Lambda(t)$ and the third derivative vanishes for all values of $\alpha$. Using a Taylor expansion of $\tilde{L}$ to achieve the expansion in Equation (12), this means that the regular CLT rate of convergence applies for $\alpha \leq \Lambda(t)$, whereas the estimator is at least 2-smeary for $\alpha_t = \Lambda(t)$, if $\tilde{L}_{t,m}$ still takes its global minimum at $\delta = 0$. According to Lemma 3.3, this is satisfied, if $t > \delta(m)$.

Example 8. In this example we consider the diffusion means of a bimodal Brownian normal distribution arising from the Brownian motion on $S^2$ visualized in Figure 1a. We sample from a Brownian normal distribution on $S^2$ with variance $\sigma = 0.3$ in both poles $\mu$ and $-\mu$, with total probability mass 0.8 in the north pole $\mu$ and 0.2 in the south pole $-\mu$. The estimated variance $\tilde{V}$ for the estimators of the Fréchet mean and the diffusion $t$-mean for $t = 0.4, 0.6, 1, 2, 4$ is plotted in Figure 1b. This variance is scaled by sample size $n$ and the dotted line represents the convergence rate of a 2-smeary estimator and the black line the standard CLT rate of $\sqrt{n}$. The weight $\alpha = 0.2$ was chosen such that the Fréchet estimator is smeary to exemplify the behavior of the diffusion estimator in such a case. Figure 1b clearly indicates that the diffusion estimators have lower convergence rates and that an increase in $t$ leads to a faster converging diffusion estimator.

### 4.1. Small $t$ asymptotics

For complete Riemannian manifolds, the uniform convergence Hsu (2002, Theorem 5.2.1) on compact subsets

$$
\lim_{t \to 0} -2t \ln p(x, y, t) = \text{dist}(x, y)^2
$$

connects the logarithmic heat kernel and the squared distance for arbitrary points $x, y \in M$. For fixed $t > 0$, the scalar $2t$ does not affect the minima, so the sets $E_t(X)$ and $\text{argmin}_{y \in M} \mathbb{E}[-2t \ln p(X, y, t)]$ coincide. For $t = 0$ we have the set of classical Fréchet means

$$
M = \text{argmin}_{y \in M} \mathbb{E}[	ext{dist}(X, y)^2].
$$

We set

$$
\rho_t(x, y) := -2t \log p(x, y, t) \quad \text{and} \quad F_t(y) := \mathbb{E}[\rho_t(X, y)] \quad \text{for} \quad t > 0, \quad \text{as well as} \quad 
\rho_0(x, y) := \text{dist}(x, y)^2 \quad \text{and} \quad F_0(y) := \mathbb{E}[\rho_0(X, y)],
$$
and assume that \(M\) allows for an exhaustion by countably many compact sets \(K_j \subseteq M\) such that for every \(y \in M\) there is \(T(y) > 0\) such that the following local tightness condition holds

for every \(\varepsilon > 0\) there is \(j(\varepsilon) \in \mathbb{N}\) so for all \(j \geq j(\varepsilon), t \in [0, T(y)]\): \(\mathbb{E}[\rho_t(X, y) \cdot 1\{X \in K_j\}] < \varepsilon\).  

(15)

**Theorem 4.7.** Assume that \(M\) is a complete Riemannian manifold that allows for an exhaustion by countably many compact sets satisfying the local tightness condition (15), this is the case if \(M\) is compact, then

\[ M \neq \emptyset. \]

Further, for every sequence \(0 < t_k \to 0 \ (k \to \infty)\) the sets of diffusion means converge to the set of classical Fréchet means in the sense of Ziezold, i.e.

\[ \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} E_{t_k} \subseteq M. \]

If additionally

1. every closed bounded subset of \(\bigcup_{k=1}^{\infty} E_{t_k}\) is compact (the Heine-Borel property),
2. there exist \(y_0 \in M\) and \(C > 0\) such that \(P(\rho_t(X, y_0) < C) > 0\) and for every sequence \(y_n \in M\) with \(d(y_0, y_n) \to \infty\) there is a sequence \(R_n \to \infty\) with \(\rho_t(x, y_0) > R_n\) for all \(x \in M\) with \(\rho_t(x, y_0) < C\) (coercivity),

then convergence takes place in the sense of Bhattacharya and Patrangenaru, i.e. for every \(\varepsilon > 0\) there is a number \(n = n(\varepsilon, \omega) > 0\) such that

\[ \bigcup_{k=n}^{\infty} E_{t_k} \subset B(M, \varepsilon), \]

where \(B(M, \varepsilon)\) is the ball of radius \(\varepsilon\) around \(M\).

**Proof.** A self contained proof is given in Supplement A.3. \(\square\)

### 4.2. Estimating \(t\)

In previous sections we have discussed the estimation of the diffusion means for fixed time-parameter \(t\) and it has been exemplified that increasing \(t\) in terms of diffusion mean set and asymptotic rate of the diffusion estimator. In this section, we will discuss the estimation of \(t\) for fixed mean.

A way to select an appropriate \(t\) is to apply the same maximum likelihood method that was used to define the diffusion means. We define the diffusion variance set \(T(y)\) at a point \(y \in M\) to be

\[ T(y) = \arg \min_{t > 0} L_t(y). \]

Letting \(\mu_s\) be a diffusion \(s\)-mean for \(s > 0\), the derivative \(\frac{d}{dt} L_t(\mu_s)\) can be written in terms of the Laplace Beltrami operator \(\Delta\) due to \(p\) being a solution to the heat equation, i.e. \(\frac{d}{dt} p = \frac{1}{2} \Delta p:\)

\[ \frac{d}{dt} \int_{M} -\ln p(x, \mu_s, t) d\mathbb{P}^X(x) = -\int_{M} \frac{d}{dt} \frac{p(x, \mu_s, t)}{p(x, \mu_s, t)} d\mathbb{P}^X(x) = -\int_{M} \frac{1}{2} \Delta p(x, \mu_s, t) \frac{p(x, \mu_s, t)}{p(x, \mu_s, t)} d\mathbb{P}^X(x). \]

(16)
See also e.g. Hsu (2002). This means that the local minima of \( L_t(\mu_s) \) w.r.t. \( t \) can be found using the gradient descent algorithm

\[
t_{n+1} = t_n + \frac{\beta}{2} \int_{\mathcal{M}} \frac{\Delta p(x, \mu_s, t)}{p(x, \mu_s, t)} d\mathbb{P}^X(x)
\]

with learning rate \( \beta > 0 \). As an illustration, we revisit the two pole distribution from Example 6.

**Example 9.** Let \( X \) have the two-pole distribution as in Example 6. The Laplace Beltrami operator on \( S^2 \) with respect to the coordinates of Equation (8) is

\[
\Delta_{S^2} f(\theta, \phi) = \frac{1}{\sin(\phi + \pi)} \frac{d}{d\phi} \left( \sin(\phi + \pi) \frac{d}{d\phi} f \right) + \frac{1}{\sin(\theta + \pi/2)} \frac{d^2}{d\theta^2} f.
\]

(17)

Fixing \( \mu_t = \mu = (0,1,0) \) and recalling that the heat kernel only depends on \( \phi = -\pi/2 + \delta \), we get an expression for the Laplace Beltrami operator of the heat kernel

\[
\Delta_{S^2} p(\mu, s, t) = \frac{1}{\sin(\pi/2 + \delta)} \frac{d}{d\delta} \left( \sin(\pi/2 + \delta) \frac{d}{d\delta} h_{t,2}(\cos \delta) \right)
\]

\[
= \sin \delta^2 \frac{d^2}{d\delta^2} h_{t,2}(\cos \delta) - \frac{\cos \delta^2 - \sin \delta^2}{\cos \delta} h_{t,2}(\cos \delta).
\]

Thus, for a fixed \( \alpha \in [0, \frac{1}{2}] \) we have

\[
\frac{d}{dt} L_t(\mu) = -\alpha \frac{h_{t,2}'(-1)}{h_{t,2}(-1)} + (1 - \alpha) \frac{h_{t,2}'(1)}{h_{t,2}(1)} = -\alpha \ell_{t,2}'(-1) + (1 - \alpha) \ell_{t,2}'(1).
\]

We now consider \( \alpha = \Lambda(s) \) for some \( s > 0 \). We then recognize this expression from the derivative of the log-likelihood function in Equation (10) and conclude that \( \frac{d}{dt} L_t(\mu)\big|_{t=s} = 0 \), the value which leads to a smeary \( \mu \). The derivative \( \frac{d}{dt} L_t(\mu)\big|_{t=r} \) is positive for \( r > s \) and negative for \( r < s \), which indicates a local minimum such that the diffusion variance set at \( \mu \) is \( \mathcal{T}(\mu) = \{s\} \).

Joint estimation of \( \mu \) and \( t \) can therefore only lead to smeariness for \( t = s \) or non-uniqueness for \( t < s \) of the diffusion mean. Joint estimation of parameters is treated numerically in Figure 5a below.

### 4.3. Removing smeariness: Joint estimation of \( \mu \) and \( t \)

Analogous to the Gaussian MLE in Euclidean space, it is possible to estimate \( \mu_t \) and \( t \) jointly, to get

\[
(\mu_n, t_n) = \arg\min_{t > 0, \mu \in \mathcal{M}} L_{t,n}(\mu).
\]

(18)

Such an estimation was performed on a data set on \( S^1 \) by Hansen, Eltzner and Sommer (2021). It was found that the diffusion mean \( \mu_n \) jointly estimated with \( t_n \) exhibits a considerably lower magnitude of finite sample smeariness (introduced in Section 5.1) than the Fréchet mean.

In fact, it turns out that non-smeariness of diffusion means with jointly estimated \( \mu \) and \( t \) is a very general phenomenon as shown in the following result. We state it for the population mean, but the result holds similarly for sample estimators minimizing (18).

**Theorem 4.8.** Assume \( \mu, t \) minimizes the likelihood (1). Then

\[
\Delta_{\mu} \mathbb{E}[-\ln p(X, \mu, t)] = \mathbb{E}[\| \nabla_x \ln p(x, \mu, t)_{x=X} \|^2].
\]
Proof. Since $p$ is a solution to the heat equation, we can continue (16) to see that $-\ln p$ satisfies

$$\frac{d}{dt} \int_M -\ln p(x, \mu, t) d\mathbb{P}^X(x) = - \int_M \frac{1}{2} \Delta_x p(x, \mu, t) d\mathbb{P}^X(x),$$

$$= - \int_M \frac{1}{2} \Delta_x \ln p(x, \mu, t) + \frac{1}{2} \left\| \nabla_x \ln p(x, \mu, t)_{x=X} \right\|^2 d\mathbb{P}^X(x).$$

Because of symmetry of the heat kernel, $\Delta_\mu \ln p(x, \mu, t) = \Delta_x \ln p(x, \mu, t)$. At a critical point for $t$, the left-hand side vanishes and the result follows. \qed

As a consequence of this result, the expected value of the Hessian trace of the negative log-likelihood is non-zero if $\left\| \nabla_x \ln p(x, \mu, t)_{x=X} \right\|^2$ is non-zero with positive probability. This happens unless $X$ is only supported on the critical points of the log-likelihood, as in the two pole distribution from Examples 6, 7, 9. This result does not imply that all eigenvalues of the expected Hessian are strictly positive but it means that some are and the rest could be 0 because negative values cannot happen at a local minimum. Positive eigenvalues of the expected Hessian correspond to $r = 2$ in the Taylor expansion in Theorem 4.5. The phenomenon where some, but not all eigenvalues are zero, i.e. smeariness occurs in those directions only, has been introduced by Eltzner (2022), Tran, Eltzner and Huckemann (2021) as directional smeariness. These consideration yield at once the following.

Corollary 4.9. Let $X_1, \ldots, X_n$ be an i.i.d. random sample on $M$. If for every $\mu$ and $t$ one has $\sum_{j=1}^n \left\| \nabla_x \ln p(x, \mu, t)_{x=X_j} \right\|^2 > 0$, then $\mu$ is at most directionally smearable.

We here briefly discuss the generality of Corollary 4.9 and possible ways to strengthen the result. On real analytic manifolds, the heat kernel is real analytic so that $-\ln p(x, \mu, t)$ can only have isolated critical points. This happens for the two-pole example as mentioned above, but it is “rare” in practice. E.g. if the data distribution is non-singular, it is almost sure not the case.

The Hessian trace being positive does not rule out directional smeariness. In order to strengthen this result, we need to consider the second order derivative terms in the Laplacian individually. One possibility is here to use particular a structure of $M$, e.g. if $M$ has a product Riemannian structure, the Hessian trace will generally be positive on each of these factors individually. Even more generally, with the anisotropic Brownian-like flows considered in Sommer and Swayne (2017), the diffusion variance can be fitted by maximum likelihood in all directions individually, corresponding to estimation of a full covariance matrix. This provides additional information on the directional derivative terms in the Laplacian individually, and we expect to be able to strengthen Corollary 4.9 in this setting in future work.

We perform a numerical study of joint parameter estimation for the two pole distribution discussed in Example 9 and the Brownian normal distribution. The results are displayed in Figure 5.

For the Brownian normal distribution we recover the true variance, as displayed in Figure 5b, which illustrates asymptotic consistency of the estimator. For the two-pole distribution, the variance $t$ estimated jointly with the mean $\mu$ is converges the variance $t = s$ leading to smeariness, as seen in Figure 5a. This remarkable result does not contradict Corollary 4.9, since the two pole distribution features a critical point for $\nabla_x \ln p(x, \mu, t)$ for $x$ being the south pole and thus violates the assumptions of the corollary. While the fact that the mean, if jointly estimated with the diffusion variance, is smearable in the two-pole model is undesirable, the non-uniqueness of the Fréchet mean in the same model is arguable even worse. Manually fixing a higher $t$ than estimated, leads to a more benign, non-smeary asymptotic behavior of $\mu$. This illustrates that diffusion means can have preferable asymptotic behavior to the Fréchet mean even in pathological cases.
Diffusion means in geometric spaces

Figure 5. The plots display estimates of the diffusion variance $\hat{t}$. The left plot show estimates on the two-pole distribution for various south pole weights $\alpha$ (yielding smeariness) and the right panel shows estimates from a Brownian normal distribution for various diffusion variances $t$.

5. Application

5.1. Reducing finite sample smeariness (FSS) in geomagnetic data

In contrast to smeariness, i.e. that the variance of estimators $\hat{\mu}_n$ based on a sample $X_1,\ldots,X_n$ of a population mean $\mu$ scales not asymptotically with inverse sample size $1/n$, there are distributions where this happens within finite sample size regime only. This phenomenon has been called finite sample smeariness (FSS) by Hundrieser, Eltzner and Huckemann (2020) and is measured, given an underlying metric $d(\cdot,\cdot)$, by the variance modulation that has been introduced by Pennec (2019)

$$m_n := \frac{n\mathbb{E}[d(\hat{\mu}_n,\mu)^2]}{\sum_{j=1}^{n} d(X_j,\mu)^2}.$$

While $m_n = 1$ is constant in $n$ in case of Euclidean spaces, in Figures 1b, 6 and 7 the estimators of the curves $n \mapsto m_n$ initially increase as in case of smeariness and settle (sometimes after a moderate decrease) at a constant larger than one. Presence of FSS, i.e. of $m_n > 1$ makes, among others, quantile based tests inapplicable, in contrast to bootstrap tests, having lowered power, though, see Hundrieser, Eltzner and Huckemann (2020). This effect is stronger the higher the initial overshoot and the higher the asymptotic horizontal.

Spherical statistics has a natural application to geolocation data, as they can be considered as observed random variables on the 2-sphere $S^2$. To show the occurrence of finite sample smeariness, as described by Hundrieser, Eltzner and Huckemann (2020), on real data, we here include an example of estimating both the Fréchet mean and diffusion $t$-mean for $t = 1$ and $t = 2$ on data sets of magnetic pole locations during geomagnetic pole reversals presented in McElhinny and Lock (1996) and downloadable from ftp://ftp.ngdc.noaa.gov/geomag/Paleomag/access/ver3.5, where each curve in the panels of Figure 6 represent a separate data set. These data sets were investigated by Eltzner (2022), who found finite sample smeariness of the Fréchet mean. The distribution of the magnetic pole locations during the reversals is similar to the bimodal Brownian normal distribution of Example 8, typically with lower variance. The limit of zero variance corresponds to the two pole distribution of Example 6.
Figure 6. Bootstrap variance modulation curves (19) of the estimated Fréchet mean and diffusion mean for \( t = 1 \) and \( t = 2 \) of eight magnetic pole data sets from McElhinny and Lock (1996), each represented by a different color. The dashed line indicates 2-smeariness while the solid black horizontal indicates 0-smeariness.

Figure 6 displays the \( k \)-out-of-\( n \) bootstrap variance modulations as introduced in Eltzner (2022), Hundrieser, Eltzner and Huckemann (2020)

\[
m_k := \frac{k \text{Var}[\mu_n^k]}{\text{Var}[\{X_j\}_{j=1}^n]} \tag{19}
\]

of the Fréchet mean and diffusion mean for \( t = 1 \) and \( t = 2 \) of eight different data sets. It is evident from the plots that the choice of \( t \) has an effect on FSS, which is consistent with the results from the previous examples: By increasing \( t \), asymptotic variance is reduced and initial, partly strong, overshoots are eventually removed, i.e. effects of FSS are reduced by increasing \( t \).

Lastly, we jointly estimate the diffusion mean and diffusion variance for the data sets of magnetic pole locations during geomagnetic pole reversals McElhinny and Lock (1996), which yields the results displayed in Figure 7. The results indicate a lower maximal variance modulation for the diffusion mean than for the Fréchet mean. Above bootstrap sample sizes of \( k = 100 \) all curves appear compatible with the standard CLT rate. The modulation curves for the diffusion means appear close to monotonically growing. In the notation introduced by Hundrieser, Eltzner and Huckemann (2020) this means that there is no Type II FSS. The estimated \( t \) parameters lie approximately in the range \([1.3, 3.1]\), with \( t \)-estimates being smaller for the data sets with faster convergence to the CLT rate.

In Table 1 we compare the variances of \( n \)-out-of-\( n \) bootstrap means for the Fréchet mean and the diffusion mean with jointly estimated diffusion variance. These are relevant for the power of hypothesis tests, smaller variance translating to higher power. The diffusion mean variance is clearly smaller for all data sets except data sets 062 and 063, where both variances are of similar size. This indicates that in case of a finite sample smeary Fréchet means, diffusion means achieve higher power in hypothesis tests.

In Figure 8 we illustrate the bootstrapped samples for data sets 044, for which the diffusion mean has much lower variance than the Fréchet mean, and 062, for which variances are similar. One can see that the distribution of diffusion means appear much more isotropic around the sample mean.

These results indicate that a joint estimation of \( \mu \) and \( t \) generally leads to a diffusion mean \( \mu_n \) which exhibits considerably lower magnitude of FSS than the Fréchet mean.
Figure 7. Bootstrap variance modulation curves (19) of the estimated Fréchet mean and diffusion mean for optimal $t$-values of eight magnetic pole data sets from McElhinny and Lock (1996), each represented by a different color. The dashed line indicates 2-smeariness while the solid black horizontal indicates Euclidean 0-smeariness.

Table 1. Variances of $n$-out-of-$n$ bootstrap means for the Fréchet mean (3rd column) and the diffusion means (4th column), with original sample size $n$ (2nd column) of the corresponding data sets (1st column), and the ratio diffusion mean variance over Fréchet mean variance (5th column) where the mean and diffusion variance $t$ (6th column) have been estimated jointly.

<table>
<thead>
<tr>
<th>Data set</th>
<th>sample size $n$</th>
<th>Fréchet mean variance</th>
<th>diffusion mean variance</th>
<th>ratio</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>006</td>
<td>107</td>
<td>1.118</td>
<td>0.323</td>
<td>0.289</td>
<td>1.809</td>
</tr>
<tr>
<td>012</td>
<td>40</td>
<td>0.403</td>
<td>0.281</td>
<td>0.697</td>
<td>1.601</td>
</tr>
<tr>
<td>044</td>
<td>145</td>
<td>0.394</td>
<td>0.119</td>
<td>0.302</td>
<td>1.322</td>
</tr>
<tr>
<td>062</td>
<td>154</td>
<td>0.482</td>
<td>0.487</td>
<td>1.010</td>
<td>2.303</td>
</tr>
<tr>
<td>063</td>
<td>84</td>
<td>1.127</td>
<td>1.049</td>
<td>0.931</td>
<td>2.429</td>
</tr>
<tr>
<td>107</td>
<td>50</td>
<td>1.023</td>
<td>0.473</td>
<td>0.462</td>
<td>1.754</td>
</tr>
<tr>
<td>122</td>
<td>239</td>
<td>1.932</td>
<td>1.173</td>
<td>0.607</td>
<td>3.031</td>
</tr>
<tr>
<td>141</td>
<td>60</td>
<td>0.824</td>
<td>0.406</td>
<td>0.493</td>
<td>1.685</td>
</tr>
</tbody>
</table>

5.2. Power of hypothesis tests

To see that the reduction in FSS is of practical importance, we compare the results of hypothesis tests for the pole position on the 8 geomagnetism data sets discussed above and on the wind direction data set discussed in Hundrieser, Eltzner and Huckemann (2020). We perform the two sample $T^2$ tests for heteroscedastic data using bootstrapped covariances and quantiles designed to achieve high power as discussed in detail in Eltzner and Huckemann (2017). We display the resulting sets of p-values in terms of pp-plots in Figure 9.

For the wind direction data, we considered data sets for the two cities separately and only took into account tests between data sets of which at least one displayed considerable FSS with a variance modulation of at least 5. This was the case for 3 out of 20 data sets for Basel, leading to 54 tests.
Figure 8. Bootstrap means (black dots) for data sets 044, where diffusion means have a much lower variance than Fréchet means, and 062, where variances are comparable. The red stars indicate the sample means. Panel (a) shows that the bootstrapped Fréchet means are near a small circle, which is consistent with the circular mode of the sample mean distribution in case of smeariness. In contrast, the bootstrapped diffusion means displayed in Panel (b) feature a fairly concentrated normal distribution consistent with a non-smeary mean. In Panel (c) the distribution of bootstrapped Fréchet means is again very anisotropic, featuring a tail of almost antipodal outliers, indicating smeariness. While the bootstrapped diffusion means displayed in Panel (d) have a similar high overall variance, the distribution is rather isotropic around the sample mean, indicating isotropic FSS.

considered and for 5 out of 20 data sets for Göttingen, leading to 85 tests considered, for a total number of 139 hypothesis tests.

It is immediately obvious from Figure 9 that the power of the test is much higher when using diffusion means with simultaneously estimated variance, highlighting the practical usefulness of diffusion means in a realistic scenario for data on positively curved spaces.

**Funding**

P. Hansen and S. Sommer gratefully acknowledge support by the Novo Nordisk Foundation grant NNF18OC0052000 and the Villum Foundation grants 22924 and 40582. B. Eltzner and S. Huckemann gratefully acknowledge funding by the DFG SFB 1456.
Supplementary Material


References


Diffusion means in geometric spaces


*Received April 2022 and revised November 2022*