Recursive Karcher Expectation Estimators And Geometric Law of Large Numbers

Jeffrey Ho **Guang Cheng** **Hesamoddin Salehian** **Baba C. Vemuri**
Department of CISE
University of Florida, Gainesville, FL 32611

Abstract

This paper studies a form of law of large numbers on $\mathbf{P}_n$, the space of $n \times n$ symmetric positive-definite matrices equipped with Fisher-Rao metric. Specifically, we propose a recursive algorithm for estimating the Karcher expectation of an arbitrary distribution defined on $\mathbf{P}_n$, and we show that the estimates computed by the recursive algorithm asymptotically converge in probability to the correct Karcher expectation. The steps in the recursive algorithm mainly consist of making appropriate moves on geodesics in $\mathbf{P}_n$, and the algorithm is simple to implement and it offers a tremendous gain in computation time of several orders in magnitude over existing non-recursive algorithms. We elucidate the connection between the more familiar law of large numbers for real-valued random variables and the asymptotic convergence of the proposed recursive algorithm, and our result provides an example of a new form of law of large numbers for random variables taking values in a Riemannian manifold. From the practical side, the computation of the mean of a collection of symmetric positive-definite (SPD) matrices is a fundamental ingredient in many algorithms in machine learning, computer vision and medical imaging applications. We report an experiment using the proposed recursive algorithm for K-means clustering, demonstrating the algorithm's efficiency, accuracy and stability.

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

1 Introduction

Estimating the expectation (mean) of a distribution is undoubtedly the most important and fundamental step in any statistical analysis. Given a sequence $\mathbf{X}_1, \mathbf{X}_2, ...\mathbf{X}_k$ of i.i.d. samples from a probability measure dP on $\mathbb{R}^l$, a simple estimator $\mathbf{m}_k$ for the mean $\mathbf{m}$ of dP is given by

$$\mathbf{m}_k = \frac{\mathbf{X}_1 + ... + \mathbf{X}_k}{k}. \tag{1}$$

The validity of the estimator of course is guaranteed by the (weak) law of large numbers, which states that the estimator $\mathbf{m}_k$ converges to the true mean $\mathbf{m}$ in probability. This well-known result is perhaps first taught at a college-level probability class, and for practitioners in statistics, AI and machine learning, it is so deeply grounded that many times we are using it without our immediate awareness of it.

The problem of interest in this paper is concerned with random variables taking values in a space that does not have an additive (vector space) structure, and in particular, general distributions defined on (Riemannian) manifolds. This absence of additive structure has an important consequence since it means the formula in Equation 1 that computes the estimator is no longer valid and in fact, it is not clear how the mean should be defined at all. Specifically, let (Ω, ω) denote a probability space with probability measure ω. A real-valued random variable $\mathbf{X}$ (more generally, a vector-valued random variable) is a measurable function defined on Ω taking values in $\mathbb{R}$ ($\mathbb{R}^n$, $n > 1$):

$$\mathbf{X} : \Omega \to \mathbb{R}\,(\mathbb{R}^n).$$

The distribution of the random variable $\mathbf{X}$ is the push-forward probability measure $dP_{\mathbf{X}} = \mathbf{X}^*(\omega)$ on $\mathbb{R}$, and its expectation $\mathbb{E}X$ is defined by the integral

$$\mathbb{E}\mathbf{X} = \int_\Omega \mathbf{X}\, d\omega = \int_{\mathbb{R}} x\, dP_{\mathbf{X}}(x). \tag{2}$$

One important reason that the above integral can be defined is that $\mathbf{X}$ takes value in a vector space, a space

that allows additions of two points. However, if the target of the random variable $\mathbf{X}$ does not have an apparent additive structure, the definition of the expectation through the above integral is no longer viable, and in particular, the absence of additive structure must be substituted by a useful structure of the target space in order to properly and appropriately define the expectation.

A manifold-valued ($\mathcal{M}$-valued) random variable $\mathbf{X}$ is a random variable that takes values in a manifold $\mathcal{M}$. Familiar examples of important manifolds in statistics include spheres, Stiefel and Grassman manifolds used in directional statistics [2, 6], and the space $\mathbf{P}_n$ of $n \times n$ symmetric positive-definite matrices that parameterizes the space of all non-degenerated Gaussian distributions in $\mathbb{R}^n$. If the manifold $\mathcal{M}$ is assumed to be Riemannian, then it is possible to compensate for the lack of additive structure with its intrinsic geometric structure. Let $\mathbf{d}_{\mathcal{M}}(\mathbf{x}, \mathbf{y})$ denote the Riemannian distance between two points $\mathbf{x}, \mathbf{y} \in \mathcal{M}$. The integral in Equation 2 can be generalized to manifold-valued random variable $\mathbf{X}$ by defining its **Karcher expectation** as [4]

$$\mathbb{KE}\,\mathbf{X} \;=\; \min_{\mu^* \in \mathcal{M}} \int_{\Omega} \mathbf{d}^2_{\mathcal{M}}(\mu^*, \mathbf{X})\, d\omega. \tag{3}$$

Note that, by definition, Karcher expectation of $\mathbf{X}$ is indeed a point on $\mathcal{M}$, and for an arbitrary manifold $\mathcal{M}$, the minimum of the above integral is generally not unique, in contrast to the uniqueness of mean in the Euclidean case. Nevertheless, [4] provides several characterizations of the manifolds $\mathcal{M}$ for which the Karcher expectation is unique. In particular, for a complete Riemannian manifold with negative sectional curvature, the Karcher expectation is unique, and this result applies to the manifold $\mathbf{P}_n$ of $n \times n$ symmetric positive-definite matrices equipped with the Fisher-Rao metric, which is different from the Euclidean (Frobenius) metric.

While Karcher's result guarantees that for any distribution defined on $\mathbf{P}_n$, its Karcher expectation with respect to the Fisher-Rao metric is unique, it is not clear how the cherished mean estimator given in Equation 1 should be generalized to estimate the Karcher expectation from a collection of i.i.d. samples. A direct approach would be to interpret $\mathbf{m}_n$ in Equation 1 as the finite mean and accordingly define the n^{th} Karcher estimator $\mathbf{m}_n$ as

$$\mathbf{m}_k = \min_{\mu^* \in \mathbf{P}_k} \sum_{i=1}^{k} \mathbf{d}^2(\mu^*, \mathbf{X}_i). \tag{4}$$

However, this approach is undesirable because the computation of $\mathbf{m}_k$ requires an optimization, and for large number of samples, it is usually unappetizing. Instead, we will generalize a slightly different but entirely equivalent form of Equation 1,

$$\mathbf{m}_k = \frac{(k-1)\,\mathbf{m}_{k-1} + \mathbf{X}_k}{k}. \tag{5}$$

This incremental form is advantageous because it involves only two points and in an Euclidean space $\mathbb{R}^n$, we can interpret it geometrically as moving an appropriate distance away from $\mathbf{m}_{k-1}$ towards $\mathbf{X}_k$ on the straight line joining $\mathbf{X}_k$ and $\mathbf{m}_{k-1}$. Based on this idea, we propose a recursive algorithm for computing Karcher estimator that is entirely based on traversing geodesics in $\mathbf{P}_n$ and without requiring any optimization such as in Equation 4. *The main theorem proved in this paper shows that in the limit, as the number of samples goes to infinity, the recursive Karcher estimator converges to the (unique) Karcher expectation in probability.* It is this result that provides us with a *new form of law of large numbers*, and the main difference between the Euclidean and non-Euclidean settings is that the additions are now replaced by the geometric operations of moves on the geodesics. Another distinguishing feature is that the finite estimators computed by Equations (4) and (5) are the same in Euclidean space but generally quite different in non-Euclidean spaces, and in particular, the result also depends on the ordering of the sequence $\mathbf{X}_1, \mathbf{X}_2, ...\mathbf{X}_k$. Asymptotically, it is immaterial because the limit will always converge to the unique Karcher expectation. Furthemreo, our recursive algorithm based on Equation (5) leads to a gain in computation time of several orders in magnitude over existing non-recursive algorithms based on Equation 4.

The rest of the paper is organized as follows. In section 2, we will present a few important and relevant background material on the Riemannian manifold $\mathbf{P}_n$ with the Fisher-Rao metric that are needed in the statement and proof of the main theorem. The recursive estimator for Karcher expectation on $\mathbf{P}_n$ and the main theorem is presented in section 3, with the proof (which is surprisingly simple) given in section 4. The last section provides several experimental results that include an application to K-means clustering on $\mathbf{P}_n$.

2 Preliminaries

In this section, we gather together relevant and well-known differential geometric properties of $\mathbf{P}_n$ and we use [3] as our main reference (and also [8]). Let $\mathbf{P}_n$ denote the space of $n \times n$ symmetric positive-definite (SPD) matrices, and as a manifold, the tangent space $\mathbf{T}_{\mathbf{M}}$ at a point $\mathbf{M} \in \mathbf{P}_n$ can be identified with the space of $n \times n$ symmetric matrices. Furthermore, $\mathbf{P}_n$ admits

an action of the general linear group $\mathbf{GL}(n)$ ($n \times n$ non-singular matrices) according to the formula:

$$\forall g \in \mathbf{GL}(n), \forall \mathbf{M} \in \mathbf{P}_n,\ g^*\mathbf{M} = g\mathbf{M}g^\top.$$

This action is transitive, and there is a $\mathbf{GL}(n)$-invariant (affine-invariant) Riemannian metric defined by

$$< \mathbf{U}, \mathbf{V} >_{\mathbf{M}} = \mathbf{Tr}(\mathbf{M}^{-1}\mathbf{U}\mathbf{M}^{-1}\mathbf{V}), \tag{6}$$

where $\mathbf{Tr}$ denotes the trace and $< \mathbf{U}, \mathbf{V} >_{\mathbf{T_M}}$ denotes the inner-product of two tangent vectors $\mathbf{U}, \mathbf{V} \in \mathbf{T_M}$. The invariance of the Riemannian metric is easily checked using the action defined in the equation above: for any $g \in \mathbf{GL}(n)$,

$$< g^*\mathbf{U}, g^*\mathbf{V} >_{g*\mathbf{M}} = < \mathbf{U}, \mathbf{V} >_{\mathbf{M}}.$$

This particular Riemannian metric on $\mathbf{P}_n$ is special in two ways. First, this metric is the Fisher-Rao metric when $\mathbf{P}_n$ is considered as the parameter domain for the set of all normal distributions with a fixed mean [1] and its importance and relevance in statistics follow from this connection. Second, this metric has been studied quite extensively in differential geometry since the mid-twentieth century [**?**], and its various important properties have long been established, including that it is a Riemannian symmetric space and is geodesically complete with constant negative sectional curvature. In particular, there are closed-form formulas for computing the geodesics and Riemannian distances: for any two points $\mathbf{M}, \mathbf{N} \in \mathbf{P}_n$, their Riemannian distance is given by the formula

$$\mathbf{d}^2(\mathbf{M}, \mathbf{N}) = \mathbf{Tr}((\mathbf{Log}(\mathbf{M}^{-1}\mathbf{N}))^2),$$

where $\mathbf{Log}$ is the matrix logarithm, and the unique geodesic $\gamma(s)$ joining $\mathbf{M}, \mathbf{N}$ is given by the formula

$$\gamma(s) = \mathbf{M}^{\frac{1}{2}}(\mathbf{M}^{-\frac{1}{2}}\mathbf{N}\mathbf{M}^{-\frac{1}{2}})^s\mathbf{M}^{\frac{1}{2}}. \tag{7}$$

We note that in the above formula, $\gamma(0) = \mathbf{M}, \gamma(1) = \mathbf{N}$, and $\mathbf{M}^{-\frac{1}{2}}\mathbf{N}\mathbf{M}^{-\frac{1}{2}}$ is indeed an SPD matrix. If $\mathbf{M}^{-\frac{1}{2}}\mathbf{N}\mathbf{M}^{-\frac{1}{2}} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ denote its eigen-decomposition with $\mathbf{D}$ diagonal, then the fractional exponent $(\mathbf{M}^{-\frac{1}{2}}\mathbf{N}\mathbf{M}^{-\frac{1}{2}})^s$ can be computed as $\mathbf{U}\mathbf{D}^s\mathbf{U}$, where $\mathbf{D}^s$ exponentiates the diagonal entries of $\mathbf{D}$. An example of Equation 7 that will become useful later is the geodesic $\gamma(s)$ between the identity matrix $\mathbf{I}$ and a diagonal matrix $\mathbf{D}$. Let

$$\mathbf{D} = \begin{pmatrix} D_1 & & \\ & \ddots & \\ & & D_n \end{pmatrix} = \begin{pmatrix} e^{d_1} & & \\ & \ddots & \\ & & e^{d_n} \end{pmatrix},$$

where $d_i = \log D_i$ for $i = 1, ..., n$. The path $\gamma(s)$ defined by

$$\gamma(s) = \begin{pmatrix} e^{sd_1} & & \\ & \ddots & \\ & & e^{sd_n} \end{pmatrix} \tag{8}$$

is the unique geodesic joining the identity $\mathbf{I}$ and $\mathbf{D}$ such that $\gamma(0) = \mathbf{I}, \gamma(1) = \mathbf{D}$.

For a probability measure $dP(\mathbf{x})$ on $\mathbf{P}_n$, the Riemannian distance $\mathbf{d}$ can be used to define its Karcher expectation (which is unique according to [4]) and its variance:

$$\mathbf{M}_P = \arg\min_{\mu \in \mathbf{P}_n} \int_{\mathbf{P}_n} \mathbf{d}^2(\mu, \mathbf{x})\, dP(\mathbf{x}),$$
$$\mathbf{Var}(P) = \int_{\mathbf{P}_n} \mathbf{d}^2(\mathbf{M}_P, \mathbf{x})\, dP(\mathbf{x}).$$

3 Algorithm and Theorem

The law of large numbers for real-valued random variables states that for a sequence of identically and independently distributed random variables $\mathbf{X}_1, \mathbf{X}_2, ...$, the sequence $\mathbf{S}_n = (\mathbf{X}_1 + \mathbf{X}_2 + ... + \mathbf{X}_k)/k$ converges to $\mathbb{E}\mathbf{X}_1$ in probability as $k \to \infty$

$$\mathbf{S}_k = (\mathbf{X}_1 + \mathbf{X}_2 + ... + \mathbf{X}_k)/k \to \mathbb{E}\mathbf{X}_1.$$

Or equivalently, the sequence

$$\mathbf{S}_{k+1} = \frac{k\mathbf{S}_k + \mathbf{X}_{k+1}}{k+1} = \frac{k}{k+1}\mathbf{S}_k + \frac{1}{k+1}\mathbf{X}_{k+1}$$

converges to $\mathbb{E}\mathbf{X}_1$ in probability as $k \to \infty$. The above equation can be interpreted geometrically as the update formula for $\mathbf{S}_{k+1}$ moving the current estimator $\mathbf{S}_k$ towards the new sample $\mathbf{X}_{k+1}$ along the line joining $\mathbf{S}_k$ and $\mathbf{X}_{k+1}$ with the size of the move given by the weights $\frac{1}{k+1}, \frac{k}{k+1}$. Since we are seeking a geometric generalization of the law of large numbers, this particular geometric interpretation using straight lines (geodesics) can be applied to any space that has geodesics connecting any pair of points. For the space $\mathbf{P}_n$, this immediately leads to the following recursive algorithm:

Recursive Karcher Expectation Estimator Let $\mathbf{X}_1, \mathbf{X}_2, ... \in \mathbf{P}_n$ be i.i.d. samples from a probability distribution $P(\mathbf{x})$ on $\mathbf{P}_n$. The recursive Karcher expectation estimator $\mathbf{M}_k$ is defined as

$$\mathbf{M}_1 = \mathbf{X}_1, \tag{9}$$
$$\mathbf{M}_{k+1} = \mathbf{M}_k^{\frac{1}{2}}(\mathbf{M}_k^{-\frac{1}{2}}\mathbf{X}_{k+1}\mathbf{M}_k^{-\frac{1}{2}})^{w_{k+1}}\mathbf{M}_k^{\frac{1}{2}}, \tag{10}$$

where $w_{k+1} = \frac{1}{k+1}$. Note that, formally, $\mathbf{M}_k$ corresponds to $\mathbf{S}_k$ above and the update formula for $\mathbf{M}_{k+1}$ in Equation 10 simply moves $\mathbf{M}_k$ along the geodesic (See Equation 7) joining $\mathbf{M}_k$ and $\mathbf{X}_{k+1}$ with the size of the move given by w_{k+1}. This is in complete agreement with the geometric interpretation of the Euclidean update formula discussed above. Perhaps more surprisingly, this simple recursive algorithm can be shown to

converge to the correct Karcher expectation according to the following main theorem of the paper, whose proof will be given in the next section:

Theorem 1 (Geometric Law of Large Numbers)**.** *For any arbitrary distribution defined on* $\mathbf{P}_n$ *and a sequence of i.i.d samples* $\mathbf{X}_1, \mathbf{X}_2, ...$*, the recursive Karcher expectation estimator converges to the Karcher expectation in probability as* $k \to \infty$*.*

4 Proof of the Main Theorem

Let P denote the given distribution on $\mathbf{P}_n$ with Karcher expectation $\mathbf{M} \in \mathbf{P}_n$ and $\mathbf{X}_1, \mathbf{X}_2, ...$ a sequence of i.i.d. samples. Let $\mathbf{M}_k$ denote the k^{th} Karcher estimator according to Equation 10 and denote by P_k its distribution. The main point of the proof is to show that the integral

$$\int_{\mathbf{P}_n} \mathbf{d}^2(\mathbf{x}, \mathbf{M}) dP_k(\mathbf{x}) \leq \frac{1}{k} \mathbf{Var}(P) \tag{11}$$

converges to zero as $k \to \infty$. Notice that the above integral computes the variance of the k^{th} Karcher estimator $\mathbf{M}_k$ with respect to the true Karcher expectation $\mathbf{M}$, and the bound shows that this variance converges to zero as $k \to \infty$, which implies the convergence of the recursive Karcher estimator $\mathbf{M}_k$ to $\mathbf{M}$.

It turns out that the proof relies heavily on the following inequality:

Theorem 2 (Master Inequality)**.** *Let* $\mathbf{R}, \mathbf{S}, \mathbf{T}$ *be three arbitrary points in* $\mathbf{P}_n$*, and let* $\gamma(s)$ *denote the geodesic joining* $\mathbf{R}$ *and* $\mathbf{S}$ *such that* $\gamma(0) = \mathbf{R}, \gamma(1) = \mathbf{S}$*. Then, for all* $s \in [0,\ 1]$*,*

$$\begin{aligned}\mathbf{d}^2(\gamma(s),\, \mathbf{T}) \leq\ & (1-s)\mathbf{d}^2(\mathbf{R}, \mathbf{T}) + s\, \mathbf{d}^2(\mathbf{S}, \mathbf{T}) \\ & - s(1-s)\mathbf{d}^2(\mathbf{R}, \mathbf{S}).\end{aligned}$$

We remark that the above inequality relating the distances between three points $\mathbf{R}, \mathbf{S}, \gamma(t)$ on a geodesic and an arbitrary point $\mathbf{T}$ is a general property of any Riemannian manifold with negative sectional curvature (see [5]). Here, we give a self-contained proof for $\mathbf{P}_n$ that uses only its algebraic property and the **GL**-invariance of the metric. In fact, we show a stronger result for $\mathbf{P}_n$ that

$$\begin{aligned}\mathbf{d}^2(\gamma(s),\, \mathbf{T}) =\ & (1-s)\mathbf{d}^2(\mathbf{R}, \mathbf{T}) + s\, \mathbf{d}^2(\mathbf{S}, \mathbf{T}) \\ & - s(1-s)\mathbf{d}^2(\mathbf{R}, \mathbf{S}).\end{aligned}$$

Proof. The theorem will follow if we can show that the function $\Gamma(s)$

$$\begin{aligned}\Gamma(s) =\ & (1-s)\mathbf{d}^2(\mathbf{R}, \mathbf{T}) + s\, \mathbf{d}^2(\mathbf{S}, \mathbf{T}) \\ & - s(1-s)\mathbf{d}^2(\mathbf{R}, \mathbf{S}) - \mathbf{d}^2(\gamma(s)\, \mathbf{T}),\end{aligned}$$

has zero second derivative, $\Gamma''(s) = 0$, on an open interval containing the closed unit interval $[0,\ 1]$. Since $\Gamma(0) = \Gamma(1) = 0$, the result follows.

Since the inequality concerns only three points, we can use the **GL**-invariance of the distance to simplify the calculation by transforming the three points $\mathbf{R}, \mathbf{S}, \mathbf{T}$ to $\mathbf{I}, \mathbf{D}, \mathbf{Q}$, where $\mathbf{I}$ is the identity, $\mathbf{D}$ a diagonal matrix and $\mathbf{Q}$ is some matrix in $\mathbf{P}_n$. This can be easily accomplished by first applying $\mathbf{R}^{-\frac{1}{2}}$ to $\mathbf{R}, \mathbf{S}, \mathbf{T}$ to get $\mathbf{I}, \mathbf{R}^{-\frac{1}{2}}\mathbf{S}\mathbf{R}^{-\frac{1}{2}}, \mathbf{R}^{-\frac{1}{2}}\mathbf{T}\mathbf{R}^{-\frac{1}{2}}$. Let $\mathbf{U}\mathbf{D}\mathbf{U}^\top = \mathbf{R}^{-\frac{1}{2}}\mathbf{S}\mathbf{R}^{-\frac{1}{2}}$ denote the eigen-decomposition, and applying $\mathbf{U}^\top$ to the transformed triple to obtain $\mathbf{I}, \mathbf{D}, \mathbf{U}^\top\mathbf{R}^{-\frac{1}{2}}\mathbf{T}\mathbf{R}^{-\frac{1}{2}}\mathbf{U} = \mathbf{Q}$. Since the distance is **GL**-invariant, it suffices to prove the result for the normalized triple $\mathbf{I}, \mathbf{D}, \mathbf{Q}$ with $\Gamma(s)$

$$\begin{aligned}\Gamma(s) =\ & (1-s)\mathbf{d}^2(\mathbf{I}, \mathbf{Q}) + s\, \mathbf{d}^2(\mathbf{D}, \mathbf{Q}) \\ & - s(1-s)\mathbf{d}^2(\mathbf{I}, \mathbf{D}) - \mathbf{d}^2(\gamma(s),\, \mathbf{Q}).\end{aligned}$$

By direct calculation, we have

$$\Gamma''(s) = 2\, \mathbf{d}^2(\mathbf{I}, \mathbf{D}) - \frac{d^2}{ds^2}\mathbf{d}^2(\gamma(s),\, \mathbf{Q}).$$

The geodesic $\gamma(s)$ joining $\mathbf{I}$ and $\mathbf{D}$ is given in Equation 8 and we will denote $\gamma(s) = \mathbf{D}_s$. Since $\mathbf{d}^2(\mathbf{D}_s, \mathbf{Q}) = \mathbf{Tr}(\mathbf{Log}^2(\mathbf{D}_s^{-1}\mathbf{Q}))$, we let $\mathbf{W}_s = \mathbf{Log}(\mathbf{D}_s^{-1}\mathbf{Q}))$.

$$\frac{d}{ds}\mathbf{Tr}(\mathbf{W}_s^2) = 2\mathbf{Tr}(\mathbf{W}_s \frac{d\mathbf{W}_s}{ds}),$$

and

$$\frac{d^2}{ds^2}\mathbf{Tr}(\mathbf{W}_s^2) = 2\mathbf{Tr}((\frac{d\mathbf{W}_s}{ds})^2 + \mathbf{W}_s \frac{d^2\mathbf{W}_s}{ds^2}).$$

We have

$$\frac{d\mathbf{W}_s}{ds} = (\mathbf{D}_s^{-1}\mathbf{Q}))^{-1} \frac{d\mathbf{D}_s^{-1}}{ds}\mathbf{Q} = \mathbf{Q}^{-1}\mathbf{D}_s \frac{d\mathbf{D}_s^{-1}}{ds}\mathbf{Q},$$

and

$$\frac{d^2\mathbf{W}_s}{ds^2} = \mathbf{Q}^{-1}\frac{d\mathbf{D}_s}{ds}\frac{d\mathbf{D}_s^{-1}}{ds}\mathbf{Q} + \mathbf{Q}^{-1}\mathbf{D}_s \frac{d^2\mathbf{D}_s^{-1}}{ds^2}\mathbf{Q}.$$

Using Equation 8, we further have

$$\frac{d\mathbf{D}_s}{ds} = \begin{pmatrix} d_1 e^{sd_1} & & \\ & \ddots & \\ & & d_n e^{sd_n} \end{pmatrix}, \tag{12}$$

$$\frac{d\mathbf{D}_s^{-1}}{ds} = \begin{pmatrix} -d_1 e^{-sd_1} & & \\ & \ddots & \\ & & -d_n e^{-sd_n} \end{pmatrix}, \tag{13}$$

$$\frac{d^2\mathbf{D}_s^{-1}}{ds^2} = \begin{pmatrix} d_1^2 e^{-sd_1} & & \\ & \ddots & \\ & & d_n^2 e^{-sd_n} \end{pmatrix}. \tag{14}$$

Using the above three equations and Equation 8, we have

$$\frac{d\mathbf{D}_s}{ds}\frac{d\mathbf{D}_s^{-1}}{ds} + \mathbf{D}_s\frac{d^2\mathbf{D}_s^{-1}}{ds^2} = 0,$$

and hence $d^2\mathbf{W}_s/ds^2 = 0$. Furthermore, since

$$\mathbf{D}_s\frac{d\mathbf{D}_s^{-1}}{ds^2} = \begin{pmatrix} -d_1 & & \\ & \ddots & \\ & & -d_n \end{pmatrix},$$

we have

$$\mathbf{Tr}((\frac{d\mathbf{W}_s}{ds})^2) = d_1^2 + ... + d_n^2 = \mathbf{d}^2(\mathbf{I}, \mathbf{D}).$$

This shows that $\Gamma''(s) = 0$ for all $s \in [0, 1]$. □

The master inequality in turn implies the following inequality:

Proposition 1. *Let P be an arbitrary distribution defined on $\mathbf{P}_n$, and $\mathbf{M}$ denote its Karcher expectation. Then for any $\mathbf{y} \in \mathbf{P}_n$,*

$$\int_{\mathbf{P}_n} [\,\mathbf{d}^2(\mathbf{x}, \mathbf{y}) - \mathbf{d}^2(\mathbf{x}, \mathbf{M})]\, dP(\mathbf{x}) \geq \mathbf{d}^2(\mathbf{y}, \mathbf{M}). \quad (15)$$

Note that we already know that

$$\int_{\mathbf{P}_n} \mathbf{d}^2(\mathbf{x}, \mathbf{y}) dP(\mathbf{x}) \geq \int_{\mathbf{P}_n} \mathbf{d}^2(\mathbf{x}, \mathbf{M}) dP(\mathbf{x}),$$

and the difference between the two sides are in fact lower-bounded by the squared distance between $\mathbf{y}$ and $\mathbf{M}$.

Proof. Let $\gamma(s)$ denote the geodesic joining $\mathbf{y}$ and $\mathbf{M}$ such that $\gamma(1) = \mathbf{y}$ and $\gamma(0) = \mathbf{M}$. Consider the function $\lambda(s)$,

$$\lambda(s) = \int_{\mathbf{P}_n} \mathbf{d}^2(\mathbf{x}, \gamma(s)) dP(\mathbf{x}) - \int_{\mathbf{P}_n} \mathbf{d}^2(\mathbf{x}, \mathbf{M}) dP(\mathbf{x}).$$

Note that $\lambda(0) = 0$ and to complete the proof, we need to show that $\lambda(1) \geq \mathbf{d}^2(\mathbf{y}, \mathbf{M})$. For each $0 \leq s \leq 1$, applying the master inequality with $\mathbf{x} = \mathbf{T}, \mathbf{R} = \mathbf{M}, \mathbf{S} = \mathbf{y}$ and integrating, we have

$$0 \leq \lambda(s) \leq s\lambda(1) - s(1-s)\mathbf{d}^2(\mathbf{y}, \mathbf{M}).$$

This is true for all $0 \leq s \leq 1$ and in particular,

$$\lambda(1) \geq (1-s)\mathbf{d}^2(\mathbf{y}, \mathbf{M}),$$

for all $s > 0$. Letting $s \to 0$, we have the desired inequality

$$\lambda(1) \geq \mathbf{d}^2(\mathbf{y}, \mathbf{M}).$$

□

With the above two inequalities in hand, the proof of the main theorem is straightforward:

Proof. We will prove the inequality in Equation 11 by induction as it holds trivially for $k = 1$. By Equation 10, each $\mathbf{M}_{k+1} = \gamma(\frac{1}{k+1})$ where $\gamma(t)$ is the geodesic joining $\mathbf{M}_k, \mathbf{X}_{k+1}$ such that $\gamma(0) = \mathbf{M}_k, \gamma(1) = \mathbf{X}_{k+1}$. Using the master inequality, we have

$$\mathbf{d}^2(\mathbf{M}_{k+1}, \mathbf{M}) \leq \frac{k}{k+1}\mathbf{d}^2(\mathbf{M}_k, \mathbf{M}) + \frac{1}{k+1}\mathbf{d}^2(\mathbf{X}_{k+1}, \mathbf{M}) - \frac{k}{(k+1)^2}\mathbf{d}^2(\mathbf{M}_k, \mathbf{X}_{k+1}).$$

Taking the expectations, we have

$$\mathbf{Var}(\mathbf{M}_{k+1}) \leq \frac{k}{k+1}\mathbf{Var}(\mathbf{M}_k) + \frac{1}{k+1}\mathbf{Var}(P) - \frac{k}{(k+1)^2}\mathbb{E}_{\mathbf{M}_k}\mathbb{E}_{\mathbf{X}_{k+1}}\mathbf{d}^2(\mathbf{M}_k, \mathbf{X}_{k+1}),$$

where $\mathbf{Var}(\mathbf{M}_{k+1}), \mathbf{Var}(\mathbf{M}_k)$ are variances of $\mathbf{M}_{k+1}, \mathbf{M}_k$ with respect to $\mathbf{M}$, respectively. Since

$$\mathbb{E}_{\mathbf{X}_{k+1}}\mathbf{d}^2(\mathbf{M}_k, \mathbf{X}_{k+1}) = \int_{\mathbf{P}_n} \mathbf{d}^2(\mathbf{M}_k, \mathbf{x})\, dP(\mathbf{x})$$

and using Equation 15, we obtain

$$\mathbb{E}_{\mathbf{X}_{k+1}}\mathbf{d}^2(\mathbf{M}_k, \mathbf{X}_{k+1}) \geq \int_{\mathbf{P}_n} \mathbf{d}^2(\mathbf{X}_{k+1}, \mathbf{M})\, dP(\mathbf{x}) + \mathbf{d}^2(\mathbf{M}_k, \mathbf{M}) = \mathbf{Var}(P) + \mathbf{d}^2(\mathbf{M}_k, \mathbf{M}).$$

This implies

$$\mathbb{E}_{\mathbf{M}_k}\mathbb{E}_{\mathbf{X}_{k+1}}\mathbf{d}^2(\mathbf{M}_k, \mathbf{X}_{k+1}) \geq \mathbf{Var}(P) + \mathbf{Var}(\mathbf{M}_k).$$

Putting these together, we have

$$\mathbf{Var}(\mathbf{M}_{k+1}) \leq \frac{k}{k+1}\mathbf{Var}(\mathbf{M}_k) + \frac{1}{k+1}\mathbf{Var}(P) - \frac{k}{(k+1)^2}(\mathbf{Var}(\mathbf{M}_k) + \mathbf{Var}(P)),$$

with the right-hand-side equals

$$\frac{k^2}{(k+1)^2}\mathbf{Var}(\mathbf{M}_k) + \frac{1}{(k+1)^2}\mathbf{Var}(P).$$

By induction hypothesis, it is in turn upper-bounded by

$$\frac{k^2}{(k+1)^2}\frac{1}{k}\mathbf{Var}(P) + \frac{1}{(k+1)^2}\mathbf{Var}(P) = \frac{1}{k+1}\mathbf{Var}(P).$$

□

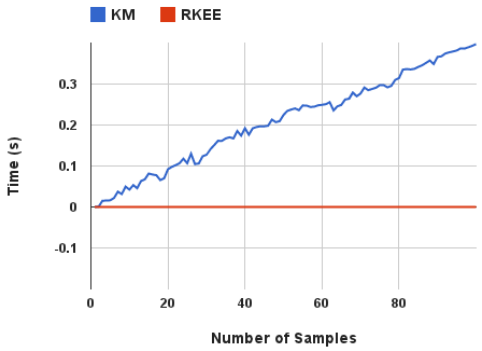

Figure 1: Running time comparison of the recursive (red) versus non-recursive (blue) Karcher expectation estimators for data on $\mathbf{P}_5$.

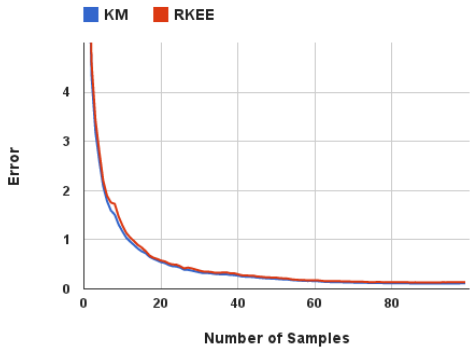

Figure 2: Error comparison of the recursive (red) versus non-recursive (blue) Karcher expectation estimators for data on $\mathbf{P}_5$.

5 Experiments

In this section, we demonstrate the accuracy and efficiency of our proposed recursive algorithm through a set of experiments.

5.1 Performance of the Recursive Karcher Expectation Estimator for Symmetric Distributions

We illustrate the performance of our recursive estimator on a set of random samples on $\mathbf{P}_n$ drawn from a symmetric distribution, and compare the accuracy and computational efficiency of the recursive Karcher expectation estimator, RKEE, and the non-recursive Karcher mean, KM, for the given dataset. To this end, a set of 100 i.i.d samples from a log-Normal distribution [7] on $\mathbf{P}_5$ is generated, and the Karcher expectation is computed using RKEE incrementally as well as the non-recursive method. We set the expectation and the variance of log-Normal distribution to the identity matrix and one, respectively. The error in estimation is measured by the geodesic distance from each estimated point to the identity. Further, for each new sample, the computation time for each method is recorded. Figure 1 illustrates the significant difference in running time between RKEE and KM. It can be seen that the time taken by our method is considerably shorter than the non-recursive method, and is almost constant as a function of the number of samples.

The errors of the two estimators are shown in Figure 2. It can be seen that the recursive estimator provides roughly the same accuracy as its non-recursive counterpart. Furthermore, for large numbers of samples, the recursive estimation error converges to zero. Therefore, the recursive algorithm performs more accurately as the number of data becomes larger.

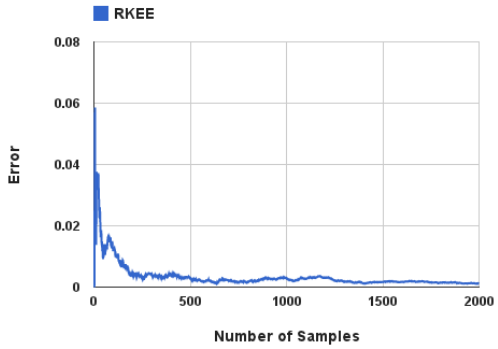

Figure 3: Convergence of the recursive Karcher expectation estimator for an asymmetric distribution on $\mathbf{P}_3$.

5.2 Performance of the Recursive Karcher Expectation Estimator for Asymmetric Distributions

For asymmetric distributions, we use a mixture of two log-Normal distributions and repeat the same experiment as above. The first distribution in the mixture is centered at the identity matrix with the variance 0.25, and the second component is centered at a randomly-chosen matrix with variance 0.12. A set of 800 samples are drawn from this distribution for the experiment. To measure the error, we compute the gradient vector of the objective function in Equation 4 ($k = 800$) and it norm. Figure 3 illustrates the error of RKEE for this asymmetric distribution. It can be seen that as the number of samples increases, the error converges to zero. This demonstrates our algorithm's accuracy and stability.

5.3 Application to K-means Clustering

In this section we evaluate the performance of our proposed recursive algorithm for K-means clustering. K-means clustering is of fundamental importance for many applications in computer vision and machine learning. Due to the lack of a closed-form formula for computing the Karcher mean, mean computation is the most time consuming step in applying K-means to SPD matrices, since at the end of each iteration the mean for each estimated cluster needs to be recomputed. The experimental results in this section demonstrate that, for SPD matrices, our recursive estimator can significantly speed up the clustering process without any observable degradation in its accuracy when compared with the non-recursive method.

RKEE is applied to K-means clustering for SPD matrices as follows. At the end of each iteration, we only consider matrices whose cluster assignments have changed For each of these "moving" samples, the source cluster center is updated by subtracting the sample, and the destination cluster center is updated by adding the new matrix. Both of these updates can be efficiently performed using our recursive formula given in Equation 10, with appropriate weights. A set of experiments are performed using different scenarios to illustrate the effectiveness of our method. In each experiment, a set of random samples from mixtures of log-Normal distributions on $\mathbf{P}_n$ are generated and used as inputs to the K-means algorithm. In the first experiment, we increase the number of samples and compare the accuracy and running time of recursive and non-recursive estimates for each case. In the second experiment, we increase the number of clusters, i.e. number of log-Normal components, to observe how each estimator behaves. In the last experiment, we evaluate the performance of each algorithm with respect to matrix dimension. To measure the clustering error, the geodesic distance between each estimated cluster center and its true value is computed and they are summed over all cluster centers.

Figures 4 and 5, respectively, compare the running time and the clustering error of each method with increasing number of samples. It is evident that the recursive formula outperforms the non-recursive method, while the errors for both methods are very similar. Moreover, as the number of samples increases, recursive algorithm improves in accuracy. In Figures 6 and 7, with increasing number of clusters, the proposed method is still far more computationally efficient than the non-recursive version. Also Figure 8 illustrates a significant difference in running time between these two methods, while Figure 9 shows that the errors for both methods are roughly the same. These experiments verify that the proposed recursive method

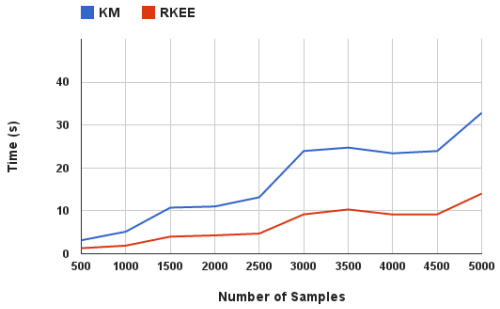

Figure 4: Comparison of running time for K-means clustering based on recursive and non-recursive Karcher expectation estimators, for varying number of samples from three clusters on $\mathbf{P}_3$

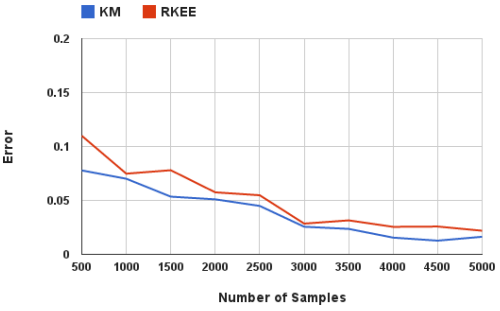

Figure 5: Comparison of accuracy for K-means clustering based on recursive and non-recursive Karcher expectation estimators, for varying number of samples from three clusters on $\mathbf{P}_3$

is far more computationally efficient for K-means clustering with SPD matrices.

6 Conclusions

In this paper, we have presented a novel recursive Karcher expectation estimator for symmetric positive-definite (SPD) matrix-variate random variables. The validity of the recursive estimator is provided by the paper's main theorem that guarantees, asymptotically, the estimators converge to the Karcher expectation. Because the update formula for the estimator uses the geodesics, we interpret this result as a *geometric form of law of large numbers*. The novel recursive Karcher expecation estimator is used for computing the cluster centers in a K-means clustering algorithm applied to SPD manifold-valued data, and experimental results clearly demonstrate the significant improvement in computational efficiency over the non-recursive counter part.

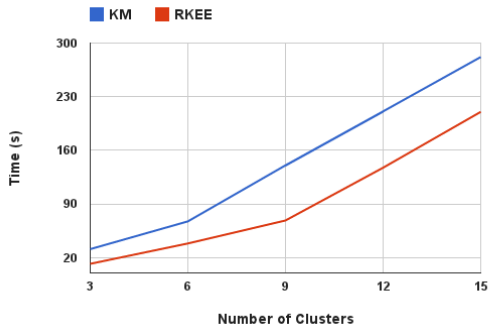

Figure 6: Comparison of running time for K-means clustering based on recursive and non-recursive Karcher expectation estimators, for 5000 SPD matrices from varying number of clusters on $\mathbf{P}_3$

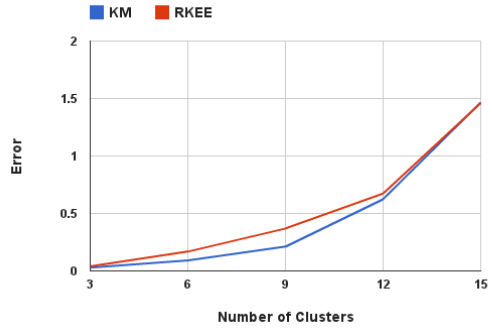

Figure 7: Comparison of accuracy for K-means clustering based on recursive and non-recursive Karcher expectation estimators, for 5000 SPD matrices from varying number of clusters on $\mathbf{P}_3$.

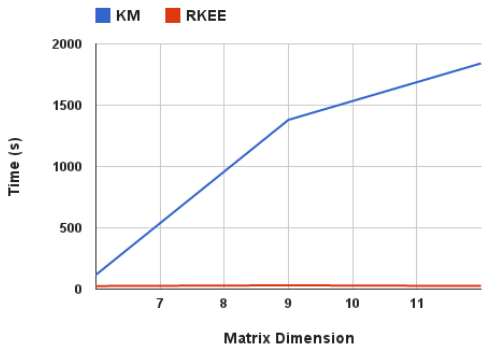

Figure 8: Comparison of running time for K-means clustering based on recursive and non-recursive Karcher expectation estimators, for 2000 samples from three clusters with varying sizes

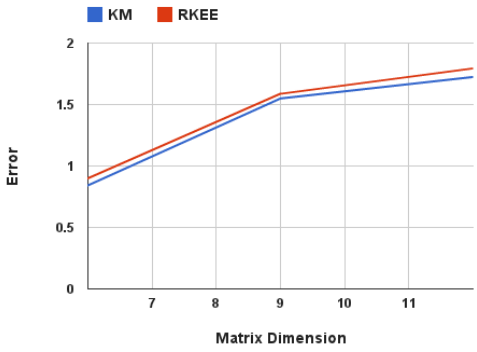

Figure 9: Comparison of accuracy for K-means clustering based on recursive and non-recursive Karcher expectation estimators, for 2000 samples from three clusters with varying sizes.

Acknowledgments

This research was in part funded by the NIH grant NS066340 to BCV.

References

[1] S. Amari and H. Nagaoka. *Methods for Information Geometry*. American Mathematical Society, 2007.

[2] Y. Chikuse. *Statistics on Special Manifolds*. Springer, 2003.

[3] S. Helgason. *Differantial Geometry, Lie Groups, and Symmetric Spaces*. American Mathematical Society, 2001.

[4] H. Karcher. Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math*, 30:509–541, 1977.

[5] J. Lurie. Lecture notes on the theory of hadamard spaces (metric spaces of nonpositive curvature). http://www.math.harvard.edu/ lurie/ papers/hadamard.pdf.

[6] K. Mardia and P. Jupp. *Directional Statistics*. Wiley, 1999.

[7] A. Schwartzman. *Random ellipsoids and false discovery rates: Statistics for diffusion tensor imaging data*. PhD thesis, Stanford University, 2006.

[8] A. Terras. *Harmonic Analysis on Symmetric Spaces and Applications*. Springer-Verlag, 1985.