Statistics on Manifolds

P. Thomas Fletcher[*]
* University of Virginia, Departments of Electrical & Computer Engineering and Computer Science, Charlottesville, VA USA
Corresponding: ptf8v@virginia.edu

Abstract

Statistical analysis of data on a Riemannian manifold extends fundamental concepts from multivariate statistical analysis in vector spaces by using the metric structure. The first example of generalizing a classical statistic to the manifold setting is the Fréchet mean, which minimizes the sum-of-squared geodesic distances to the data. Such a geometric least-squares principle also leads to extensions of principal components analysis and regression on manifolds. From these geometric concepts, we extend to a probabilistic viewpoint by defining normal distributions on Riemannian manifolds. This leads to probabilistic modeling and inference through both maximum likelihood and Bayesian approaches.

1. Introduction

This chapter provides a review of basic statistics for data on Riemannian manifolds, including generalizations of the concepts of a mean, principal component analysis (PCA), and regression. Definitions for these statistics in Euclidean space all somehow rely on the vector space operations in $\mathbb{R}^d$. For example, the arithmetic mean is defined using vector addition and scalar multiplication. The inherent difficulty in defining statistics on general Riemannian manifolds is the lack of vector space operations in these spaces.

One avenue for analyzing manifold-valued data is through geometry. Because a Riemannian manifold has a distance metric, we can think of model fitting as a least-squares problem, that is, minimizing the sum-of-squared distances from our data to the model. While least-squares problems in Euclidean space often have closed-form solutions, e.g., linear regression, solving least-squares problems in Riemannian manifolds typically requires some form of iterative optimization.

For PCA and regression analysis, a further complication arises in that the underlying models in $\mathbb{R}^d$ are defined as linear subspaces, which are also not available in Riemannian manifolds. In these cases, geodesic curves provide the natural generalization of straight lines to manifolds. Therefore, the natural generalization of linear regression to manifolds is *geodesic regression*, in which a geodesic curve is fit to data with an associated real-valued explanatory variable. In the case of PCA, the first principal component may now be replaced with a *principal geodesic* that best fits the data using just one dimension. Higher-order principal components are defined as *principal*

geodesic subspaces, which are generated as the image under the exponential map of linear subspaces of a tangent space.

In addition to the geometric perspective, another avenue to define manifold statistics is through probability. In traditional Euclidean statistics, least-squares fitting is equivalent to maximum likelihood estimation under a Gaussian distribution assumption of the errors. Such a probabilistic interpretation is also possible on manifolds through the definition of a Riemannian normal distribution law. We show how this distribution provides a unifying framework for probabilistic interpretation of several models of manifold data, including the Fréchet mean, geodesic regression, and principal geodesic analysis.

Throughout this chapter, let $y_1, \ldots, y_N \in M$ denote a set of data on a Riemannian manifold. From a statistical viewpoint, we will consider these data as coming from a realization of a random sample, i.e., draws from a set of N independent, identically distributed (i.i.d.) random variables. However, we will often consider data points and their statistical analysis from a purely geometric perspective, without referring to random variables.

2. The Fréchet Mean

For Euclidean data, the sample mean is the de facto point estimate of the center of a data set. It is the simplest statistic to define, yet also perhaps the most fundamentally important one. The sample mean of a set of points $y_1, \ldots, y_N \in \mathbb{R}^d$ is given by their arithmetic average,

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i. \tag{2.1}$$

This definition for the mean depends on the vector space operations of Euclidean space. In general, a Riemannian manifold will not be a vector space, and this definition for the mean will not be directly applicable. For data on a manifold embedded in Euclidean space, $\mathcal{M} \subset \mathbb{R}^d$, we could consider applying the linear mean equation using the vector operations of the ambient space, $\mathbb{R}^d$. However, the resulting mean point may not land on $\mathcal{M}$. The following two examples demonstrate how the arithmetic mean of data on an embedded manifold can fail to be on the manifold.

Example 1 (Linear Mean for the Sphere)**.** The 2D sphere has a natural embedding in $\mathbb{R}^3$ as the set of all unit-length vectors, i.e., $S^2 \equiv \{y \in \mathbb{R}^3 : \|y\| = 1\}$. Given a set of points on the sphere, $y_1, \ldots, y_N \in S^2$, their linear average in $\mathbb{R}^3$, $\bar{y}$, will not in general be a point on S^2. Take, for example, the points $y_1 = (1, 0, 0)$ and $y_2 = (0, 1, 0)$. Their mean, $\bar{y} = (0.5, 0.5, 0)$, has norm $\|\bar{y}\| = \sqrt{2}/2$, and thus does not lie on S^2.

Example 2 (Linear Mean for GL(k))**.** The space of $k \times k$ matrices with non-zero determinant form a Lie group known as the general linear group, denoted GL(k). This is a connected, open subset of $\mathrm{M}_{(k,k)}$. However, averaging under the usual vector space operations of $\mathrm{M}_{(k,k)} \equiv \mathbb{R}^{k\times k}$ does not preserve the non-degeneracy of GL(k). Take, for example, the two matrices:

$$y_1 = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \qquad y_2 = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}.$$

Both y_1 and y_2 have determinant equal to one, and are thus in GL(2), but their linear average, $\bar{y} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, has determinant zero.

Given that the formula (2.1) is not defined for general Riemannian manifolds, we may then ask if there are defining properties of the mean point in Euclidean space that can be generalized to the manifold setting. The equation for the Euclidean mean can be derived from several different principles:

1. **Algebraic:** The arithmetic mean is the unique point such that the *residuals sum to zero*:

   $$(y_1 - \bar{y}) + \cdots + (y_N - \bar{y}) = 0.$$

   Note that this definition uses only the vector space properties of $\mathbb{R}^d$.
2. **Geometric:** It is a *least-squares* centroid of the data points. That is, it minimizes the sum-of-squared distances to the data,

   $$\bar{y} = \arg\min_{y\in\mathbb{R}^d} \sum_{i=1}^{N} \|y_i - y\|^2.$$

3. **Probabilistic:** If the y_i are realizations of i.i.d. multivariate normal random variables, $Y_i \sim N(\mu, \Sigma)$, then $\bar{y}$ is a *maximum-likelihood estimate* of the mean parameter μ. That is,

   $$\bar{y} = \arg\max_{\mu\in\mathbb{R}^d} \prod_{i=1}^{N} p(y_i; \mu, \Sigma),$$

   where $p(\cdot; \mu, \Sigma)$ is the pdf for the multivariate normal distribution, $N(\mu, \Sigma)$.

The algebraic characterization of the mean point does not generalize to Riemannian manifolds, again because it is dependent on a vector space structure. However, the geometric and probabilistic characterizations can be generalized. In this section we consider the geometric characterization of the mean point on a Riemannian manifold. This concept of a mean point is due to Maurice Fréchet [Fré48], and is thus known as the *Fréchet mean*. Later, in Section 5 we will see how this is related to a probabilistic

interpretation. Now consider a set of data, $y_1, \ldots, y_N$, on a Riemannian manifold, M. The geometric characterization of the Euclidean sample mean can be generalized to Riemannian manifolds as the *sample Fréchet mean*, which is the minimizer of the sum-of-squared distances to the data,

$$\bar{y} = \arg\min_{y \in M} \sum_{i=1}^{N} d(y, y_i)^2, \tag{2.2}$$

where d denotes the geodesic distance on M. Fréchet actually introduced a much more general concept of expectation of a probability measure on a metric space, of which, the sample Fréchet mean on a manifold is a special case.

2.1. Existence and Uniqueness of the Fréchet Mean

Because the Fréchet mean is defined via an optimization, the first natural questions are whether a solution to this optimization exists and if it is unique. We begin by giving an example where the Fréchet mean does not exist.

Example 3. The "punctured plane" is $\mathbb{R}^2$ with the origin $(0, 0)$ removed. As an open set of $\mathbb{R}^2$, this is a manifold, and it is a Riemannian manifold when given the same Euclidean metric as $\mathbb{R}^2$. However, it is not a complete manifold, as geodesics (which are still straight lines) cannot pass through the missing point at the origin. Any set of points where the Fréchet mean in $\mathbb{R}^2$ would be $(0, 0)$ do not have a Fréchet mean in the punctured plane, for example, $y_1 = (1, 0), y_2 = (-1, 0)$.

It turns out that the key ingredient missing for the punctured plane is completeness of the metric. In fact, completeness of a distance metric is sufficient to guarantee existence of the Fréchet mean, as shown in the next theorem. Note that this holds for *any* complete metric space, not only those that are Riemannian manifolds.

Theorem 1 (Existence of the Fréchet mean)**.** *Let M be a complete metric space. Then the Fréchet mean of any finite set of points $y_1, \ldots, y_N \in M$ exists.*

Proof. Define the sum-of-squared distance function, $F(y) = \sum_{j=1}^{N} d(y_i, y_j)^2$. We show that a global minimum of F exists (but it may not be unique). Denote the diameter of the point set by $r = \max_{i,j} d(y_i, y_j)$. Let $K = \cup_{i=1}^{N} \bar{B}_r(y_i)$, where $\bar{B}_r(y_i)$ is the closed metric ball of radius r centered at x_i. By the completeness of X, K is a closed set, bounded in diameter by $2r$, and so is a compact set. Therefore, the restriction of the sum-of-squared distance function to the set K attains a minimum within K. Now consider a point $y \in X$ such that $y \notin K$. Then $F(y) > nr^2$. However, this must be larger than the minimum within K because $F(y_i) = \sum_{j=1}^{N} d(y_i, y_j)^2 \leq nr^2$. □

Even when a Fréchet mean of a set of data exists, it may not be unique. That is, there may be multiple points that achieve the minimum in (2.2). A simple example of this is given on the 2D sphere.

Example 4 (Non-uniqueness of the Fréchet mean on S^2)**.** Consider the unit sphere S^2 embedded in $\mathbb{R}^3$, with two data points at the north and south pole: $y_1 = (0, 0, 1), y_2 = (0, 0, -1)$. Then the Fréchet mean is the set of points on the equator $\bar{y} = \{(\cos\theta, \sin\theta, 0) : \theta \in [0, 2\pi)\}$.

Conditions for the uniqueness of the Fréchet mean were given by Karcher [Kar77] and later refined by Kendall [Ken90]. The following result is due to Asfari [Asf11].

Theorem 2 (Uniqueness of the Fréchet mean)**.** *Let M be a complete Riemannian manifold with sectional curvature bounded above by Δ, and let* inj(M) *denote the injectivity radius of M. If data $y_1, \ldots, y_N \in M$ are contained in a geodesic ball of radius*

$$r = \begin{cases} \frac{1}{2}\min\left\{\text{inj}(M), \frac{\pi}{\sqrt{\Delta}}\right\}, & \text{if } \Delta > 0, \\ \frac{1}{2}\,\text{inj}(M), & \text{if } \Delta \leq 0, \end{cases}$$

then the Fréchet mean $\bar{y}$ is unique.

Example 5 (2D constant curvature manifolds)**.** To better understand this uniqueness theorem, we consider the examples of constant curvature manifolds of dimension two.

- ($\Delta = 0$) For the Euclidean plane, $\mathbb{R}^2$, the injectivity radius is infinite and sectional curvature is equal to 0. Therefore, the theorem states that any set of data in $\mathbb{R}^2$ has a unique Fréchet mean.
- ($\Delta = 1$) For the 2-sphere, S^2, the injectivity radius is π and sectional curvature is equal to 1. The theorem then gives a bound of $r = \frac{pi}{2}$, meaning any set of data contained in an open hemisphere of S^2 will have a unique Fr'echet mean.
- ($\Delta = -1$) For the hyperbolic plane, H^2, the injectivity radius is infinite and sectional curvature is -1. Then, the theorem states that, like Euclidean space, any set of data in H^2 will have a unique Fr'echet mean.

Further theoretical results of the sample Fréchet mean were developed by Bhattacharya and Patrangenaru [BP03, BP05]. They established asymptotic consistency of the sample Fréchet mean and proved a central limit theorem.

2.2. Estimation of the Fréchet Mean

The Fréchet mean is defined by the minimization problem (2.2). The squared-distance function from a point $y \in M$ on a Riemannian manifold is smooth away from the cut

locus of y. As such, a natural strategy for computing the Fréchet mean is by gradient descent optimization, first proposed by Pennec [Pen99]. This gradient descent algorithm also appeared in [BF01] for the case of spheres and [Moa02] for the case of rotations.

First, consider the squared-distance function from a single point $x \in M$,

$$F_x(y) = d(y, x)^2.$$

As a consequence of the Gauss lemma (see [dC92]), the gradient of the squared-distance function is given by

$$\text{grad}F_x(y) = -2\text{Log}_y(x).$$

Then the gradient descent, with some step size $\tau > 0$, proceeds as follows:

Algorithm 1: Fréchet Mean

Input: $y_1, \ldots, y_N \in M$
Output: $\bar{y} \in M$, the Fréchet mean
Initialize: $\bar{y}_0 = x_1$
while $\|v\| > \epsilon$ **do**
  $v = \frac{\tau}{N} \sum_{i=1}^{N} \text{Log}_{\bar{y}_j} y_i$
  $\bar{y}_{j+1} = \text{Exp}_{\bar{y}_j} v$

3. Covariance and Principal Geodesic Analysis

The covariance of a vector-valued random variable y in $\mathbb{R}^d$ is defined as

$$\text{Cov}(y) = E\left[(y - E[y])\,(y - E[y])^T\right].$$

This definition clearly relies on the vector space structure of $\mathbb{R}^d$, i.e., vector transpose and matrix multiplication operations. Therefore, it does not apply directly as written to a manifold-valued random variable. However, we can rewrite this equation by recalling that the Riemannian log map in Eucliean space is given by $\text{Log}_y x = (x - y)$. Then, the covariance of y is equivalently

$$\text{Cov}(y) = E\left[\left(\text{Log}_{E[y]} y\right)\left(\text{Log}_{E[y]} y\right)^T\right].$$

This equation can now be directly generalized to a Riemannian manifold by replacing the Euclidean expectation, $E[y]$, with Fréchet expectation. For a random sample,

$y_1, y_2, \ldots, y_n \in M$, the sample covariance matrix is given by

$$S = \frac{1}{n} \sum_{i=1}^{n} \left(\mathrm{Log}_{\bar{y}} y_i\right) \left(\mathrm{Log}_{\bar{y}} y_i\right)^T . \tag{2.3}$$

3.1. Principal Component Analysis

The covariance matrix encodes the variability of multivariate data, however, it is often difficult to interpret or make use of it directly. A more convenient breakdown of the variability of high-dimensional data is given by principal component analysis (PCA), a method whose origins go back to Pearson [Pea01] and Hotelling [Hot33]. See the book [Jol86a] for a comprehensive review of PCA. The objectives of principal component analysis are (1) to efficiently parameterize the variability of data and (2) to decrease the dimensionality of the data parameters. In this section we review PCA for Euclidean data, $y_1, \ldots, y_N \in \mathbb{R}^d$, with mean $\bar{y}$, before describing how it can be generalized to manifolds in the next section.

There are several different ways to describe PCA. The definitions given here may not necessarily be standard, but they are helpful as the basis for the generalization to Riemannian manifolds. The goal of PCA is to find a sequence of nested linear subspaces, $V_1, \ldots, V_d$, through the mean that best approximate the data. This may be formulated in two ways, both resulting in the same answer. The first is a least-squares approach, where the objective is to find the linear subspaces such that the sum-of-squares of the residuals to the data are minimized. More precisely, the linear subspace V_k is defined by a basis of orthonormal vectors, i.e., $V_k = \mathrm{span}(\{v_1, \ldots, v_k\})$, which are given by

$$v_k = \arg\min_{\|v\|=1} \sum_{i=1}^{N} \|y_i^k - \langle y_i^k, v\rangle\, v\|^2, \tag{2.4}$$

where the y_i^k are defined recursively by

$$y_i^1 = y_i - \bar{y},$$
$$y_i^k = y_i^{k-1} - \langle y_i^{k-1}, v_{k-1}\rangle\, v_{k-1}$$

Simply put, the point y_i^k is obtained by removing from $(y_i - \bar{y})$ the contributions of the previous directions, $v_1, \ldots, v_{k-1}$. In other words, the point y_i^k is the projection of $(y_i - \mu)$ onto the subspace perpendicular to V_{k-1}.

The other way of defining principal component analysis is as the subspaces through the mean that maximize the total variance of the projected data. The total variance for

a set of points $y_1, \ldots, y_N$ is defined as

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \|y_i - \bar{y}\|^2.$$

Then the linear subspaces $V_k = \text{span}(\{v_1, \ldots, v_k\})$ are given by the vectors

$$v_k = \arg\max_{\|v\|=1} \sum_{i=1}^{N} \langle y_i^k, v \rangle^2, \tag{2.5}$$

where the y_i^k are defined as above. It can be shown (see [Jol86a]) that both definitions of PCA, i.e., (2.4) and (2.5), give the same results thanks to the Pythagorean theorem.

The computation of the spanning vectors v_k proceeds as follows. First, the linear average of the data is computed as

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

Next, the sample covariance matrix of the data is computed as

$$S = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})(y_i - \bar{y})^T.$$

This is the unbiased estimate of the covariance matrix, that is, $N - 1$ is used in the denominator instead of N. The covariance matrix is a symmetric, positive-semidefinite quadratic form, that is, $S = S^T$, and for any $x \in \mathbb{R}^d$ the inequality $x^T S\, x \geq 0$ holds. Therefore, the eigenvalues of S are all real and nonnegative. Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of S ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, and let $v_1, \ldots, v_d$ be the correspondingly ordered eigenvectors. When repeated eigenvalues occur, there is an ambiguity in the corresponding eigenvectors, i.e., there is a hyperplane from which to choose the corresponding eigenvectors. This does not present a problem as any orthonormal set of eigenvectors may be chosen. These directions are the solutions to the defining PCA equations, (2.4) and (2.5), and are called the *principal directions* or *modes of variation*.

Any data point y_i can be decomposed as

$$y_i = \bar{y} + \sum_{k=1}^{d} \alpha_{ik} v_k,$$

for real coefficients $\alpha_{ik} = \langle y_i - \bar{y}, v_k \rangle$. The α_{ik} for fixed i are called the *principal components* of y_i. The total variation of the data is given by the sum of the eigenvalues, $\sigma^2 = \sum_{k=1}^{d} \lambda_k$. The dimensionality of the data can be reduced by discarding the principal directions that contribute little to the variation, that is, choosing an $m < d$ and

projecting the data onto V_m, giving the approximation

$$\tilde{y}_i = \bar{y} + \sum_{k=1}^{m} \alpha_{ik} v_k.$$

One method for choosing the cut-off value m is based on the percentage of total variation that should be preserved.

3.2. Principal Geodesic Analysis

Principal geodesic analysis (PGA) [FLJ03, FLPJ04] generalizes PCA to handle data $y_1, \ldots, y_N$ on a connected, complete manifold M. The goal of PGA, analogous to PCA, is to find a sequence of nested geodesic submanifolds that maximize the projected variance of the data. These submanifolds are called the *principal geodesic submanifolds*.

Let $T_{\bar{y}}M$ denote the tangent space of M at the Fréchet mean $\bar{y}$ of the y_i. Let $U \subset T_{\bar{y}}M$ be a neighborhood of 0 such that projection is well-defined for all geodesic submanifolds of $\mathrm{Exp}_{\bar{y}}(U)$. We assume that the data is localized enough to lie within such a neighborhood. The principal geodesic submanifolds are defined by first constructing an orthonormal basis of tangent vectors $v_1, \ldots, v_n \in T_{\bar{y}}M$ that span the tangent space $T_{\bar{y}}M$. These vectors are then used to form a sequence of nested subspaces $V_k = \mathrm{span}(\{v_1, \ldots, v_k\}) \cap U$. The principal geodesic submanifolds are the images of the V_k under the exponential map: $H_k = \mathrm{Exp}_{\bar{y}}(V_k)$. The first principal direction is chosen to maximize the projected variance along the corresponding geodesic:

$$v_1 = \arg\max_{\|v\|=1} \sum_{i=1}^{N} \|\mathrm{Log}_{\bar{y}}(\pi_H(y_i))\|^2, \tag{2.6}$$
$$\text{where} \quad H = \mathrm{Exp}_{\bar{y}}(\mathrm{span}(\{v\}) \cap U).$$

The remaining principal directions are defined recursively as

$$v_k = \arg\max_{\|v\|=1} \sum_{i=1}^{N} \|\mathrm{Log}_{\bar{y}}(\pi_H(y_i))\|^2, \tag{2.7}$$
$$\text{where} \quad H = \mathrm{Exp}_{\bar{y}}(\mathrm{span}(\{v_1, \ldots, v_{k-1}, v\}) \cap U).$$

Just as is the case with PCA, we can alternatively define PGA through a least squares fit to the data. In this setting, the first principal direction is chosen to minimize the sum-of-squared geodesic distance from the data to the corresponding geodesic:

$$v_1 = \arg\min_{\|v\|=1} \sum_{i=1}^{N} \|\mathrm{Log}_{y_i}(\pi_H(y_i))\|^2,$$
$$\text{where} \quad H = \mathrm{Exp}_{\bar{y}}(\mathrm{span}(\{v\}) \cap U).$$

The remaining principal directions are defined recursively as

$$v_k = \arg\min_{\|v\|=1} \sum_{i=1}^{N} \|\mathrm{Log}_{y_i}(\pi_H(y_i))\|^2,$$
$$\text{where} \quad H = \mathrm{Exp}_{\bar{y}}(\mathrm{span}(\{v_1, \ldots, v_{k-1}, v\}) \cap U).$$

3.3. Estimation: Tangent Approximation and Exact PGA

Neither the variance maximization (2.6), (2.7) nor the residual minimization (3.2), (3.2) formulation of PGA has a closed form solution for general manifolds. Therefore, Fletcher et al. [FLPJ04] proposes to approximate PGA in the tangent space to the Fréchet mean of the data. This is done by first mapping the x_i to the tangent space $T_\mu M$ using the Log map. Linear distances in $T_\mu M$ between points close to the origin are similar to the geodesic distances between the corresponding points in M under the Exp map. Therefore, if the data are highly concentrated about the Fréchet mean, the PGA optimization problem is well-approximated by the PCA optimization problem of the Log map transformed points. Fletcher et al. [FLPJ04] give an explicit expression in the case of the sphere for the approximation error between projections in the tangent space versus on the manifold. This suggests the following tangent space approximation algorithm to PGA.

Algorithm 2: Tangent approximation to PGA

Input: Data $y_1, \ldots, y_N \in M$
Output: Principal directions, $v_k \in T_\mu M$, variances, $\lambda_k \in \mathbb{R}$
$\bar{y}$ = Fréchet mean of $\{y_i\}$ (Algorithm 1)
$u_i = \mathrm{Log}_{\bar{y}}(y_i)$
$\mathbf{S} = \frac{1}{N-1} \sum_{i=1}^{N} u_i u_i^T$
$\{v_k, \lambda_k\}$ = eigenvectors/eigenvalues of $\mathbf{S}$.

Later work developed algorithms for exact optimization of the variance maximization formulas (2.6), (2.7) for PGA. This was first worked out for the special case of SO(3) by Said et al. [SCBS07] and then for general Riemannian manifolds by Sommer et al. [SLHN10, SLN14]. This algorithm, often referred to as exact PGA, proceeds by gradient ascent. This requires derivatives of the Riemannian Exp and Log maps, which are given by Jacobi fields. These derivatives are also used in geodesic regression and will be covered in the next section. Chakraborty et al. [CSV16] developed an efficient algorithm for exact PGA on constant curvature manifolds, using closed-form solutions for distances and projections onto geodesic submanifolds. Salehian et al. [SVV14] present an incremental algorithm for computing PGA by updating the parameters with

eac newly introduced data point. This has two advantages: (1) it reduces the memory cost over the standard batch mode PGA algorithms, and (2) it allows new data to be easily added later, without recomputing the entire PGA.

3.4. Further Extensions of PCA to Manifolds

Geodesic PCA [HHM10] solves a similar problem to the sum-of-squared residual minimization formulation of PGA (3.2), (3.2), with the exception that the geodesic principal components are not constrained to pass through the Fréchet mean. In the case of data in Euclidean space, the hyperplanes that best fit the data always pass through the mean. However, in the case of data on a manifold with nontrivial curvature, removing the constraint that geodesics pass through the mean can lead to more flexibility in fitting data. For data on a sphere, S^d, principal nested spheres (PNS) [JDM12] finds a series of nested subspheres of decreasing dimension that best fit the data. In contrast to PGA, the principal spheres are not constrained to be geodesic spheres (i.e., they can have smaller radius than the original sphere). Also, instead of building up from low dimension to high, PNS iteratively finds nesed spheres starting from the full dimension d and removing one dimension at a time. One consequence of this is that the 0-dimensional principal nested sphere is not necessarily the Fréchet mean. Eltzner et al. [EJH14] extend PNS to polyspheres (products of multiple spheres) by developing a procedure for deforming a polysphere into a single sphere where PNS can then be applied. Banerjee et al. [BJV17] present a version of PGA that is robust to outliers, along with an exact algorithm to compute it.

4. Regression Models

Regression analysis is a fundamental statistical tool for determining how a measured variable is related to one or more potential explanatory variables. The most widely used regression model is linear regression, due to its simplicity, ease of interpretation, and ability to model many phenomena. However, if the response variable takes values on a nonlinear manifold, a linear model is not applicable. Several works have studied regression models on manifolds, where the goal is to fit a curve on a manifold that models the relationship between a scalar parameter and data on the manifold. This is typically done by a least squares fit, similar to the Fréchet mean definition in (2.2), except now the optimization is over a certain class of curves on the manifold rather than a point. That is, given manifold data $y_1, \ldots, y_N \in M$ with corresponding real data $x_1, \ldots, x_N \in \mathbb{R}$, the regression problem is to find a curve $\hat{\gamma}(x) \in M$ such that

$$\hat{\gamma} = \arg\min_{\gamma \in \Gamma} \sum_{i=1}^{N} d(\gamma(x_i), y_i)^2, \tag{2.8}$$

where Γ is a space of curves on M.

In this chapter we will focus on nonparametric kernel regression on Riemannian manifolds [DFBJ07] and geodesic regression [Fle11, Fle12], i.e., where Γ is the space of parameterized geodesics on M. Niethammer et al. [NHV11] independently proposed geodesic regression for the case of diffeomorphic transformations of image time series. Hinkle et al. [JH14] use constant higher-order covariant derivatives to define intrinsic polynomial curves on a Riemannian manifold for regression. Shi et al. [SSL+09] proposed a semiparametric model for manifold response data, which also has the ability to handle multiple covariates.

A closely related problem to the regression problem is that of fitting smoothing splines to manifold data. The typical objective function for smoothing splines is a combination of a data matching term and a regularization term for the spline curve. For example, Su et al. [SDK+12] proposed a smoothing spline where the data matching is the same least squares objective as the regression problem (2.8), leading to a smoothing splines optimization of the form

$$\hat{\gamma} = \arg\min_{\gamma \in \Gamma} \sum_{i=1}^{N} d(\gamma(x_i), y_i)^2 + \lambda \mathcal{R}(\gamma), \tag{2.9}$$

where $\mathcal{R}$ is some regularization functional, and $\lambda > 0$ is a weighting between regularization and data fitting. In this case, the search space may be the space of all continuous curve segments, $\Gamma = C([0, 1], M)$. Jupp and Kent [JK87] proposed solving the smoothing spline problem on a sphere by unrolling onto the tangent space. This unrolling method was later extended to shape spaces by Kume [KDL07]. Smoothing splines on the group of diffeomorphisms has been proposed as growth models by Miller et al. [Mil04] and as second-order splines by Trouvé et al. [TV10]. A similar paradigm is used by Durrleman et al. [DPT+09] to construct spatiotemporal image atlases from longitudinal data. Yet another related problem is the spline *interpolation* problem, where the data matching term is dropped and the regularization term is optimized subject to constraints that the curve pass through specific points. The pioneering work of Noakes et al. [NHP89] introduced the concept of a cubic spline on a Riemannian manifold for interpolation. Crouch and Leite [CL95] investigated further variational problems for these cubic splines and for specific classes of manifolds, such as Lie groups and symmetric spaces. Buss and Fillmore [BF01] defined interpolating splines on the sphere via weighted Fréchet averaging.

4.1. Regression in Euclidean Space

4.1.1. Multilinear Regression

Before formulating geodesic regression on general manifolds, we begin by reviewing multiple linear regression in $\mathbb{R}^d$. Here we are interested in the relationship between a non-random *independent* variable $X \in \mathbb{R}$ and a random *dependent* variable Y taking

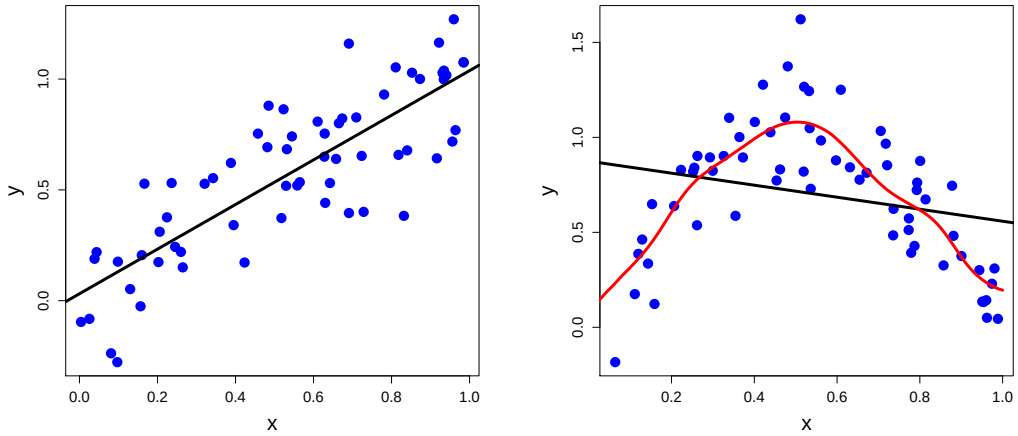


Figure 2.1 Comparison of linear (black) and nonparametric (red) regressions. When the data follows a linear trend (left), a linear regression model is favored due to its ease of interpretation. However, when the data trend is nonlinear (right), nonparametric regression models will fit better.

values in $\mathbb{R}^d$. A multiple linear model of this relationship is given by

$$Y = \alpha + X\beta + \epsilon, \tag{2.10}$$

where $\alpha \in \mathbb{R}^d$ is an unobservable *intercept* parameter, $\beta \in \mathbb{R}^d$ is an unobservable *slope* parameter, and ϵ is an $\mathbb{R}^d$-valued, unobservable random variable representing the error. Geometrically, this is the equation of a one-dimensional line through $\mathbb{R}^d$ (plus noise), parameterized by the scalar variable X. For the purposes of generalizing to the manifold case, it is useful to think of α as the starting point of the line and β as a velocity vector.

Given realizations of the above model, i.e., data $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}^d$, for $i = 1, \ldots, N$, the least squares estimates, $\hat{\alpha}, \hat{\beta}$, for the intercept and slope are computed by solving the minimization problem

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{(\alpha,\beta)} \sum_{i=1}^{N} \|y_i - \alpha - x_i\beta\|^2 . \tag{2.11}$$

This equation can be solved analytically, yielding

$$\hat{\beta} = \frac{\frac{1}{N}\sum x_i y_i - \bar{x}\bar{y}}{\sum x_i^2 - \bar{x}^2},$$
$$\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta},$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the x_i and y_i, respectively. If the errors in the model are drawn from distributions with zero mean and finite variance, then these estimators are unbiased and consistent. Furthermore, if the errors are homoscedastic (equal variance) and uncorrelated, then the Gauss-Markov theorem states that they will have minimal mean-squared error amongst all unbiased linear estimators.

4.1.2. Univariate Kernel Regression

Before reviewing the manifold version, we give a quick overview of univariate kernel regression as developed by Nadaraya [Nad64] and Watson [Wat64]. As in the linear regression setting, we are interested in finding a relationship between data $x_1, \ldots, x_N \in \mathbb{R}$, coming from an independent variable X, and data $y_1, \ldots, y_N \in \mathbb{R}$, representing a dependent variable Y. The model of their relationship is given by

$$Y = f(X) + \epsilon,$$

where f is an arbitrary function, and ϵ is a random variable representing the error. Contrary to linear regression, the function f is not assumed to have any particular parametric form.

Instead, the function f is estimated from the data by local weighted averaging.

$$\hat{f}_h(x) = \frac{\sum_{i=1}^{N} K_h(x - x_i) y_i}{\sum_{i=1}^{N} K_h(x - x_i)}.$$

In this equation, K is a function that satisfies $\int K(t)\, dt = 1$ and $K_h(t) = \frac{1}{h} K(\frac{t}{h})$, with bandwidth parameter $h > 0$. This is the estimation procedure shown in Figure 4.1 (red curves).

4.2. Regression on Riemannian Manifolds

4.2.1. Geodesic Regression

Let $y_1, \ldots, y_N$ be points on a smooth Riemannian manifold M, with associated scalar values $x_1, \ldots, x_N \in \mathbb{R}$. The goal of geodesic regression is to find a geodesic curve γ on M that best models the relationship between the x_i and the y_i. Just as in linear regression, the speed of the geodesic will be proportional to the independent parameter corresponding to the x_i. Estimation will be set up as a least-squares problem, where we want to minimize the sum-of-squared Riemannian distances between the model and the data. A schematic of the geodesic regression model is shown in Figure 2.2.

Notice that the tangent bundle TM serves as a convenient parameterization of the set of possible geodesics on M. An element $(p, v) \in TM$ provides an intercept p and a slope v, analogous to the α and β parameters in the multiple linear regression model (2.10). In fact, β is a vector in the tangent space $T_\alpha \mathbb{R}^d \cong \mathbb{R}^d$, and thus (α, β) is an element of the tangent bundle $T\mathbb{R}^d$. Now consider an M-valued random variable Y and

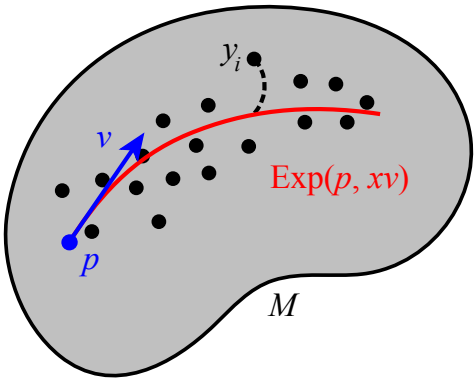


Figure 2.2 Schematic of the geodesic regression model.

a non-random variable $X \in \mathbb{R}$. The generalization of the multiple linear model to the manifold setting is the *geodesic model*,

$$Y = \mathrm{Exp}(\mathrm{Exp}(p, Xv), \epsilon), \tag{2.12}$$

where ϵ is a random variable taking values in the tangent space at $\mathrm{Exp}(p, Xv)$. Notice that for Euclidean space, the exponential map is simply addition, i.e., $\mathrm{Exp}(p, v) = p + v$. Thus, the geodesic model coincides with (2.10) when $M = \mathbb{R}^d$.

Least Squares Estimation

Consider a realization of the model (2.12): $(x_i, y_i) \in \mathbb{R} \times M$, for $i = 1, \ldots, N$. Given this data, we wish to find estimates of the parameters $(p, v) \in TM$. First, define the sum-of-squared error of the data from the geodesic given by (p, v) as

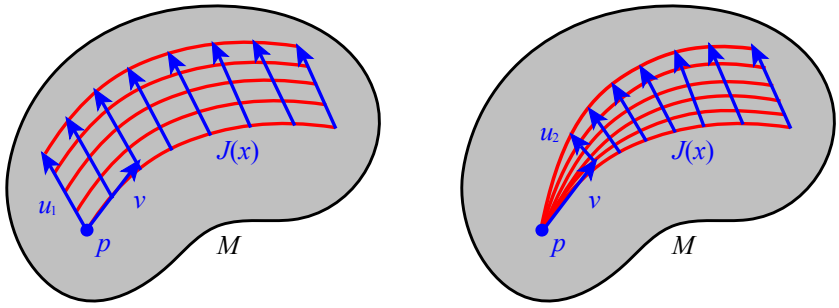
$$E(p, v) = \frac{1}{2} \sum_{i=1}^{N} d(\mathrm{Exp}(p, x_i v), y_i)^2. \tag{2.13}$$

Following the ordinary least squares minimization problem given by (2.11), we formulate a least squares estimator of the geodesic model as a minimizer of the above sum-of-squares energy, i.e.,

$$(\hat{p}, \hat{v}) = \arg\min_{(p,v)} E(p, v). \tag{2.14}$$

Again, notice that this problem coincides with the ordinary least squares problem when $M = \mathbb{R}^d$.

Unlike the linear setting, the least squares problem in (2.14) for a general manifold M will typically not yield an analytic solution. Instead we derive a gradient descent algorithm. Computation of the gradient of (2.13) will require two parts: the derivative



$d_p \operatorname{Exp}$ $d_v \operatorname{Exp}$

Figure 2.3 Jacobi fields as derivatives of the exponential map.

of the Riemannian distance function and the derivative of the exponential map. Fixing a point $p \in M$, the gradient of the squared distance function is $\nabla_x d(p,x)^2 = -2\operatorname{Log}_x(p)$ for $x \in V(p)$.

The derivative of the exponential map $\operatorname{Exp}(p,v)$ can be separated into a derivative with respect to the initial point p and a derivative with respect to the initial velocity v. To do this, first consider a variation of geodesics given by $c_1(s,t) = \operatorname{Exp}(\operatorname{Exp}(p, su_1), tv(s))$, where $u_1 \in T_pM$ defines a variation of the initial point along the geodesic $\eta(s) = \operatorname{Exp}(p, su_1)$. Here we have also extended $v \in T_pM$ to a vector field $v(s)$ along η via parallel translation. This variation is illustrated on the left side of Figure 2.3. Next consider a variation of geodesics $c_2(s,t) = \operatorname{Exp}(p, su_2 + tv)$, where $u_2 \in T_pM$. (Technically, u_2 is a tangent to the tangent space, i.e., an element of $T_v(T_pM)$, but there is a natural isomorphism $T_v(T_pM) \cong T_pM$.) The variation c_2 produces a "fan" of geodesics as seen on the right side of Figure 2.3.

Now the derivatives of $\operatorname{Exp}(p,v)$ with respect to p and v are given by

$$d_p \operatorname{Exp}(p,v) \cdot u_1 = \frac{d}{ds} c_1(s,t)\Big|_{s=0} = J_1(1)$$
$$d_v \operatorname{Exp}(p,v) \cdot u_2 = \frac{d}{ds} c_2(s,t)\Big|_{s=0} = J_2(1),$$

where $J_i(t)$ are *Jacobi fields* along the geodesic $\gamma(t) = \operatorname{Exp}(p,tv)$. Jacobi fields are solutions to the second order equation

$$\frac{D^2}{dt^2} J(t) + R(J(t), \gamma'(t))\,\gamma'(t) = 0, \tag{2.15}$$

where R is the Riemannian curvature tensor. For more details on the derivation of the Jacobi field equation and the curvature tensor, see for instance [dC92]. The initial conditions for the two Jacobi fields above are $J_1(0) = u_1, J_1'(0) = 0$ and $J_2(0) =$

$0, J_2'(0) = u_2$, respectively. If we decompose the Jacobi field into a component tangential to γ and a component orthogonal, i.e., $J = J^\top + J^\perp$, the tangential component is linear: $J^\top(t) = u_1^\top + tu_2^\top$. Therefore, the only challenge is to solve for the orthogonal component.

Finally, the gradient of the sum-of-squares energy in (2.13) is given by

$$\nabla_p E(p, v) = -\sum_{i=1}^{N} d_p \operatorname{Exp}(p, x_i v)^\dagger \operatorname{Log}(\operatorname{Exp}(p, x_i v), y_i),$$
$$\nabla_v E(p, v) = -\sum_{i=1}^{N} x_i\, d_v \operatorname{Exp}(p, x_i v)^\dagger \operatorname{Log}(\operatorname{Exp}(p, x_i v), y_i),$$

where we have taken the adjoint of the exponential map derivative, e.g., defined by $\langle d_p \operatorname{Exp}(p, v)u, w\rangle = \langle u, d_p \operatorname{Exp}(p, v)^\dagger w\rangle$. As we will see in the next section, formulas for Jacobi fields and their respective adjoint operators can often be derived analytically for many useful manifolds.

R^2 Statistics and Hypothesis Testing

In regression analysis the most basic question one would like to answer is whether the relationship between the independent and dependent variables is significant. A common way to test this is to see if the amount of variance explained by the model is high. For geodesic regression we will measure the amount of explained variance using a generalization of the R^2 statistic, or coefficient of determination, to the manifold setting. To do this, we first define predicted values of y_i and the errors ϵ_i as

$$\hat{y}_i = \operatorname{Exp}(\hat{p}, x_i \hat{v}),$$
$$\hat{\epsilon}_i = \operatorname{Log}(\hat{y}_i, y_i),$$

where $(\hat{p}, \hat{v})$ are the least squares estimates of the geodesic parameters defined above. Note that the $\hat{y}_i$ are points along the estimated geodesic that are the best predictions of the y_i given only the x_i. The $\hat{\epsilon}_i$ are the residuals from the model predictions to the true data.

Now to define the total variance of data, $y_1, \ldots, y_N \in M$, we use the Fréchet variance, intrinsically defined by

$$\operatorname{var}(y_i) = \min_{y \in M} \frac{1}{N} \sum_{i=1}^{N} d(y, y_i)^2.$$

The unexplained variance is the variance of the residuals, $\operatorname{var}(\hat{\epsilon}_i) = \frac{1}{N}\sum \|\hat{\epsilon}_i\|^2$. From the definition of the residuals, it can be seen that the unexplained variance is the mean squared distance of the data to the model, i.e., $\operatorname{var}(\hat{\epsilon}_i) = \frac{1}{N}\sum d(\hat{y}_i, y_i)^2$. Using these two

variance definitions, the generalization of the R^2 statistic is then given by

$$R^2 = 1 - \frac{\text{unexplained variance}}{\text{total variance}} = 1 - \frac{\text{var}(\hat{\epsilon}_i)}{\text{var}(y_i)}. \tag{2.16}$$

Fréchet variance coincides with the standard definition of variance when $M = \mathbb{R}^d$. Therefore, it follows that the definition of R^2 in (2.16) coincides with the R^2 for linear regression when $M = \mathbb{R}^d$. Also, because Fréchet variance is always nonnegative, we see that $R^2 \leq 1$, and that $R^2 = 1$ if and only if the residuals to the model are exactly zero, i.e., the model perfectly fits the data. Finally, it is clear that the residual variance is always smaller than the total variance, i.e., $\text{var}(\hat{\epsilon}_i) \leq \text{var}(y_i)$. This is because we could always choose $\hat{p}$ to be the Fréchet mean and $v = 0$ to achieve $\text{var}(\hat{\epsilon}_i) = \text{var}(y_i)$. Therefore, $R^2 \geq 0$, and it must lie in the interval $[0, 1]$, as is the case for linear models.

We now describe a permutation test for testing the significance of the estimated slope term, $\hat{v}$. Notice that if we constrain v to be zero in (2.14), then the resulting least squares estimate of the intercept, $\hat{p}$, will be the Fréchet mean of the y_i. The desired hypothesis test is whether the fraction of unexplained variance is significantly decreased by also estimating v. The null hypothesis is $H_0 : R^2 = 0$, which is the case if the unexplained variance in the geodesic model is equal to the total variance. Under the null hypothesis, there is no relationship between the X variable and the Y variable. Therefore, the x_i are exchangeable under the null hypothesis, and a permutation test may randomly reorder the x_i data, keeping the y_i fixed. Estimating the geodesic regression parameters for each random permutation of the x_i, we can calculate a sequence of R^2 values, $R^2_1, \ldots, R^2_m$, which approximate the sampling distribution of the R^2 statistic under the null hypothesis. Computing the fraction of the R^2_k that are greater than the R^2 estimated from the unpermuted data gives us a p-value.

4.2.2. Kernel Regression on Manifolds

The regression method of Davis et al. [DFBJ07] generalizes the Nadaraya-Watson kernel regression method to the case where the dependent variable lives on a Riemannian manifold, i.e., $y_i \in M$. Here the model is given by

$$Y = \text{Exp}(f(X), \epsilon),$$

where $f : \mathbb{R} \to M$ defines a curve on M, and $\epsilon \in T_{f(X)}M$ is an error term. As in the univariate case, there are no assumptions on the parametric form of the curve f.

Motivated by the definition of the Nadaraya-Watson estimator as a weighted averaging, the *manifold kernel regression estimator* is defined using a weighted Fréchet sample mean as

$$\hat{f}_h(x) = \arg\min_{y \in M} \frac{\sum_{i=1}^N K_h(x - x_i) d(y, y_i)^2}{\sum_{i=1}^N K_h(x - x_i)}.$$

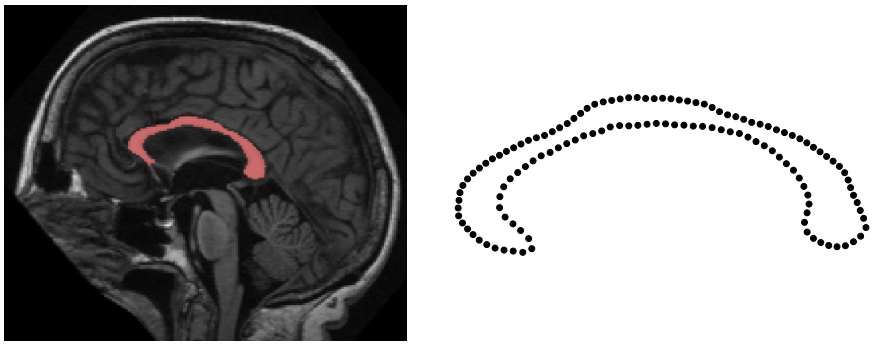


Figure 2.4 Corpus callosum segmentation and boundary point model for one subject.

Notice that when the manifold under study is a Euclidean vector space, equipped with the standard Euclidean norm, the above minimization results in the Nadaraya-Watson estimator.

4.3. Example of Regression on Kendall Shape Space

We now give examples of both geodesic and nonparametric kernel regression in Kendall shape space (see Chapter **??**). The goal of our statistical analysis is to understand the relationship between age and the shape of the corpus callosum. The corpus callosum is the major white matter bundle connecting the two hemispheres of the brain. A midsagittal slice from a magnetic resonance image (MRI) with segmented corpus callosum is shown in Figure 2.4. The data used is derived from the OASIS brain database (`www.oasis-brains.org`), and this regression analysis originally appears in [Fle12].

The data consisted of MRI from 32 subjects with ages ranging from 19-90 years old. The corpus callosum was segmented in a midsagittal slice using the ITK SNAP program (`www.itksnap.org`). These boundaries of these segmentations were sampled with 128 points using ShapeWorks (`www.sci.utah.edu/software.html`). This algorithm generates a sampling of a set of shape boundaries while enforcing correspondences between different point models within the population. An example of a segmented corpus callosum and the resulting boundary point model is shown in Figure 2.4. Each of these preprocessing steps were done without consideration of the subject age, to avoid any bias in the data generation.

The statistical significance of the estimated trend was tested using the permutation test described in Section 4.2.1, using 10,000 permutations. The p-value for the significance of the slope estimate, $\hat{v}$, was $p = 0.009$. The coefficient of determination (for the unpermuted data) was $R^2 = 0.12$. The low R^2 value must be interpreted carefully. It says that age only describes a small fraction of the shape variability in the corpus callosum. This is not surprising: we would expect the intersubject variability in corpus

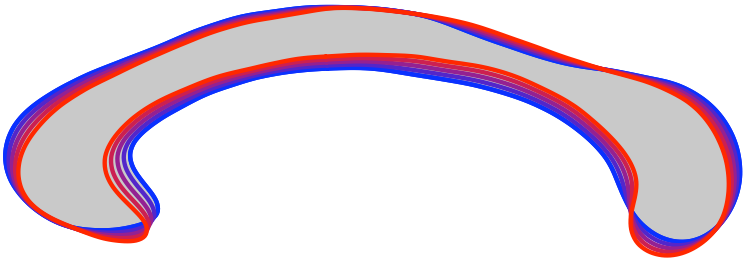


Figure 2.5 Geodesic regression of the corpus callosum. The estimated geodesic is shown as a sequence of shapes from age 19 (blue) to age 90 (red).

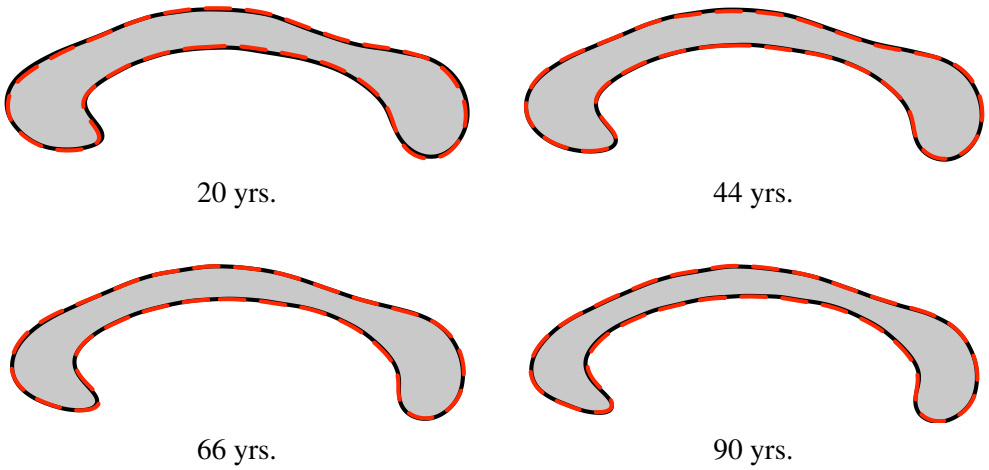


Figure 2.6 Comparison of geodesic regression (solid black) and nonparametric kernel regression (dashed red) of the corpus callosum shape versus age.

callosum shape to be difficult to fully describe with a single variable (age). However, this does not mean that the age effects are not important. In fact, the low p-value says that the estimated age changes are highly unlikely to have been found by random chance.

Next, we computed a nonparametric kernel regression of the corpus callosum versus age, as described in Section 4.2.2. The kernel regression was performed on the same Kendall shape space manifold and the bandwidth was chosen automatically using the cross-validation procedure described in Section [DFBJ07]. Next, the resulting corpus callosum shape trend generated by the kernel regression method was compared to the result of the geodesic regression. This was done by again generating shapes from the geodesic model $\hat{\gamma}(x_k)$ at a sequence of ages, x_k, and overlaying the corresponding generated shapes from the kernel regression model at the same ages. The results are

plotted for ages x_k = 20, 44, 66, and 90. Both regression methods give strikingly similar results. The two regression models at other values of ages, not shown, are also close to identical. This indicates that a geodesic curve does capture the relationship between age and corpus callosum shape, and that the additional flexibility offered by the nonparametric regression does not change the estimated trend. However, even though both methods provide a similar estimate of the trend, the geodesic regression has the advantage that it is simpler to compute and easier to interpret, from the standpoint of the R^2 statistic and hypothesis test demonstrated above.

5. Probabilistic Models

5.1. Normal Densities on Manifolds

In this section we review probabilistic formulations for geodesic regression and PGA. Before defining these models, we first consider a basic definition of a manifold-valued normal distribution and give procedures for maximum-likelihood estimation of its parameters. There is no standard definition of a normal distribution on manifolds, mainly because different properties of the multivariate normal distribution in $\mathbb{R}^d$ may be generalized to manifolds by different definitions. Grenander [Gre63] defines a generalization of the normal distribution to Lie groups and homogeneous spaces as a solution to the heat equation. Pennec [Pen06] defines a generalization of the normal distribution in the tangent space to a mean point via the Riemannian Log map. The definition that we use here, introduced in [Fle12], and also used in [JH14, ZF13], generalizes the connection between least-squares estimation of statistical models and maximum-likelihood estimation under normally distributed errors.

Consider a random variable y taking values on a Riemannian manifold M, defined by the probability density function (pdf)

$$p(y;\mu,\tau) = \frac{1}{C(\mu,\tau)} \exp\left(-\frac{\tau}{2} d(\mu,y)^2\right), \tag{2.17}$$

$$C(\mu,\tau) = \int_M \exp\left(-\frac{\tau}{2} d(\mu,y)^2\right) dy, \tag{2.18}$$

where $C(\mu,\tau)$ is a normalizing constant. We term this distribution a *Riemannian normal distribution*, and use the notation $y \sim N_M(\mu,\tau^{-1})$ to denote it. The parameter $\mu \in M$ acts as a location parameter on the manifold, and the parameter $\tau \in \mathbb{R}_+$ acts as a dispersion parameter, similar to the precision of a Gaussian. This distribution has the advantages that (a) it is applicable to any Riemannian manifold, (b) it reduces to a multivariate normal distribution (with isotropic covariance) when $M = \mathbb{R}^d$, and (c) much like the Euclidean normal distribution, maximum-likelihood estimation of parameters gives rise to least-squares methods when M is a Riemannian homogeneous space, as shown next.

5.1.1. Maximum-Likelihood Estimation of μ

Returning to the Riemannian normal density in (2.17), the maximum-likelihood estimate of the mean parameter, μ, is given by

$$\begin{aligned}\hat{\mu} &= \arg\max_{\mu\in M} \sum_{i=1}^{N} \ln p(y_i;\mu,\tau) \\ &= \arg\min_{\mu\in M} N \ln C(\mu,\tau) + \frac{\tau}{2}\sum_{i=1}^{N} d(\mu,y_i)^2.\end{aligned}$$

This minimization problem clearly reduces to the least-squares estimate, or Fréchet mean in (2.2), if the normalizing constant $C(\mu,\tau)$ does not depend on the μ parameter. As shown in [Fle12], this occurs when the manifold M is a Riemannian homogeneous space, which means that for any two points $x, y \in M$, there exists an isometry that maps x to y. This is because the integral in (2.18) is invariant under isometries. More precisely, given any two points $\mu, \mu' \in M$, there exists an isometry $\phi : M \to M$, with $\mu' = \phi(\mu)$, and we have

$$\begin{aligned}C(\mu,\tau) &= \int_M \exp\left(-\frac{\tau}{2}d(\mu,y)^2\right)dy \\ &= \int_M \exp\left(-\frac{\tau}{2}d(\phi(\mu),\phi(y))^2\right)d\phi(y) \\ &= C(\mu',\tau).\end{aligned}$$

Thus, in the case of a Riemannian homogeneous space, the normalizing constant can be written as

$$C(\tau) = \int_M \exp\left(-\frac{\tau}{2}d(\mu,y)^2\right)dy, \tag{2.19}$$

and we have equivalence of the MLE and Fréchet mean, i.e., $\hat{\mu} = \bar{y}$.

Two properties of the Riemannian normal distribution are worth emphasizing at this point. First, the requirement that M be a Riemannian homogeneous space is important. Without this, the normalizing constant $C(\mu,\tau)$ may be a function of μ, and if so, the MLE will not coincide with the Fréchet mean. For example, a Riemannian normal distribution on an anisotropic ellipsoid (which is not a homogeneous space) will have a normalizing constant that depends on μ. Second, it is also important that the Riemannian normal density be isotropic, unlike the normal law in [Pen06], which includes a covariance matrix in the tangent space to the mean. Again, a covariance tensor field would need to be a function of the mean point, μ, which would cause the normalizing constant to change with μ. That is, unless the covariant derivative of the covariance field was zero everywhere. Unfortunately, such tensor fields are not always possible on general homogeneous spaces. For example, the only symmetric,

second-order tensor fields with zero covariant derivatives on S^2 are isotropic.

5.1.2. Estimation of the Dispersion Parameter, τ

Maximum-likelihood estimation of the dispersion parameter, τ, can also be done using gradient ascent. Unlike the case for estimation of the μ parameter, now the normalizing constant is a function of τ, and we must evaluate its derivative. We can rewrite the integral in (2.19) in normal coordinates, which can be thought of as a polar coordinate system in the tangent space, $T_\mu M$. The radial coordinate is defined as $r = d(\mu, y)$, and the remaining $n-1$ coordinates are parameterized by a unit vector v, i.e., a point on the unit sphere $S^{n-1} \subset T_\mu M$. Thus we have the change-of-variables, $\phi(rv) = \mathrm{Exp}(\mu, rv)$. Now the integral for the normalizing constant becomes

$$C(\tau) = \int_{S^{n-1}} \int_0^{R(v)} \exp\left(-\frac{\tau}{2} r^2\right) |\det(d\phi(rv))| \, dr \, dv, \tag{2.20}$$

where $R(v)$ is the maximum distance that $\phi(rv)$ is defined. Note that this formula is only valid if M is a complete manifold, which guarantees that normal coordinates are defined everywhere except possibly a set of measure zero on M.

The integral in (2.20) is difficult to compute for general manifolds, due to the presence of the determinant of the Jacobian of ϕ. However, for symmetric spaces this change-of-variables term has a simple form. If M is a symmetric space, there exists a orthonormal basis $u_1, \ldots, u_n$, with $u_1 = v$, such that

$$|\det(d\phi(rv))| = \prod_{k=2}^{d} f_k(r), \tag{2.21}$$

where $\kappa_k = K(u_1, u_k)$ denotes the sectional curvature, and f_k is defined as

$$f_k(r) = \begin{cases} \frac{1}{\sqrt{\kappa_k}} \sin(\sqrt{\kappa_k} r) & \text{if } \kappa_k > 0, \\ \frac{1}{\sqrt{-\kappa_k}} \sinh(\sqrt{-\kappa_k} r) & \text{if } \kappa_k < 0, \\ r & \text{if } \kappa_k = 0. \end{cases}$$

Notice that with this expression for the Jacobian determinant there is no longer a dependence on v inside the integral in (2.20). Also, if M is simply connected, then $R(v) = R$ does not depend on the direction v, and we can write the normalizing constant as

$$C(\tau) = A_{n-1} \int_0^R \exp\left(-\frac{\tau}{2} r^2\right) \prod_{k=2}^{d} |\kappa_k|^{-1/2} f_k(\sqrt{|\kappa_k|} r) dr,$$

where A_{n-1} is the surface area of the $n-1$ hypersphere, S^{n-1}. While this formula works only for simply connected symmetric spaces, other symmetric spaces could be handled by lifting to the universal cover, which is simply connected, or by restricting

the definition of the Riemannian normal pdf in (2.17) to have support only up to the injectivity radius, i.e., $R = \min_v R(v)$.

The derivative of the normalizing constant with respect to τ is

$$C'(\tau) = A_{n-1} \int_0^R \frac{r^2}{2} \exp\left(-\frac{\tau}{2} r^2\right) \prod_{k=2}^{d} |\kappa_k|^{-1/2} f_k(\sqrt{|\kappa_k|} r) dr. \tag{2.22}$$

Both $C(\tau)$ and $C'(\tau)$ involve only a one-dimensional integral, which can be quickly and accurately approximated by numerical integration. Finally, the derivative of the log-likelihood needed for gradient ascent is given by

$$\frac{d}{d\tau} \sum_{i=1}^{N} \ln p(y_i; \mu, \tau) = -N \frac{C'(\tau)}{C(\tau)} - \frac{1}{2} \sum_{i=1}^{N} d(\mu, y_i)^2.$$

5.1.3. Sampling from a Riemannian Normal Distribution

In this section, we describe a Markov Chain Monte Carlo (MCMC) method for sampling from a Riemannian normal distribution with given mean and dispersion parameters, (μ, τ). From (2.20) we see that the Riemannian normal density is proportional to an isotropic Gaussian density in $T_\mu M$ times a change-of-variables term. This suggests using an independence sampler with an isotropic Gaussian as the proposal density.

More specifically, let $y \sim N_M(\mu, \tau^{-1})$, and let $\phi(rv) = \mathrm{Exp}(\mu, rv)$ be normal coordinates in the tangent space $T_\mu M$. Then the density in (r, v) is given by

$$f(r, v) \propto \begin{cases} \exp\left(-\frac{\tau}{2} r^2\right) |\det(d\phi(rv))| & r \le R(v), \\ 0 & \text{otherwise.} \end{cases}$$

Notice that the density is zero beyond the cut locus. For the independence sampler, we will not need to compute the normalization constant. We will then use an isotropic (Euclidean) Gaussian in $T_\mu M$ as the proposal density, which in polar coordinates is given by

$$g(r, v) \propto r \exp\left(-\frac{\tau}{2} r^2\right).$$

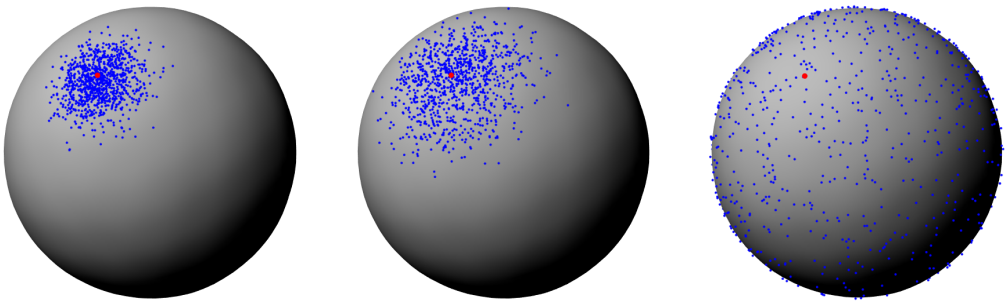
An iteration of the independence sampler begins with the previous sample (r, v) and generates a proposal sample $(\tilde{r}, \tilde{v})$ from g, which is accepted with probability

$$\begin{aligned} \alpha((\tilde{r}, \tilde{v}), (r, v)) &= \min\left\{1, \frac{f(\tilde{r}, \tilde{v}) g(r, v)}{f(r, v) g(\tilde{r}, \tilde{v})}\right\} \\ &= \min\left\{1, \left|\frac{r \det(d\phi(\tilde{r}\tilde{v}))}{\tilde{r} \det(d\phi(rv))}\right|\right\}, \end{aligned} \tag{2.23}$$

So, the acceptance probability reduces to simply a ratio of the Log map change-of-variables factors, which for symmetric spaces can be computed using (2.21). The final

Algorithm 3: Independence sampler for the Riemannian normal distribution

Input: Parameters μ, τ
Draw initial sample (r, v) from g
for $i = 1$ **to** S **do**
  Sample proposal $(\tilde{r}, \tilde{v})$ from g
  Compute the acceptance probability $\alpha((\tilde{r}, \tilde{v}), (r, v))$ using (2.23)
  Draw a uniform random number $u \in [0, 1]$
  if $\tilde{r} \leq R(\tilde{v})$ **AND** $u \leq \alpha((\tilde{r}, \tilde{v}), (r, v))$ **then**
    Accept: Set $y_i = \mathrm{Exp}_(\mu, \tilde{r}\tilde{v})$, and set $(r, v) = (\tilde{r}, \tilde{v})$
  else
    Reject: Set $y_i = \mathrm{Exp}_(\mu, rv)$



$\tau = 50$
Accept rate = 0.995

$\tau = 20$
Accept rate = 0.991

$\tau = 1$
Accept rate = 0.833

Figure 2.7 Samples from a Riemannian normal density on S^2 for various levels of τ. Samples are in blue, and the mean parameter, μ, is shown in red.

MCMC procedure is given by Algorithm 3.

5.1.4. Sphere Example

We now demonstrate the above procedures for sampling from Riemannian normal densities and ML estimation of parameters on the two-dimensional sphere, S^2. Figure 5.1.3 shows example samples generated using the independence sampler in Algorithm 3 for various levels of τ. Notice that the sampler is efficient (high acceptance rate) for larger values of τ, but less efficient for smaller τ as the distribution approaches a uniform distribution on the sphere. This is because the proposal density matches the true density well, but the sampler rejects points beyond the cut locus, which happen more frequently when τ is small and the distribution is approaching the uniform distribution on the sphere.

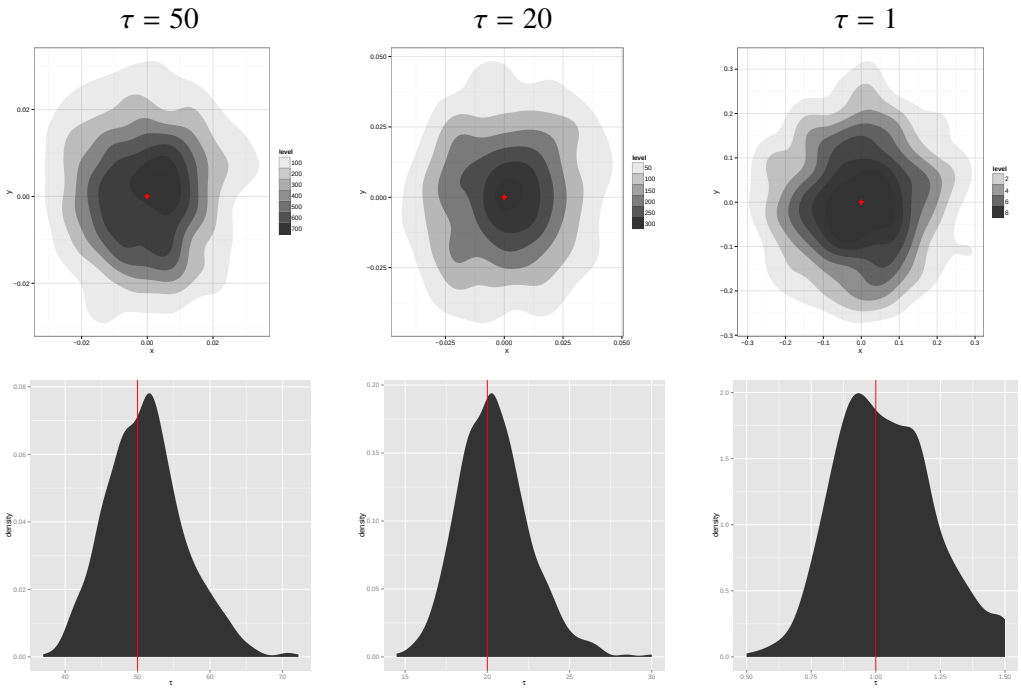


Figure 2.8 Monte Carlo simulation of the MLEs, $\hat{\mu}$ (top row), and $\hat{\tau}$ (bottom row). The true parameter values are marked in red.

Next, to test the ML estimation procedures, we used the independence sampler to repeatedly generate $N = 100$ random points on S^2 from a $N_{S^2}(\mu, \tau)$ density, where $\mu = (0, 0, 1)$ was the north pole, and again we varied $\tau = 1, 20, 50$. Then we computed the MLEs, $\hat{\mu}, \hat{\tau}$, using the gradient ascent procedures above. Each experiment was repeated 1000 times, and the results are summarized in Figure 5.1.4. For the $\hat{\mu}$ estimates, we plot a kernel density estimate of the points $\text{Log}_{\mu}\hat{\mu}$. This is a Monte Carlo simulation of the sampling distribution of the $\hat{\mu}$ statistic, mapped into the tangent space of the true mean, $T_{\mu}M$, via the Log map. Similarly, the corresponding empirical sampling distribution of the $\hat{\tau}$ statistics are plotted as kernel density estimates. While the true sampling distributions are unknown, the plots demonstrate that the MLEs have reasonable behavior, i.e., they are distributed about the true parameter values, and their variance decreases as τ increases.

5.2. Probabilistic Principal Geodesic Analysis

Principal component analysis (PCA) [Jol86b] has been widely used to analyze high-dimensional Euclidean data. Tipping and Bishop proposed probabilistic PCA (PPCA)

[TB99], which is a latent variable model for PCA. A similar formulation was independently proposed by Roweis [Row98]. The main idea of PPCA is to model an n-dimensional Euclidean random variable y as

$$y = \mu + Bx + \epsilon, \tag{2.24}$$

where μ is the mean of y, x is a q-dimensional latent variable, with $x \sim N(0, I)$, B is an $n \times q$ factor matrix that relates x and y, and $\epsilon \sim N(0, \sigma^2 I)$ represents error. We will find it convenient to model the factors as $B = W\Lambda$, where the columns of W are mutually orthogonal, and Λ is a diagonal matrix of scale factors. This removes the rotation ambiguity of the latent factors and makes them analogous to the eigenvectors and eigenvalues of standard PCA (there is still of course an ambiguity of the ordering of the factors). We now generalize this model to random variables on Riemannian manifolds.

5.2.1. Probability Model

The PPGA model for a random variable y on a smooth Riemannian manifold M is

$$y|x \sim N_M\left(\mathrm{Exp}(\mu, z), \tau^{-1}\right), z = W\Lambda x, \tag{2.25}$$

where $x \sim N(0, 1)$ are again latent random variables in $\mathbb{R}^q$, μ here is a base point on M, W is a matrix with q columns of mutually orthogonal tangent vectors in $T_\mu M$, Λ is a $q \times q$ diagonal matrix of scale factors for the columns of W, and τ is a scale parameter for the noise. In this model, a linear combination of $W\Lambda$ and the latent variables x forms a new tangent vector $z \in T_\mu M$. Next, the exponential map shoots the base point μ by z to generate the location parameter of a *Riemannian normal distribution*, from which the data point y is drawn. Note that in Euclidean space, the exponential map is an addition operation, $\mathrm{Exp}(\mu, z) = \mu + z$. Thus, PPGA coincides with (2.24), the standard PPCA model, when $M = \mathbb{R}^d$.

5.2.2. Inference

We develop a maximum likelihood procedure to estimate the parameters $\theta = (\mu, W, \Lambda, \tau)$ of the PPGA model defined in (2.25). Given observed data $y_i \in \{y_1, ..., y_N\}$ on M, with associated latent variable $x_i \in \mathbb{R}^q$, and $z_i = W\Lambda x_i$, we formulate an expectation maximization (EM) algorithm. Since the expectation step over the latent variables does not yield a closed-form solution, we develop a HMC method to sample x_i from the posterior $p(x|y;\theta)$, the log of which is given by

$$\log \prod_{i=1}^{N} p(x_i|y_i;\theta) \propto -N \log C - \sum_{i=1}^{N} \frac{\tau}{2} d\left(\mathrm{Exp}(\mu, z_i), y_i\right)^2 - \frac{\|x_i\|^2}{2}, \tag{2.26}$$

and use this in a Monte Carlo Expectation Maximization (MCEM) scheme to estimate θ. The procedure contains two main steps:

5.2.3. E-step: HMC

For each x_i, we draw a sample of size S from the posterior distribution (2.26) using HMC with the current estimated parameters θ^k. Denote x_{ij} as the jth sample for x_i, the Monte Carlo approximation of the Q function is given by

$$Q(\theta|\theta^k) = E_{x_i|y_i;\theta^k}\left[\prod_{i=1}^{N} \log p(x_i|y_i;\theta^k)\right] \approx \frac{1}{S}\sum_{j=1}^{S}\sum_{i=1}^{N} \log p(x_{ij}|y_i;\theta^k). \tag{2.27}$$

Hamiltonian Monte Carlo (HMC) [DKPR87] is a powerful gradient-based Markov Chain Monte Carlo sampling method that is applicable to a wide array of continuous probability distributions. It rigorously explores the entire space of a target distribution by utilizing Hamiltonian dynamics as a Markov transition probability. The gradient information of the log probability density is used to efficiently sample from the higher probability regions.

Next, we derive an HMC procedure to draw a random sample from the posterior distribution of the latent variables x. The first step to sample from a distribution $f(x)$ using HMC is to construct a Hamiltonian system $H(x, m) = U(x) + V(m)$, where $U(x) = -\log f(x)$ is a "potential energy", and $V(m) = -\log g(m)$ is a "kinetic energy", which acts as a proposal distribution on an auxiliary momentum variable, m. An initial random momentum m is drawn from the density $g(m)$. Starting from the current point x and initial random momentum m, the Hamiltonian system is integrated forward in time to produce a candidate point, x^*, along with the corresponding forward-integrated momentum, m^*. The candidate point x^* is accepted as a new point in the sample with probability

$$P(\text{accept}) = \min(1, \exp(-U(x^*) - V(m^*) + U(x) + V(m)).$$

This acceptance-rejection method is guaranteed to converge to the desired density $f(x)$ under fairly general regularity assumptions on f and g.

In the HMC sampling procedure, the potential energy of the Hamiltonian $H(x_i, m) = U(x_i) + V(m)$ is defined as $U(x_i) = -\log p(x_i|y_i;\theta)$, and the kinetic energy $V(m)$ is a typical isotropic Gaussian distribution on a q-dimensional auxiliary momentum variable, m. This gives us a Hamiltonian system to integrate: $\frac{dx_i}{dt} = \frac{\partial H}{\partial m} = m$, and $\frac{dm}{dt} = -\frac{\partial H}{\partial x_i} = -\nabla_{x_i} U$. Due to the fact that x_i is a Euclidean variable, we use a standard "leapfrog" numerical integration scheme, which approximately conserves the Hamiltonian and results in high acceptance rates. Now, the gradient with respect to each x_i is

$$\nabla_{x_i} U = x_i - \tau\Lambda W^T\{d_{z_i}\,\mathrm{Exp}(\mu, z_i)^{\dagger}\mathrm{Log}(\mathrm{Exp}(\mu, z_i), y_i)\}. \tag{2.28}$$

M-step: Gradient Ascent

In this section, we derive the maximization step for updating the parameters $\theta = (\mu, W, \Lambda, \tau)$ by maximizing the HMC approximation of the Q function in (2.27). This

turns out to be a gradient ascent scheme for all the parameters since there are no closed-form solutions.

Gradient for τ:
The gradient term for estimating τ is

$$\nabla_\tau Q = -N\frac{C'(\tau)}{C(\tau)} - \frac{1}{S}\sum_{i=1}^{N}\sum_{j=1}^{S} d(\mathrm{Exp}(\mu, z_{ij}), y_i)^2,$$

where the derivative $C'(\tau)$ is given in (2.22).

Gradient for μ:
From (2.26) and (2.27), the gradient term for updating μ is

$$\nabla_\mu Q = \frac{1}{S}\sum_{i=1}^{N}\sum_{j=1}^{S} \tau d_\mu \,\mathrm{Exp}(\mu, z_{ij})^\dagger \mathrm{Log}\Big(\mathrm{Exp}(\mu, z_{ij}), y_i\Big).$$

Gradient for Λ:
For updating Λ, we take the derivative w.r.t. each ath diagonal element Λ^a as

$$\frac{\partial Q}{\partial \Lambda^a} = \frac{1}{S}\sum_{i=1}^{N}\sum_{j=1}^{S} \tau (W^a x_{ij}^a)^T \{d_{z_{ij}} \mathrm{Exp}(\mu, z_{ij})^\dagger \mathrm{Log}(\mathrm{Exp}(\mu, z_{ij}), y_i)\},$$

where W^a denotes the ath column of W, and x_{ij}^a is the ath component of x_{ij}.

Gradient for W:
The gradient w.r.t. W is

$$\nabla_W Q = \frac{1}{S}\sum_{i=1}^{N}\sum_{j=1}^{S} \tau d_{z_{ij}} \mathrm{Exp}(\mu, z_{ij})^\dagger \,\mathrm{Log}(\mathrm{Exp}(\mu, z_{ij}), y_i) x_{ij}^T \Lambda. \tag{2.29}$$

To preserve the mutual orthogonality constraint on the columns of W, we project the gradient in (2.29) onto the tangent space at W, then updating W by shooting the geodesic on the Stiefel manifold in the negative projected gradient direction, see the detail in [EAS98].

The MCEM algorithm for PPGA is an iterative procedure for finding the subspace spanned by q principal components, shown in Algorithm 4. The computation time per iteration depends on the complexity of exponential map, log map, and Jacobi field which may vary for different manifold. Note the cost of the gradient ascent algorithm also linearly depends on the data size, dimensionality, and the number of samples drawn. An advantage of MCEM is that it can run in parallel for each data point. Since the posterior distribution (2.26) is estimated by HMC sampling, to diagnose the

Algorithm 4: Monte Carlo Expectation Maximization for PPGA

Input: Data set Y, reduced dimension q.
Initialize μ, W, Λ, σ.
while *gradient is larger than some threshold* **do**
  Sample X according to (2.28)
  Update μ, W, Λ, σ by gradient ascent.

convergence of the PPGA MCEM algorithm, we run parallel independent chains to obtain univariate quantities of the full distribution.

5.2.4. PPGA of Simulated Sphere Data

Using the generative model for PGA (2.25), we forward simulated a random sample of 100 data points on the unit sphere S^2, with known parameters $\theta = (\mu, W, \Lambda, \tau)$, shown in Table 2.1. Next, we ran the maximum likelihood estimation procedure to test whether we could recover those parameters. We initialized μ from a random uniform point on the sphere. We initialized W as a random Gaussian matrix, to which we then applied the Gram-Schmidt algorithm to ensure its columns were orthonormal. Figure 2.9 compares the ground truth principal geodesics and MLE principal geodesic analysis. A good overlap between the first principal geodesic shows that PPGA recovers the model parameters.

One advantage that the PPGA model has over the least-squares PGA formulation is that the mean point is estimated jointly with the principal geodesics. In the standard PGA algorithm, the mean is estimated first (using geodesic least-squares), then the principal geodesics are estimated second. This does not make a difference in the Euclidean case (principal components must pass through the mean), but it does in the nonlinear case. To demonstrate this, we give examples where data can be fit better when jointly estimating mean and PGA than when doing them sequentially. We compared the PPGA model with PGA and standard PCA (in the Euclidean embedding space). The noise variance τ was not valid to be estimated in both PGA and PCA. The estimation error of principal geodesics turned to be larger in PGA compared to PPGA. Furthermore, the standard PCA converges to an incorrect solution due to its inappropriate use of a Euclidean metric on Riemannian data. A comparison of the ground truth parameters and these methods is given in Table 2.1.

	μ	w	Λ	τ
Ground truth	$(-0.78, 0.48, -0.37)$	$(-0.59, -0.42, 0.68)$	0.40	100
PPGA	$(-0.78, 0.48, -0.40)$	$(-0.59, -0.43, 0.69)$	0.41	102
PGA	$(-0.79, 0.46, -0.41)$	$(-0.59, -0.38, 0.70)$	0.41	N/A
PCA	$(-0.70, 0.41, -0.46)$	$(-0.62, -0.37, 0.69)$	0.38	N/A

Table 2.1 Comparison between ground truth parameters for the simulated data and the MLE of PPGA, non-probabilistic PGA, and standard PCA.

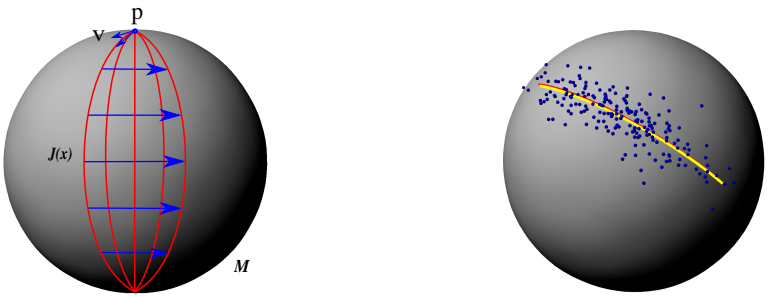


Figure 2.9 Left: Jacobi fields; Right: the principal geodesic of random generated data on unit sphere. Blue dots: random generated sphere data set. Yellow line: ground truth principal geodesic. Red line: estimated principal geodesic using PPGA.

[Asf11] B. Asfari. Riemannian l^p center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139:655–673, 2011.

[BF01] S. R. Buss and J. P. Fillmore. Spherical averages and applications to spherical splines and interpolation. *ACM Transactions on Graphics*, 20(2):95–126, 2001.

[BJV17] M. Banerjee, B. Jian, and B. C. Vemuri. Robust fréchet mean and PGA on Riemannian manifolds with applications to neuroimaging. In *International Conference on Information Processing in Medical Imaging*, 2017.

[BP03] R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. *Annals of Statistics*, 31(1):1–29, 2003.

[BP05] R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds—II. *Annals of Statistics*, 33(3):1225–1259, 2005.

[CL95] P. Crouch and F. S. Leite. The dynamic interpolation problem: on Riemannian manifolds, Lie groups, and symmetric spaces. *Journal of Dynamical and Control Systems*, 1(2):177–202, 1995.

[CSV16] R. Chakraborty, D. Seo, and B. C. Vemuri. An efficient exact-PGA algorithm for constant curvature manifolds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3976–3984, 2016.

[dC92] M. do Carmo. *Riemannian Geometry*. Birkhäuser, 1992.

[DFBJ07] B. Davis, P. T. Fletcher, E. Bullitt, and S. Joshi. Population shape regression from random design data. In *Proceedings of IEEE International Conference on Computer Vision*, 2007.

[DKPR87] S. Duane, A. Kennedy, B. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, pages 216–222, 1987.

[DPT+09] S. Durrleman, X. Pennec, A. Trouvé, G. Gerig, and N. Ayache. Spatiotemporal atlas estimation for developmental delay detection in longitudinal datasets. In *Medical Image Computing and Computer-Assisted Intervention*, pages 297–304, 2009.

[EAS98] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[EJH14] B. Eltzner, S. Jung, and S. Huckemann. Dimension reduction on polyspheres with application to skeletal representations. In *International Conference on Networked Geometric Science of Information*, pages 22–29, 2014.

[Fle11] P. T. Fletcher. Geodesic regression on Riemannian manifolds. In *MICCAI Workshop on Mathematical Foundations of Computational Anatomy*, pages 75–86, 2011.

[Fle12] P. T. Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *International Journal of Computer Vision*, pages 1–15, 2012.

[FLJ03] P. T. Fletcher, C. Lu, and S. Joshi. Statistics of shape via principal geodesic analysis on Lie groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 95–101, 2003.

[FLPJ04] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8):995–1005, 2004.

[Fré48] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré*, 10(3):215–310, 1948.

[Gre63] U. Grenander. *Probabilities on Algebraic Structures*. John Wiley and Sons, 1963.

[HHM10] S. Huckemann, T. Hotz, and A. Munk. Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, pages 1–58, 2010.

[Hot33] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 498–520, 1933.

[JDM12] S. Jung, I. L. Dryden, and J. S. Marron. Analysis of principal nested spheres. *Biometrika*, 99(3):551–568, 2012.

[JH14] S. C. Joshi J. Hinkle, P. T. Fletcher. Intrinsic polynomials for regression on riemannian manifolds. *Journal of Mathematical Imaging and Vision*, 50(1–2):32–52, 2014.

[JK87] P. E. Jupp and J. T. Kent. Fitting smooth paths to spherical data. *Applied Statistics*, 36(1):34–46, 1987.

[Jol86a] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[Jol86b] I. T. Jolliffe. *Principal Component Analysis*, volume 487. Springer-Verlag New York, 1986.

[Kar77] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Math*, 30(5):509–541, 1977.

[KDL07] A. Kume, I. L. Dryden, and H. Le. Shape-space smoothing splines for planar landmark data. *Biometrika*, 94(3):513–528, 2007.

[Ken90] W. S. Kendall. Probability, convexity, and harmonic maps with small image I: uniqueness and fine existence. *Proceedings of the London Mathematical Society*, 3(61):371–406, 1990.

[Mil04] M. Miller. Computational anatomy: shape, growth, and atrophy comparison via diffeomorphisms. *NeuroImage*, 23:S19–S33, 2004.

[Moa02] M. Moakher. Means and averaging in the group of rotations. *SIAM Journal on Matrix Analysis and Applications*, 24(1):1–16, 2002.

[Nad64] E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 10:186–190, 1964.

[NHP89] L. Noakes, G. Heinzinger, and B. Paden. Cubic splines on curved spaces. *IMA Journal of Mathematical Control and Information*, 6(4):465–473, 1989.

[NHV11] M. Niethammer, Y. Huang, and F.-X. Viallard. Geodesic regression for image time-series. In *Proceedings of Medical Image Computing and Computer Assisted Intervention*, 2011.

[Pea01] K. Pearson. On lines and planes of closest fit to points in space. *Philosophical Magazine*, 2:609–629, 1901.

[Pen99] X. Pennec. Probabilities and statistics on Riemannian manifolds: basic tools for geometric measurements. In *IEEE Workshop on Nonlinear Signal and Image Processing*, 1999.

[Pen06] X. Pennec. Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1), 2006.

[Row98] S. Roweis. EM algorithms for PCA and SPCA. *Advances in neural information processing systems*, pages 626–632, 1998.

[SCBS07] S. Said, N. Courty, N. Le Bihan, and S. J. Sangwine. Exact principal geodesic analysis for data on SO(3). In *15th European Signal Processing Conference*, pages 1701–1705, 2007.

[SDK+12] J. Su, I. L. Dryden, E. Klassen, H. Le, and A. Srivastava. Fitting smoothing splines to time-indexed, noisy points on nonlinear manifolds. *Image and Vision Computing*, 30(6):428–442, 2012.

[SLHN10] S. Sommer, F. Lauze, S. Hauberg, and M. Nielsen. Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations. In *European Conference on Computer Vision*, pages 43–56, 2010.

[SLN14] S. Sommer, F. Lauze, and M. Nielsen. Optimization over geodesics for exact principal geodesic analysis. *Advances in Computational Mathematics*, 40(2):283–313, 2014.

[SSL+09] X. Shi, M. Styner, J. Lieberman, J. Ibrahim, W. Lin, and H. Zhu. Intrinsic regression models for manifold-valued data. *J Am Stat Assoc*, 5762:192–199, 2009.

[SVV14] H. Salehian, D. Vaillancourt, and B. C. Vemuri. iPGA: Incremental principal geodesic analysis with applications to movement disorder classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2014.

[TB99] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[TV10] A. Trouvé and F.-X. Vialard. A second-order model for time-dependent data interpolation: Splines on shape spaces. In *MICCAI STIA Workshop*, 2010.

[Wat64] G. S. Watson. Smooth regression analysis. *Sankhya*, 26:101–116, 1964.

[ZF13] M. Zhang and P. T. Fletcher. Probabilistic principal geodesic analysis. In *Neural Information Processing Systems (NIPS)*, 2013.