ESTIMATING RIEMANNIAN METRIC WITH NOISE-CONTAMINATED INTRINSIC DISTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

We extend metric learning by studying the Riemannian manifold structure of the underlying data space induced by dissimilarity measures between data points. The key quantity of interest here is the Riemannian metric, which characterizes the Riemannian geometry and defines straight lines and derivatives on the manifold. Being able to estimate the Riemannian metric allows us to gain insights into the underlying manifold and compute geometric features such as the geodesic curves. We model the observed dissimilarity measures as noisy responses generated from a function of the intrinsic geodesic distance between data points. A new local regression approach is proposed to learn the Riemannian metric tensor and its derivatives based on a Taylor expansion for the squared geodesic distances. Our framework is general and accommodates different types of responses, whether they are continuous, binary, or comparative, extending the existing works which consider a single type of response at a time. We develop theoretical foundation for our method by deriving the rates of convergence for the asymptotic bias and variance of the estimated metric tensor. The proposed method is shown to be versatile in simulation studies and a real data application involving taxi trip time in New York City.

1 INTRODUCTION

The estimation of distance metric, also known as metric learning, has attracted great interest since its introduction for classification (Hastie & Tibshirani, 1996) and clustering (Xing et al., 2002). A global Mahalanobis distance is commonly used to obtain the best distance for discriminating two classes (Xing et al., 2002; Weinberger & Saul, 2009). While a global metric is often the focus of earlier works, multiple local metrics (Frome et al., 2007; Weinberger & Saul, 2009; Ramanan & Baker, 2011; Chen et al., 2019) are found to be useful because they better capture the data space geometry. There is a great body of work on distance metric learning; see, e.g., Bellet et al. (2015); Suárez et al. (2021) for recent reviews.

Metric learning is intimately connected with learning on Riemannian manifolds. Hauberg et al. (2012) connects multi-metric learning to learning the geometry structure of a Riemannian manifold, and advocates its benefits in regression and dimensional reduction tasks. Lebanon (2002; 2006); Le & Cuturi (2015) discuss Riemannian metric learning by utilizing a parametric family of metric, and demonstrate applications in text and image classification. Like in these works, our target is to learn the *Riemannian metric* instead of the *distance metric*, which fundamentally differentiates our approach from most existing works in metric learning. We focus on the nonparametric estimation of the data geometry as quantified by the Riemannian metric tensor. Contrary to distance metric learning, where the coefficient matrix for the Mahalanobis distance is constant in a neighborhood, the Riemannian metric is a smooth tensor field that allows analysis for finer structure. Our emphasis is in inference, namely learning how differences in the response measure are explained by specific differences in the predictor coordinates, rather than obtaining a metric optimal for a supervised learning task.

A related field is manifold learning which attempts to find low-dimensional nonlinear representations of apparent high-dimensional data sampled from an underlying manifold (Roweis & Saul, 2000; Tenenbaum et al., 2000; Coifman & Lafon, 2006). Those embedding methods generally start by assuming the local geometry is given, e.g., by the Euclidean distance between ambient data

points. Not all existing methods are isometric, so the geometry obtained this way can be distorted. Perraul-Joncas & Meila (2013) uses the Laplacian operator to obtain pushforward metric for the low-dimensional representations. Instead of specifying the geometry of the ambient space, we learn the geometry from noisy measures of intrinsic distances. Fefferman et al. (2020) discusses an abstract setting of this task, while our work proposes a practical estimation of the Riemannian metric tensor when coordinates are also available, and we show that our approach is numerically sound.

We suppose that data are generated from an unknown Riemannian manifold, and we have available the coordinates of the data objects. The Euclidean distance between the coordinates may not reflect the underlying geometry. Instead, we assume that we further observe similarity measures between objects, modeled as noise-contaminated intrinsic distances, that are used to characterize the intrinsic geometry on the Riemannian manifold. The targeted Riemannian metric is estimated in a data-driven fashion, which enables estimating geodesics (straight lines and locally shortest paths) and performing calculus on the manifold.

To formulate the problem, let $(\mathcal{M}, G)$ be a Riemannian manifold with Riemannian metric G, and $dist\,(\cdot,\cdot)$ be the geodesic distance induced by G which measures the true intrinsic difference between points. The coordinates of data points $x_0, x_1 \in \mathcal{M}$ are assumed known, identifying each point via a tuple of real numbers. Also observed are noisy measurements y of the intrinsic distance between data points, which we refer to as *similarity measurements* (equivalently dissimilarity). The response is modeled flexibly, and we consider the following common scenarios: (i) noisy distance, where $y = dist\,(x_0, x_1)^2 + \epsilon$ for error ϵ, (ii) similarity/dissimilarity, where $y = 0$ if the two points x_0, x_1 are considered similar and $y = 1$ otherwise, and (iii) relative comparison, where a triplet of points (x_0, x_1, x_2) are given and $y = 1$ if x_0 is more similar to x_1 than to x_2 and $y = 0$ otherwise. The binary similarity measurement is common in computer vision (e.g. Chopra et al., 2005), while the relative comparison could be useful for perceptional tasks and recommendation system (e.g. Schultz & Joachims, 2003; Berenzweig et al., 2004). We aim to estimate the Riemannian metric G and its derivatives using the coordinates and similarity measures among the data points.

The major contribution of this paper is threefold. First, we formulate a framework for probabilistic modeling of similarity measurements among data on manifold via intrinsic distances. Based on a Taylor expansion for the spread of geodesic curves in differential geometry, the local regression procedure successfully estimates the Riemannian metric and its derivatives. Second, a theoretical foundation is developed for the proposed method including asymptotic consistency. Last and most importantly, the proposed method provides a geometric interpretation for the structure of the data space induced by the similarity measurements, as demonstrated in the numerical examples.

2 Background in Riemannian Geometry

For brevity, *metric* now refers to *Riemannian metric* while *distance metric* is always spelled out. Throughout the paper, $\mathcal{M}$ denotes a d-dimensional manifold endowed with a coordinate chart (U, φ), where $\varphi : U \to \mathbb{R}^d$ maps a point $p \in U \subset \mathcal{M}$ on the manifold to its coordinate $\varphi(p) = \left(\varphi^1(p), \ldots, \varphi^d(p)\right) \in \mathbb{R}^d$. Without loss of generality, we identify a point by its coordinate as $\left(p^1, \ldots, p^d\right)$, suppressing φ for the coordinate chart. Upper-script Roman letters denote the components of a coordinate, e.g., p^i is the i-th entry in the coordinate of the point p, and γ^i is the i-th component function of a curve $\gamma : \mathbb{R} \supset [a, b] \to \mathcal{M}$ when expressed on chart U. The tangent space $T_p\mathcal{M}$ is a vector space consisting of velocities of the form $v = \gamma'(0)$ where γ is any curve satisfying $\gamma(0) = p$. The coordinate chart induces a basis on the tangent space $T_p\mathcal{M}$, as $\partial_i|_p = \partial/\partial x^i|_p$ for $i = 1, \ldots, d$, so that a tangent vector $v \in T_p\mathcal{M}$ is represented as $v = \sum_{i=1}^d v^i \partial_i$ for some $v^i \in \mathbb{R}$, suppressing the subscript p in the basis. We adopt the Einstein summation convention unless otherwise specified, namely $v^i\partial_i$ denotes $\sum_{i=1}^d v^i \partial_i$, where common pairs of upper- and lower-indices denotes a summation from 1 to d (see e.g., Lee, 2013, pp.18–19).

The Riemannian metric G on a d-dimensional manifold $\mathcal{M}$ is a smooth tensor field acting on the tangent vectors. At any $p \in \mathcal{M}$, $G(p) : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ is a symmetric bi-linear tensor/function satisfying $G(p)(v, v) \geq 0$ for any $v \in T_p\mathcal{M}$ and $G(p)(v, v) = 0$ if and only if $v = 0$. On a chart φ, the metric is represented as a d-by-d positive definite matrix that quantifies the distance traveled along infinitesimal changes in the coordinates. With an abuse of notation, the chart representation of G is given by the matrix-valued function $p \mapsto G(p) = [G_{ij}(p)]_{i,j=1}^d \in \mathbb{R}^{d\times d}$ for $p \in \mathcal{M}$, so

the distance traveled by γ at time t for a duration of dt is $[G_{ij}(\gamma(t))\dot{\gamma}^i(t)\dot{\gamma}^j(t)]^{1/2}$. The intrinsic distance induced by G, or the *geodesic distance*, is computed as

$$dist\,(p,q) = \inf_{\alpha} \int_0^1 \sqrt{\sum_{1\le i,j\le d} G_{ij}(\alpha(t))\dot{\alpha}^i(t)\dot{\alpha}^j(t)}dt, \tag{2.1}$$

for two points p, q on the manifold $\mathcal{M}$, where infimum is taken over any curve $\alpha : [0,1] \to \mathcal{M}$ connecting p to q.

A *geodesic curve* (or simply *geodesic*) is a smooth curve $\gamma : \mathbb{R} \supset [a,b] \to \mathcal{M}$ satisfying the *geodesic equations*, represented on a coordinate chart as

$$\ddot{\gamma^k}(t) + \dot{\gamma^i}(t)\dot{\gamma^j}(t)\Gamma^k_{ij} \circ \gamma(t) = 0, \text{ for } i,j,k = 1,\ldots,d, \tag{2.2}$$

where over-dots represent derivative w.r.t. t; $\Gamma^k_{ij} = \frac{1}{2}G^{kl}\left(\partial_i G_{jl} + \partial_j G_{il} - \partial_l G_{ij}\right)$ are the *Christoffel symbols* at p; and G^{kl} is the (k,l)-element of G^{-1}. Solving (2.2) with initial conditions $\gamma^i(0) = p^i$, $\dot{\gamma^i}(0) = v^i, i = 1,\ldots,d$ leads to the geodesic starting from p with initial velocity $v = v^i\partial_i \in T_p\mathcal{M}$. A geodesic traces out the generalization of a straight line on the manifold, preserving travel direction with no acceleration, and is also locally the shortest path.

Considering the shortest path γ connecting p to q and applying Taylor's expansion at $t = 0$, we obtain $dist\,(p,q)^2 \approx \sum_{1\le i,j\le d} G_{ij}(p)(q^i - p^i)(q^j - p^j)$, showing the connection between the geodesic distance and a quadratic form analogous to the Mahalanobis distance. Our estimation method is based on this approximation, and we will discuss the higher-order terms shortly which unveil finer structure of the manifold.

3 Local Regression for Similarity Measurements

3.1 Probabilistic Modeling for Similarity Measurements

Suppose that we observe N independent triplets (Y_u, X_{u0}, X_{u1}), $u = 1,\ldots,N$. Here, the X_{uj} are locations on the manifold identified with their coordinates $\left(X^1_{uj},\ldots,X^d_{uj}\right) \in \mathbb{R}^d$, $j = 1,2$, and Y_u are noisy similarity measures of the proximity of (X_{u0}, X_{u1}) in terms of the intrinsic geodesic distance $dist\,(\cdot,\cdot)$ on $\mathcal{M}$. To account for different structures of the similarity measurements, we model the response in a fashion analogous to generalized linear models. For X_{u0}, X_{u1} lying in a small neighborhood $\mathcal{U}_p \subset \mathcal{M}$ of a target location $p \in \mathcal{M}$, the similarity measure Y_u is modeled as

$$\mathbb{E}\left(Y_u | X_{u0}, X_{u1}\right) = g^{-1}\left(dist\,(X_{u0}, X_{u1})^2\right), \tag{3.1}$$

where g is a given link function that relates the conditional expectation to the squared distance.

Example 3.1. We describe below three common scenarios modeled by the general framework (3.1).

1. Continuous response being the squared geodesic distance contaminated with noise:

$$Y_u = dist\,(X_{u0}, X_{u1})^2 + \sigma(p)\varepsilon_n, \tag{3.2}$$

   where $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. mean zero random variables, and $\sigma : \mathcal{M} \to \mathbb{R}^+$ is a smooth positive function determining the magnitude of noise near the target point p. This model will be applied to model trip time as noisy measure of cost to travel between locations.

2. Binary (dis)similarity response:

$$\mathbb{P}\left(Y_u = 1 | X_{u0}, X_{u1}\right) = \text{logit}^{-1}\left(dist\,(X_{u0}, X_{u1})^2 - \hbar\,(p)\right) \tag{3.3}$$

   for some smooth function $\hbar : \mathcal{M} \to \mathbb{R}$, where $\text{logit}(\mu) = \log\left(\mu/\left(1-\mu\right)\right)$, $\mu \in (0,1)$ is the logit function. This models the scenario when there are latent labels for X_{uj} (e.g., digit 6 v.s. 9) and Y_u is a measure of whether the labels for X_{u0}, X_{u1} are in common or not. The function $\hbar\,(p)$ in (3.3) describes the homogeneity of the latent labels for points in a small neighborhood of p. The latent labels could have intrinsic variation even if measurements are made for the same data points $x = X_{u0} = X_{u1}$, and the strength of which is captured by $\hbar\,(p)$.

3. Binary relative comparison response, where we extend our model for triplets of points (X_{u0}, X_{u1}, X_{u2}), where Y_u stands for whether X_{u0} is more similar to X_{u1} than to X_{u2}:

$$\mathbb{P}\left(Y_u = 1 | X_{u0}, X_{u1}, X_{u2}\right) = \text{logit}^{-1}\left(dist\left(X_{u0}, X_{u2}\right)^2 - dist\left(X_{u0}, X_{u1}\right)^2\right), \quad (3.4)$$

so that the relative comparison Y_u reflects the comparison of squared distances.

3.2 Local Approximation of Squared Distances

We develop a local approximation for the squared distance as the key tool to estimate our model (3.1) through local regression. Proposition 3.1 provides a Taylor expansion for the squared geodesic distance between two geodesics with same starting point but different initial velocities (see Figure C.1 for visualization of the spread). For a point p on the Riemannian manifold $\mathcal{M}$, let $\exp_p : T_p\mathcal{M} \to \mathcal{M}$ denote the exponential map defined by $\exp_p(tv) = \gamma(t)$ where γ is a geodesic starting from p at time 0 with initial velocity $\gamma'(0) = v \in T_p\mathcal{M}$. For notational simplicity, we suppress the dependency on p in geometric quantities (e.g., the metric G is understood to be evaluated at p). For $i = 1, \ldots, d$, denote $\delta^i = \delta^i(t) = \gamma^i(t) - \gamma^i(0)$ as the difference in coordinate after a travel of time t along γ.

Proposition 3.1 (spread of geodesics, coordinated). Let $p \in \mathcal{M}$ and $v, w \in T_p\mathcal{M}$ be two tangent vectors at p. On a local coordinate chart, the squared geodesic distance between two geodesics $\gamma_0(t) = \exp_p(tv)$ and $\gamma_1(t) = \exp_p(tw)$ satisfies, as $t \to 0$,

$$dist\left(\gamma_0(t), \gamma_1(t)\right)^2 = \delta^i_{0-1}\delta^j_{0-1}G_{ij} + \delta^i_{0-1}\left(\delta^k_0\delta^l_0 - \delta^k_1\delta^l_1\right)\Gamma^j_{kl}G_{ij} + O(t^4) \quad (3.5)$$

where for $i, j, k, l, m = 1, \ldots, d$,

- $\delta^i_0 = \gamma^i_0(t) - p^i$, $\delta^i_1 = \gamma^i_1(t) - p^i$, and $\delta^i_{0-1} = \delta^i_0 - \delta^i_1$, i.e., δ^i_0, δ^i_1 are differences in i-th coordinates of $\gamma_0(t)$ and $\gamma_1(t)$ to the origin p, respectively, and $\delta^i_{0-1} = \delta^i_0 - \delta^i_1$ is the coordinate difference between $\gamma_0(t)$ and $\gamma_1(t)$;
- G_{ij} and Γ^j_{kl} are the elements of the metric and Christoffel symbols at p, respectively.

Proposition 3.1 is tied to a more fundamental coordinate-invariant result in differential geometry; see Meyer (1989) and Proposition C.1 for detail. Proposition 3.1 inspires the use of local regression to estimate the metric tensor and the Christoffel symbols. For X_{u0}, X_{u1} in a small neighborhood of p, write the linear predictor as

$$\eta_u := \beta^{(0)} + \delta^i_{u,0-1}\delta^j_{u,0-1}\beta^{(1)}_{ij} + \delta^k_{u,0-1}\left(\delta^i_{u0}\delta^j_{u0} - \delta^i_{u1}\delta^j_{u1}\right)\beta^{(2)}_{ijk}, \quad (3.6)$$

a function of the intercept $\beta^{(0)}$ and coefficients $\beta^{(1)}_{ij}, \beta^{(2)}_{ijk}$, where $\delta^i_{u0} = X^i_{u0} - p^i$, $\delta^i_{u1} = X^i_{u1} - p^i$, and $\delta^i_{u,0-1} = \delta^i_{u0} - \delta^i_{u1}$, for $i, j, k, l = 1, \ldots, d$, and $u = 1, \ldots, N$. The intercept term $\beta^{(0)}$ is included for capturing $\hbar(p)$ in (3.3) under that scenario and can otherwise be dropped from the model. The link function connects the linear predictor to the conditional mean via $\mu_u := g^{-1}(\eta_u) \approx \mathbb{E}\left(Y_u | X_{u0}, X_{u1}\right)$ as indicated by (3.1) and (3.5), where μ_u is seen as a function of the coefficients $\beta^{(0)}$, $\beta^{(1)}_{ij}$, and $\beta^{(2)}_{ijk}$. Therefore, upon the specification of a loss function $Q : \mathbb{R} \times \mathbb{R} \to \{0\} \cup \mathbb{R}^+$ and non-negative weights $w_1, \ldots, w_N$, the minimizers

$$(\hat{\beta}^{(0)}, \hat{\beta}^{(1)}_{ij}, \hat{\beta}^{(2)}_{ijk}) = \underset{\beta^{(0)}, \beta^{(1)}_{ij}, \beta^{(2)}_{ijk}; i,j,k}{\arg\min} \sum_{n=1}^{N} Q\left(Y_u,\ \mu_u\right) w_u, \quad (3.7)$$

$$\text{subject to } \beta^{(1)}_{ij} = \beta^{(1)}_{ji}, \quad \beta^{(2)}_{ijk} = \beta^{(2)}_{jik}, \text{ for } i, j, k, l = 1, \ldots, d, \quad (3.8)$$

are used to estimate the metric tensor and Christoffel symbols, obtaining

$$\hat{G}_{ij} = \hat{\beta}^{(1)}_{ij}, \quad \hat{\Gamma}^l_{ij} = \hat{\beta}^{(2)}_{ijk}\hat{G}^{kl}, \quad (3.9)$$

where $\hat{G}^{kl}$ is the matrix inverse of $\hat{G}$ satisfying $\hat{G}^{kl}\hat{G}_{kj} = 1_{\{j=l\}}$. The symmetry constraints (3.8) are the result of the symmetries in the metric tensor and Christoffel symbols, and are enforced by optimizing over only the lower triangular indices $1 \le i < j \le d$ without constraints. Asymptotic results guarantees the positive-definiteness of the metric estimate, as will be shown in Proposition 4.1.

To weigh the pairs of endpoints according to their proximity to the target location p, we apply kernel weights specified by

$$w_u = h^{-2d} \prod_{i=1}^{d} K\left(\frac{X_{u0}^i - p^i}{h}\right) K\left(\frac{X_{u1}^i - p^i}{h}\right) \tag{3.10}$$

for some $h > 0$ and non-negative kernel function K. The bandwidth h controls the bias–variance trade-off of the estimated Riemannian metric tensor and its derivatives.

Altering the link function g and the loss function Q in (3.7) enables flexible local regression estimation for models in Example 3.1.

Example 3.2. Consider the following loss functions for estimating the metric tensors and the Christoffel symbols when data are drawn from model (3.2)–(3.4), respectively.

1. Continuous noisy response: use squared loss $Q(y, \mu) = (y - \mu)^2$ with g being the identity link function so $\mu_u = \eta_u$.

2. Binary (dis)similarity response: use log-likelihood of Bernoulli random variable

$$Q(y, \mu) = y \log \mu + (1 - y) \log (1 - \mu), \tag{3.11}$$

   and g the logit link, so $\mu_u = \text{logit}^{-1}(\eta_u)$. The model becomes a local logistic regression.

3. Binary relative comparison response: apply the same loss function (3.11) and logit link as in the previous scenario, but here we formulate the linear predictor based on $dist(X_{u0}, X_{u2})^2 - dist(X_{u0}, X_{u1})^2 \approx \eta_{u1} - \eta_{u2}$ and

$$\mu_u = g^{-1}(\eta_{u1} - \eta_{u2}). \tag{3.12}$$

   Locally, the difference in squared distances is approximated by

$$\begin{aligned} \eta_{u1} - \eta_{u2} = &\left(\delta_{u,0-1}^i \delta_{u,0-1}^j - \delta_{u,0-2}^i \delta_{u,0-2}^j\right) \beta_{ij}^{(1)} \\ &+ \left(\delta_{u,0-1}^k \left(\delta_{u0}^i \delta_{u0}^j - \delta_{u1}^i \delta_{u1}^j\right) - \delta_{u,0-2}^k \left(\delta_{u0}^i \delta_{u0}^j - \delta_{u2}^i \delta_{u2}^j\right)\right) \beta_{ijk}^{(2)}, \end{aligned} \tag{3.13}$$

   for $\delta_{u2}^i = X_{u2}^i - p^i$ and $\delta_{u,0-2}^i = \delta_{u2}^i - \delta_{u0}^i$, $i = 1, \ldots, d$. Here η_{u1} and η_{u2} are constructed in analogy to (3.6) using (X_{u0}, X_{u1}) and (X_{u0}, X_{u2}) pair respectively.

Examples in Section 5 will further illustrate the proposed method in those scenarios. Besides the models listed, many other choices for the link g and loss function Q are feasible under this local regression framework (Fan & Gijbels, 1996), which accommodates a wide variety of data.

4 Bias and Variance of the Estimated Metric Tensor

This subsection provides asymptotic justification for model (3.2) with $\mathbb{E}(Y_u | X_{u0}, X_{u1}) = dist(X_{u0}, X_{u1})^2$ under the squared loss $Q(\mu, y) = (\mu - y)^2$ and the identity link $g(\mu) = \mu$. The estimator we analyzed here fits a local quadratic regression without intercept and approximates the squared distance by a simplified form of (3.6):

$$dist(X_{u0}, X_{u1})^2 \approx \eta_u := \delta_{u,0-1}^i \delta_{u,0-1}^j \beta_{ij}^{(1)}, \tag{4.1}$$

for $u = 1, \ldots, N$. Given a suitable order of the indices i, j for vectorization, we rewrite the formulation into a matrix form. Denote the local design matrix and regression coefficients as

$$\mathbf{D}_u = \left(\delta_{u,0-1}^1 \delta_{u,0-1}^1, \ \ldots, \ \delta_{u,0-1}^i \delta_{u,0-1}^j, \ \ldots, \ \delta_{u,0-1}^d \delta_{u,0-1}^d\right)^T, \quad \boldsymbol{\beta} = \left(\beta_{11}^{(1)}, \ldots, \beta_{ij}^{(1)}, \ldots, \beta_{dd}^{(1)}\right)^T,$$

so that the linear predictor $\eta_u = \mathbf{D}_u^T \boldsymbol{\beta}$. Further, write

$$\mathbf{D} = (\mathbf{D}_1, \ldots, \mathbf{D}_N)^T, \quad \boldsymbol{Y} = (Y_1, \ldots, Y_N)^T, \quad \mathbf{W} = \text{diag}(w_1, \ldots, w_N),$$

with weights w_u specified in (3.10). The objective function in (3.7) becomes

$$(\boldsymbol{Y} - \mathbf{D}\boldsymbol{\beta})^T \mathbf{W} (\boldsymbol{Y} - \mathbf{D}\boldsymbol{\beta}), \tag{4.2}$$

for which the minimizer is $\hat{\boldsymbol{\beta}} = \left(\mathbf{D}^T\mathbf{W}\mathbf{D}\right)^{-1}\mathbf{D}^T\mathbf{W}\boldsymbol{Y}$. Therefore

$$\text{bias}\left(\hat{\boldsymbol{\beta}}|\mathbf{D}\right) = \left(\mathbf{D}^T\mathbf{W}\mathbf{D}\right)^{-1}\mathbf{D}^T\mathbf{W}\boldsymbol{r}, \tag{4.3}$$

$$\text{var}\left(\hat{\boldsymbol{\beta}}|\mathbf{D}\right) = \left(\mathbf{D}^T\mathbf{W}\mathbf{D}\right)^{-1}\mathbf{D}^T\boldsymbol{\Sigma}\mathbf{D}\left(\mathbf{D}^T\mathbf{W}\mathbf{D}\right)^{-1}, \tag{4.4}$$

where

$$\boldsymbol{r} = \left(\mathbb{E}\left(Y_u|X_{u0}, X_{u1}\right) - \delta^i_{u,0-1}\delta^j_{u,0-1}G_{ij}\right)_{1\le u\le N}, \quad \boldsymbol{\Sigma} = \text{diag}\left(w_u^2\,\text{var}\left(Y_u|X_{u0}, X_{u1}\right)\right)_{1\le u\le N}.$$

To characterize the asymptotic bias and variance of the estimator, we assume the following conditions are satisfied in a neighborhood of the target p. These conditions are standard and analogous to those assumed in a local regression setting (Fan & Gijbels, 1996).

(A1) The joint density of endpoints (X_{u0}, X_{u1}) is positive and continuously differentiable.

(A2) The functions G_{ij}, Γ^k_{ij} are C^2-smooth for $i, j, k = 1, \ldots, d$.

(A3) The kernel K in weights (3.10) is symmetric, continuous, and has a bounded support.

(A4) $\sup_u \text{var}\left(Y_u|X_{u0}, X_{u1}\right) < \infty$.

Proposition 4.1. Under (A1)–(A4), $\text{bias}\left(\hat{\boldsymbol{\beta}}|\mathbf{X}\right) = O_p\left(h^2\right)$, $\text{var}\left(\hat{\boldsymbol{\beta}}|\mathbf{X}\right) = O_p\left(N^{-1}h^{-4-2d}\right)$, as $h \to 0$ and $Nh^{2+2d} \to \infty$, where $\mathbf{X} = \{(X_{u0}, X_{u1})\}_{u=1}^N$ is the collection of observed endpoints.

The local approximation (4.1) is similar to a local polynomial estimation of the second derivative of a $2d$-variate squared geodesic distance function, explaining the order of h in the rates for bias and variance.

5 Simulation

We illustrate the proposed method using simulated data with different types of responses as described in Example 3.1. Subsection 5.1 investigates the continuous (3.2) and the binary response (3.3) models with data generated on a unit sphere, and Subsection 5.2 focuses on the relative comparison model (3.4). We study whether the proposed method well estimates Riemannian geometric quantities, including the metric tensor, geodesics, and Christoffel symbols. Additional details are included in Appendix B of the Supplementary Materials.

5.1 Unit Sphere

We endow the d-dimensional unit sphere with the round metric, which induces the usual arc-length/great circle distance on the sphere. The coordinates are expressed under the stereographic projection coordinate $\left(x^1, \ldots, x^d\right)$, under which the round metric is written as

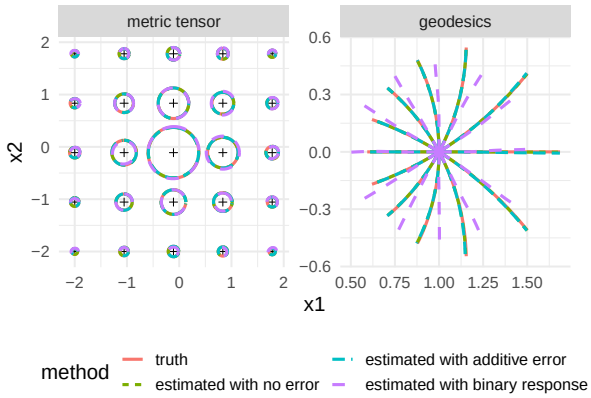
$$\mathring{G}_{ij} = 4\left(1 + \sum_{k=1}^d x^k x^k\right)^{-2} 1_{\{i=j\}}, \tag{5.1}$$

for $i, j = 1, \ldots, d$. Under the additive model (3.2) in Example 3.1, we considered either noiseless responses from (3.2) corresponding to setting $\sigma(p) = 0$, or noisy responses from (3.2) with $\sigma(p) > 0$. For simplicity of presentation, the binary similarity response model (3.3) considered the distance $dist_c(\cdot,\cdot) = \sqrt{c}\,dist(\cdot,\cdot)$ induced by the *scaled metric* $G_{ij,c} = c\mathring{G}_{ij}$ for some constant c and $i, j = 1, \ldots, d$. Intuitively, the constant c regulates the signal-to-noise ratio without changing the form of geodesics. Given the endpoints, a smaller c leads to a smaller value of geodesic distance and hence smaller variation in the linear predictors η_u, so the response Y_u will take less influence from the distance, representing a higher amount of noise.

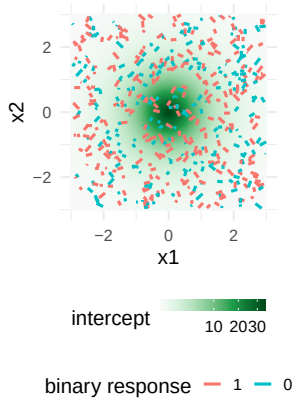
Experiments were preformed on $d = 2$-dimensional sphere and the finding are summarized in Figure 5.1. For continuous responses, the left panel of Figure 5.1a visualizes the true and estimated metric tensors via cost ellipses (A.1) and the right panel shows the corresponding geodesics by solving the geodesic equations (2.2) with true and estimated Christoffel symbols. The metrics and the geodesics were well estimated under the continuous response model without or without additive

errors, where the estimates overlap with the truth. Figure 5.1b evaluates the relative estimation errors $\left\|\hat{G} - G\right\|_F / \|G\|_F$ and $\left\|\hat{\Gamma} - \Gamma\right\|_F / \|\Gamma\|_F$ of the metric and the Christoffel symbol w.r.t. the Frobenius norm (A.2) for data from the continuous model (3.2).

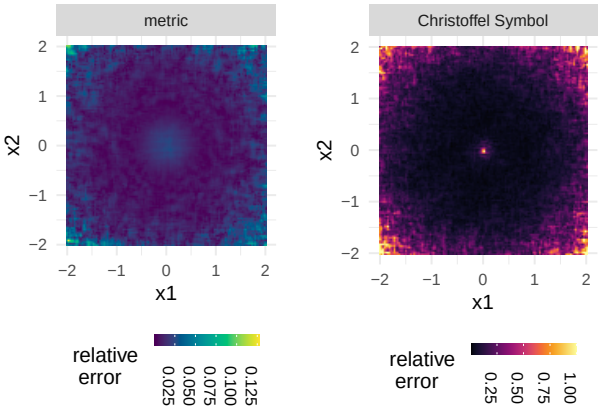
For binary responses under model (3.3), a subset of data is visualized in Figure 5.1c where the background color illustrates $\hbar$. Figure 5.1d and left panel of Figure 5.1a suggest that the intercept and the metric were reasonably estimated, while the geodesics are slightly away from the truth (Figure 5.1a, right). This indicates that the binary model has higher complexity and less information is being provided by the binary response (see also Figure B.1b in the Supplementary Materials).



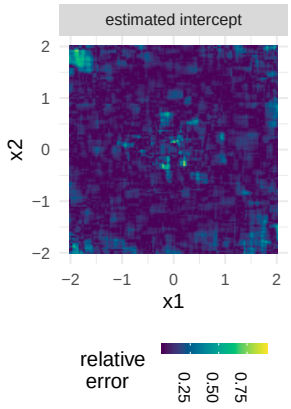
(a) Estimated and true metric tensors using ellipses representation (left) and the geodesic curves (right) starting from $(1, 0)$ with unit initial velocities pointing to 1–12 o'clock directions.



(c) Simulated data, where line segments show pairs of endpoints colored according to their binary responses. The background shows the value of $\hbar$.



(b) Relative errors in term of Frobenius norm of the estimated tensors for the continuous response model (3.2) with additive error.



(d) Errors for estimating $\hbar$ with binary responses (3.3).

Figure 5.1: Simulation results for 2-dimensional sphere under stereographic projection coordinate.

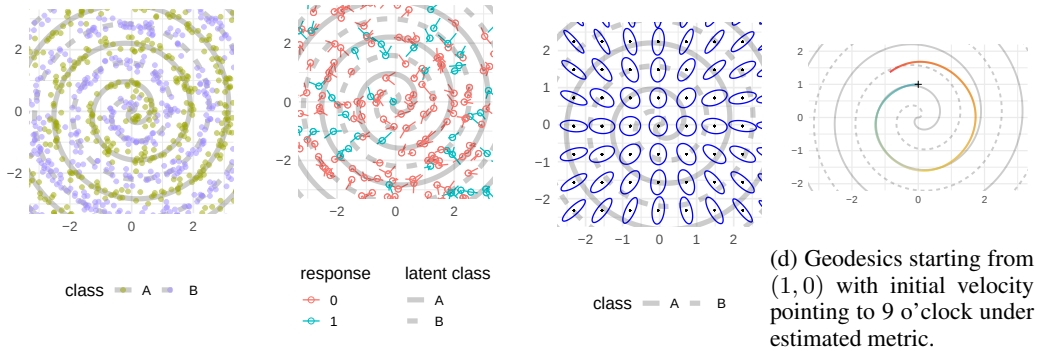
5.2 Relative Comparison on the Double Spirals

A set of $n = 7 \times 10^4$ locations $\mathcal{X} = \{X_i\}_{i=1}^n$ were generated on two spirals, corresponding to two latent classes $\mathcal{A}$ and $\mathcal{B}$. We suppose that each datum belongs to one of these latent class and lies close to the corresponding spiral (e.g., the set of green points in Figure 5.2a are have latent class $\mathcal{A}$). A total of $N = 6965312$ comparison responses were formed, where for $u = 1, \ldots, N$, Y_u is formed by comparing nearby points in a triplet (X_{u0}, X_{u1}, X_{u2}) of points in $\mathcal{X}$, visualized as wedged in Figure 5.2b, where the hollow circles in the middle correspond to X_{u0}. The relative comparisons

were generated according to $Y_u = 1$ if X_{u0}, X_{u1} belong to the same latent class and X_{u0}, X_{u2} belong to different classes; otherwise $Y = 0$.

In this example, the intrinsic distance is induced by the two latent classes, in that distances across the supports of the two classes tend to be large, and distances within the support of a single class are small. Therefore, the resulting metric tensor should reflect less cost while moving along the tangential direction of the spirals compared to perpendicular directions. We applied model (3.4) to describe the underlying geometry. We obtained the proposed estimates by minimizing the objective (3.7) with the link function (3.12) and the local approximation (3.13).

The estimated metric shown in Figure 5.2c is consistent with the interpretation of the intrinsic distance and metric induced by the class membership discussed above. Meanwhile, the estimated geodesic curve unveils the hidden circular structure of the data support as shown in Figure 5.2d. The proposed method successfully unveils geometric structures despite fitting a misspecified model to data not drawn from model (3.4).



(a) The points and their latent classes

(b) Relative comparison

(c) Estimated metric

(d) Geodesics starting from $(1, 0)$ with initial velocity pointing to 9 o'clock under estimated metric.

Figure 5.2: Simulation results for relative comparison on double spirals. Gray curves (solid and dashed) in the background represent the approximate support of the two latent classes. In (b), the tiny circles are X_{u0}, each with two segments connecting to X_{u1} and X_{u2}, colored according to Y_u.

6 New York City Taxi Trip Duration

Next, we study the geometry induced by taxi travel time in New York City (NYC) during weekday morning rush hours. New York City Taxi and Limousine Commission (TLC) Trip Record Data was accessed on May 1st, 2022[1]. We obtained trip records for yellow and green taxis, which include times and GPS coordinates for pickup and dropoff, trip distances, and trip duration. To characterize the travel time in the city, the squared trip duration (in squared seconds) is used as the response variable. We focus on 8,809,982 records between 7 a.m. to 10 a.m. on business days from May to September of 2015 in New York City areas other than the Staten Island.

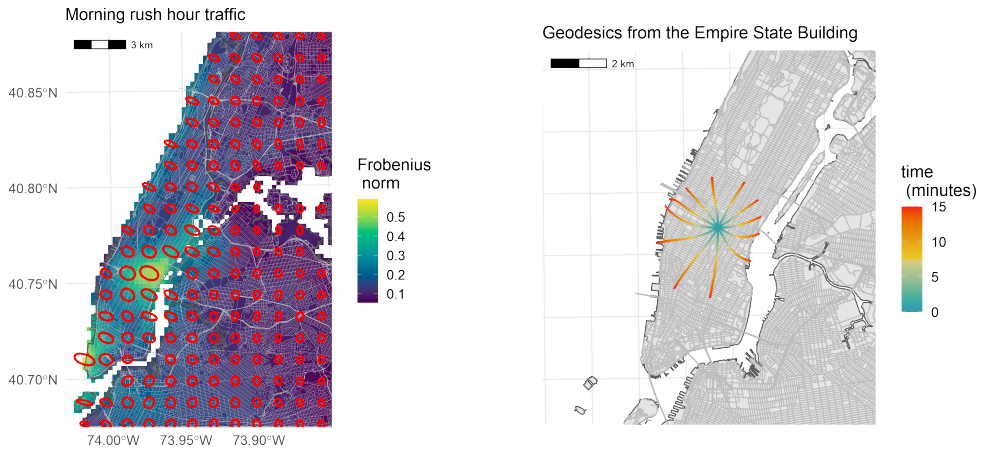
We measure the cost to travel Y_u by the squared trip duration (instead of the trip distance). Let X_{u0} and X_{u1} be the pickup and dropoff location, respectively. Targeting the continuous response model (3.2), we minimize the local objective (3.7) with $Q(y, \mu) = (y - \mu)^2$ and $g(\mu) = \mu$. For each target location p, estimation was computed using trips among the $M \leq 5 \times 10^4$ closest pickup/dropoff endpoints in the neighborhood $\mathcal{U}_p = \{(x^1, x^2) : |x^i - p^i| \leq 5 \text{ kilometers for } i = 1, 2\}$, and weights given by $h = 2.5$ kilometers and kernel K being the standard normal distribution.

Figure 6.1a shows the estimated metric for taxi travel time. The background color shows the Frobenius norm of the metric tensor, where larger values mean that longer travel time is required to pass through that location. Trips through midtown Manhattan and the financial district were estimated to be the most costly during rush hours, which is coherent to the fact that these are the city's primary

[1] https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page, though data format changed after we downloaded them.

business districts. Moreover, the cost ellipses represent the cost in time to travel a unit distance along different directions. This suggests that in Manhattan, it takes longer to drive along the east–west direction (which are generally narrower streets) compared to the north–south direction (wider avenues).

Geodesic curves in Figure 6.1b show where a 15-minutes taxi ride leads to starting from the Empire State Building. Each geodesic curve corresponds to one of 12 starting directions (1–12 o'clock). Note that we apply a continuous Riemannian manifold approximation to the city, so the geodesic curves provide approximations to the shortest paths between locations and do not necessarily conform to the road network. Travel appears to be faster in lower Manhattan than in midtown Manhattan. The spread of the geodesics differs along different directions, indicating the existence of non-constant curvature on the manifold and advocating for estimating the Riemannian metric tensor field instead of applying a single global distance metric.



(a) Estimated metric tensors for trip duration: cost ellipses and Frobenius norm (background color).

(b) Geodesics correspond to 15-minute taxi rides from the Empire State Building heading to 1–12 o'clock.

Figure 6.1: New York taxi travel time during morning rush hours.

7 Discussion

To further exploit the potential of the proposed method and the Riemannian geometry structure, proper translation of mathematical concepts such as parallelism, covariant derivative, and geodesic into the application context is crucial. Serious attempts are necessary to align and connect those concepts to the application community, such as cognition and perception research, where similarity measurements are common.

REFERENCES

Aurélien Bellet, Amaury Habrard, and Marc Sebban. Metric learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 9(1):1–151, January 2015. ISSN 1939-4608. doi: 10.2200/S00626ED1V01Y201501AIM030.

Adam Berenzweig, Beth Logan, Daniel P.W. Ellis, and Brian Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, 28(2):63–76, June 2004. ISSN 0148-9267. doi: 10.1162/014892604323112257.

Shuo Chen, Lei Luo, Jian Yang, Chen Gong, Jun Li, and Heng Huang. Curvilinear distance metric learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 539–546 vol. 1, June 2005. doi: 10.1109/CVPR.2005.202.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, July 2006. ISSN 10635203. doi: 10.1016/j.acha.2006.04.006.

Manfredo Perdigão do Carmo. *Riemannian Geometry*. Mathematics. Theory & Applications. Birkhäuser, Boston, 1992. ISBN 978-0-8176-3490-2 978-3-7643-3490-1.

Jianqing Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Number 66 in Monographs on Statistics and Applied Probability. Chapman & Hall, London ; New York, 1st ed edition, 1996. ISBN 978-0-412-98321-4.

Charles Fefferman, Sergei Ivanov, Matti Lassas, and Hariharan Narayanan. Reconstruction of a Riemannian manifold from noisy intrinsic distances. *SIAM Journal on Mathematics of Data Science*, 2(3):770–808, September 2020. doi: 10.1137/19M126829X.

Andrea Frome, Yoram Singer, Fei Sha, and Jitendra Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, October 2007. doi: 10.1109/ICCV.2007.4408839.

T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):607–616, June 1996. ISSN 01628828. doi: 10.1109/34.506411.

Søren Hauberg, Oren Freifeld, and Michael J. Black. A geometric take on metric learning. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, pp. 2024–2032. Curran Associates, Inc., 2012.

Serge Lang. *Fundamentals of Differential Geometry*. Graduate Texts in Mathematics. Springer-Verlag, New York, 1999. ISBN 978-0-387-98593-0. doi: 10.1007/978-1-4612-0541-8.

Tam Le and Marco Cuturi. Unsupervised Riemannian metric learning for histograms using Aitchison transformations. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2002–2011. PMLR, June 2015.

G. Lebanon. Metric learning for text documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):497–508, April 2006. ISSN 0162-8828. doi: 10.1109/TPAMI.2006.77.

Guy Lebanon. Learning Riemannian metrics. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, UAI'03, pp. 362–369, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 978-0-12-705664-7.

John M. Lee. *Introduction to Smooth Manifolds*. Number 218 in Graduate Texts in Mathematics. Springer, New York ; London, 2nd ed edition, 2013. ISBN 978-1-4419-9981-8 978-1-4419-9982-5.

John M. Lee. *Introduction to Riemannian Manifolds*. Springer Berlin Heidelberg, New York, NY, 2018. ISBN 978-3-319-91754-2.

Wolfgang Meyer. Toponogov's theorem and applications. *Lecture Notes, Trieste*, 1989.

Dominique Perraul-Joncas and Marina Meila. Non-linear dimensionality reduction: Riemannian metric estimation and the problem of geometric discovery. *arXiv:1305.7255 [stat]*, May 2013.

Deva Ramanan and Simon Baker. Local distance functions: A taxonomy, new algorithms, and an evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):794–806, April 2011. ISSN 1939-3539. doi: 10.1109/TPAMI.2010.127.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, December 2000. doi: 10.1126/science.290.5500.2323.

Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.

Juan Luis Suárez, Salvador García, and Francisco Herrera. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. *Neurocomputing*, 425:300–322, February 2021. ISSN 0925-2312. doi: 10.1016/j.neucom.2020.08.017.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000. doi: 10.1126/science.290.5500.2319.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. In S. Becker, S. Thrun, and K. Obermayer (eds.), *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

Supplementary Materials

CONTENTS

1 Introduction **1**

2 Background in Riemannian Geometry **2**

3 Local Regression for Similarity Measurements **3**

3.1 Probabilistic Modeling for Similarity Measurements 3

3.2 Local Approximation of Squared Distances 4

4 Bias and Variance of the Estimated Metric Tensor **5**

5 Simulation **6**

5.1 Unit Sphere 6

5.2 Relative Comparison on the Double Spirals 7

6 New York City Taxi Trip Duration **8**

7 Discussion **9**

A Additional Definition **12**

B Additional Simulation Details **13**

B.1 Unit Sphere and Round Metric 13

B.1.1 Bandwidth Selection 13

B.2 The Double Spirals 14

C Spread of Geodesics **16**

C.1 Proofs 16

D Asymptotic of the Estimated Metric Tensor **19**

A ADDITIONAL DEFINITION

A cost ellipse visualizes the metric by an ellipse

$$\mathscr{E}_p = \left\{ \left(x^1, \ldots, x^d\right) : \sum_{i,j=1}^{d} \left(x^i - p^i\right)\left(x^j - p^j\right) G^{ij} = r^2 \right\} \tag{A.1}$$

for some constant $r > 0$, which shows, approximately, the intrinsic distance on the manifold when traveling a unit length on the coordinate chart along each direction. More precisely, it shows the norm of tangent vectors $v^i \partial_i \in T_p\mathcal{M}$ subject to $\sum_{i=1}^{d} v^i v^i = r^2$ at p pointing to the corresponding direction. For example in a $d = 2$-dimensional manifold at $p = 0$, under $G = \operatorname{diag}(\lambda_1, \lambda_2)$ with $\lambda_1 > \lambda_2$ and $r = 1$. The long axis, $\left(\pm\sqrt{\lambda_1}, 0\right)$, is the norm of the tangent vector $\pm\partial_1$. Thus, the direction in which the ellipse is larger corresponds to the direction of the larger geodesic

distance. One can see (A.1) as the "inverse" of the Tissot's indicatrix, where the latter shows a local equidistance contour to the ellipse's center.

Frobenius norm of tensors is denoted as $\|\cdot\|_F$, defined as

$$\|G\|_F = \left(\sum_{i,j=1}^{d} G_{ij}G_{ij}\right)^{1/2}, \quad \|\Gamma\|_F = \left(\sum_{i,j,k=1}^{d} \Gamma_{ij}^k \Gamma_{ij}^k\right)^{1/2}, \tag{A.2}$$

for metric tensor G and Christoffel symbol Γ.

B Additional Simulation Details

This section provides further detail to complete Section 5 of the main text including how we generated the simulated data, and more figures.

B.1 Unit Sphere and Round Metric

The stereographic coordinate of the d-dimensional sphere $\mathbb{S}^d$ identifies points on the sphere by mapping it to its stereographic projection in $\mathbb{R}^d$ from the north pole. The round metric on the sphere $\mathbb{S}^d$ is the metric induced by embedding of $\mathbb{S}^d \hookrightarrow \mathbb{R}^{d+1}$. For detail, see for example page 30 of Lee (2013) and chapter 3 of Lee (2018). We generated endpoints X_{u0}, X_{u1} uniformly in the coordinate chart $(-3,3)\times(-3,3)$, then pair the endpoints so that the difference in coordinates of the endpoints $|\delta_{u0}^i - \delta_{u1}^i| = |X_{u0}^i - X_{u1}^i|$ would not exceed 0.2 for $i = 1, 2$, $u = 1, \ldots, N$.

More precisely, data were generated on $d = 2$-dimensional sphere under stereographic projection coordinate. A total of $N = 5 \times 10^5$ pairs of endpoints with $X_{u0}^i, X_{u1}^i \in (-3,3)$ were generated subject to $|X_{u0}^i - X_{u1}^i| \leq 0.2$ for all $u = 1, \ldots, N; i = 1, \ldots, d$. For a reasonable signal-to-noise ratio, we set $\sigma(p) = \sigma \approx 9 \times 10^{-4}$ for all p, which is approximately one-tenth of the marginal expectation of squared distance, i.e., $\sigma \approx \mathbb{E}\, dist\left(X_{u0}, X_{u1}\right)^2/10$. The $\hbar\left(p\right)$ was set to be the average local squared distances within a local neighborhood of p under the scaled metric with $c = 300$. In the end, the responses were generated following (3.2) and (3.3) respectively.

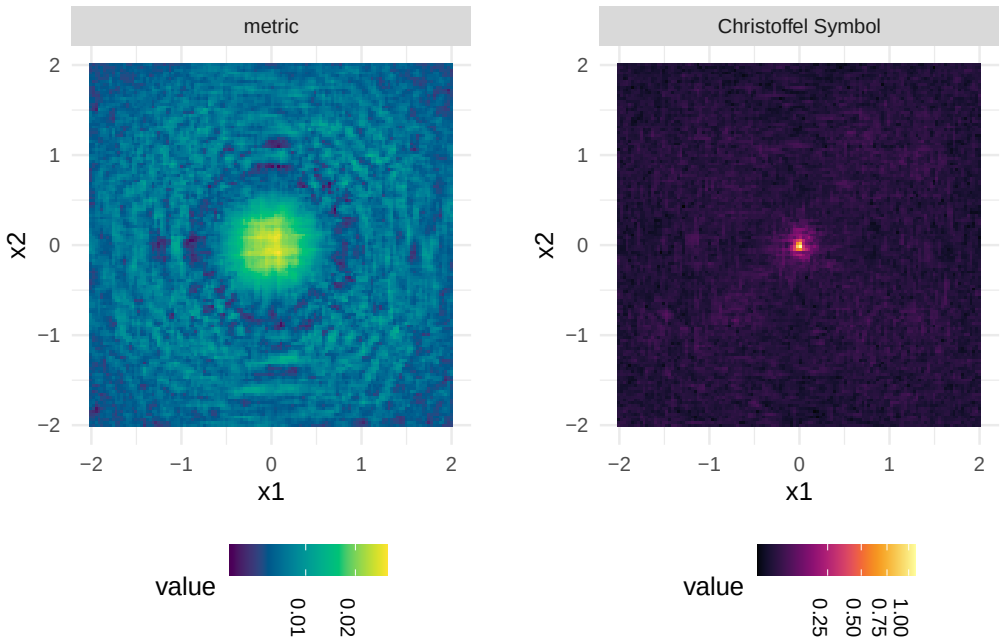
In addition, Figure B.1 illustrates the relative Frobenius error for estimated tensors using noiseless or binary responses.

B.1.1 Bandwidth Selection

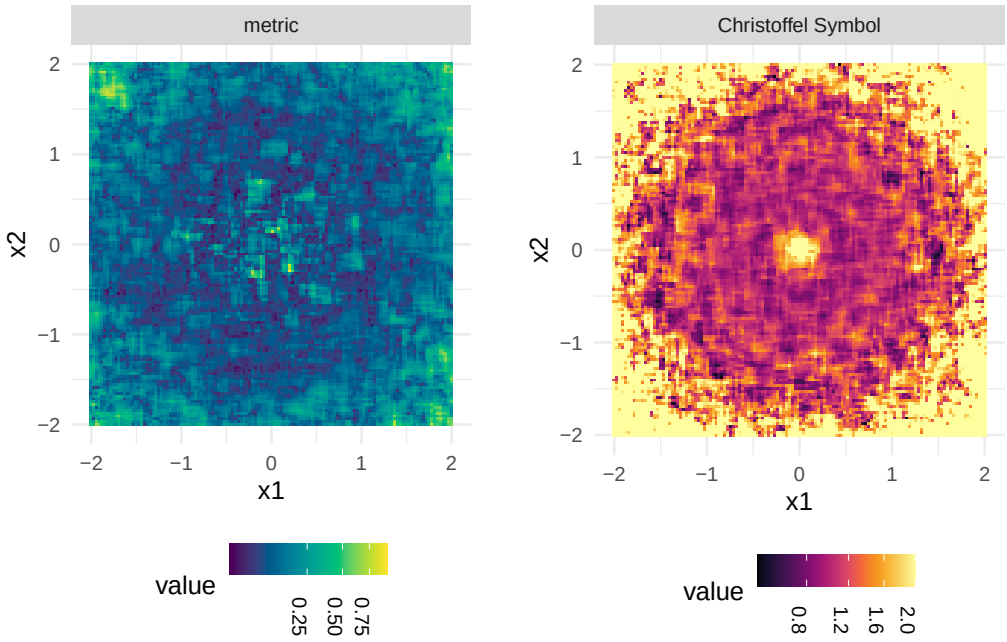
Like local regression, the proposed method relies on a neighborhood specification for optimal bias-variance trade-off. The simulation in Subsection 5.1 uses the rectangular kernel $K(x) = \mathbf{1}_{[-1,1]}(x)$ for (3.10), where $\mathbf{1}$ is the indicator function, so the estimation only utilizes observations with endpoints X_{u0}, X_{u1} are both lying in the neighborhood $\mathcal{U}_p = \left\{\left(x^1, \ldots, x^d\right) : |x^i - p^i| \leq h, i = 1, \ldots, d\right\}$ of the target point p.

We propose a train–test set scheme for data-driven bandwidth selection. To simplify computation, we only considered additive error under (3.2). A 16×16 grid $p_1, \ldots, p_{256} \in (-3,3) \times (-3,3)$ were used as target points where metric tensors were estimated, with a test set of $N_{\text{test}} = 31246$ observations that were within close proximity to the grid. Estimation of the tensors were computed w.r.t. bandwidth h utilizing a train set containing $N_{\text{train}} = 400158$ (approximately 80% of the data) randomly selected observations outside of the test set. For the test set, $\hat{Y}_{u,\text{test}} = \hat{\eta}_{u,\text{test}}$ were then computed under identity link by plugging the estimated tensors into (3.6).

The bandwidth minimizing the squared loss $\sum_{u=1}^{N_{\text{test}}} \left(Y_{u,\text{test}} - \hat{Y}_{u,\text{test}}\right)^2 / N_{\text{test}}$ is then chosen. The proposed bandwidth selection resulted in $h = 0.18$ according to the loss shown in the left panel of Figure B.2, which corresponds to an effectively local sample size of approximately 1000 in the training step (right panel of Figure B.2). These tuning parameter choices were applied in the results shown in Subsection 5.1. Other bandwidth selection methods developed for local regression (see e.g. Fan & Gijbels, 1996, section 4.10) can also be adopted here.



(a) Noiseless



(b) Binary response

Figure B.1: Relative errors w.r.t. Frobenius norm (A.2) of the estimated tensors with noiseless or binary response for 2-dimensional sphere under stereographic projection coordinate chart.

B.2 The Double Spirals

Define a class of spiral functions as $\mathcal{S} : \mathbb{R} \to \mathbb{R}^2$ with

$$t \mapsto (\cos(5t + \phi),\ t\sin(5t + \phi)) .$$

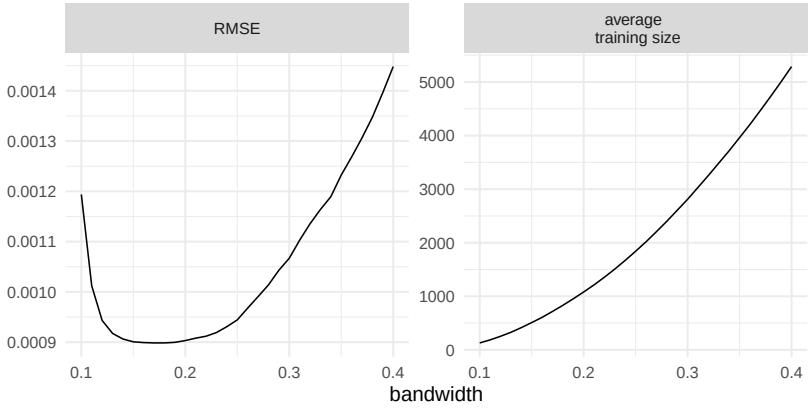


Figure B.2: Root mean squared error (RMSE) for the test set (Left) and the average number of local training observations (Right) as the bandwidth h varies.

The underlying spirals for class $\mathcal{A}$ and $\mathcal{B}$ are $\mathcal{S}_A = \mathcal{S}(\cdot, \phi = 0)$ and $\mathcal{S}_B = \mathcal{S}(\cdot, \phi = \pi)$ respectively. For endpoints, we first generate points $\mathring{X}_m$ independently and uniformly on $\mathcal{S}_A$ or $\mathcal{S}_B$, then the endpoints are generated following $\mathring{X}_m + 0.15 Z_m$ for i.i.d. standard Gaussian random variables Z_m, $m = 1, \ldots, 70000$. Provided with those candidate endpoints, we pair them to form relative comparison subject to the restriction that $|X^i_{u0} - X^i_{uj}| \leq 0.35$ for $i, j = 1, 2$, $n = 1, \ldots, N$. The responses Y_u are then generated based on the class of involving endpoints by their corresponding $\mathring{X}$ on the spirals.

For estimation, we used a larger local neighborhood $\mathcal{U}_p = \left\{ \left(x^1, x^2\right) : |x^i - p^i| \leq h \text{ for } i = 1, 2 \right\}$ with $h = \pi/2$ and weights $w_u = 1_{\{X_{u0}, X_{u1}, X_{u2} \in \mathcal{U}_p\}}$ for $u = 1, \ldots, N$ to avoid degenerate estimates.

Note that different starting points and initial velocities will generate different geodesics, not all resembling a spiral, as shown in Figure B.3.

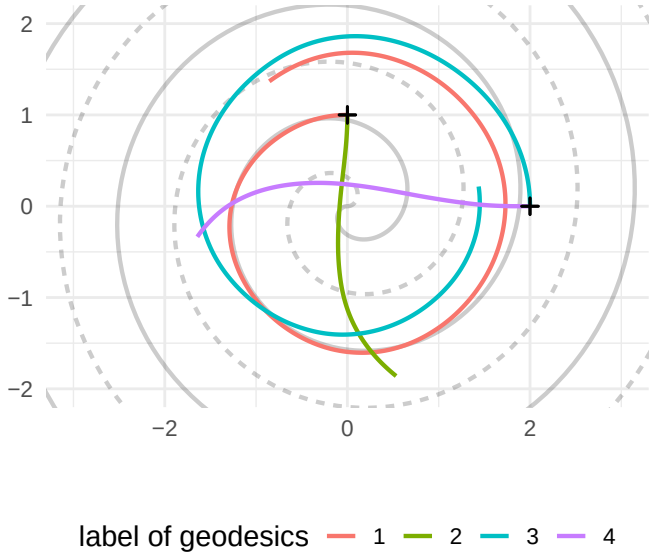


Figure B.3: Geodesics with different starting points and initial velocities under estimated metric, crosses indicate starting points.

C SPREAD OF GEODESICS

Here we provide proof to the Proposition 3.1 in the main text, which characterizes the distance between geodesics departing from a same starting point. Proposition 3.1 is a result of combining Proposition C.1 and Proposition C.2.

Proposition C.1 (spread of geodesics). Let $p \in \mathcal{M}$ and $v, w \in T_p\mathcal{M}$ be two tangent vector at p. Then the squared distance of separation satisfies Taylor expansion of

$$dist\left(\exp_p(tv), \exp_p(tw)\right)^2 = t^2 \left\|v - w\right\|^2 - \frac{1}{3}t^4 \left\langle R(v,w)w, v\right\rangle + O(t^5)$$

as $t \to 0$.

However, additional terms are introduced when computing via coordinate charts, as a result of approximating the initial velocities v and w.

Proposition C.2 (approximation of velocity). For any $p \in \mathcal{M}$, let $v \in T_p\mathcal{M}$ be a tangent vector at p and $\gamma(t) = \exp_p(tv)$ be the geodesic from p with initial velocity v. Given any local coordinate chart, write $v = v^i\partial_i$. For $i = 1, \ldots, d$, denote $\delta^i = \delta^i(t) = \gamma^i(t) - \gamma^i(0)$ as the difference in coordinate after traveling t along γ, we have

$$v^i = t^{-1}\delta^i(t) + \mathscr{R}^i(t),$$

where the remainder is

$$\mathscr{R}^i(t) = \frac{1}{2t}\delta^m\delta^n\Gamma^i_{mn} + \frac{1}{6t}\delta^m\delta^n\delta^l\left(\Gamma^k_{mn}\Gamma^i_{kl} + \partial_l\Gamma^i_{mn}\right) + O(t^3) \tag{C.1}$$

as $t \to 0$, where Γ and $\partial\Gamma$ denote the the Christoffel symbols and their derivatives at p.

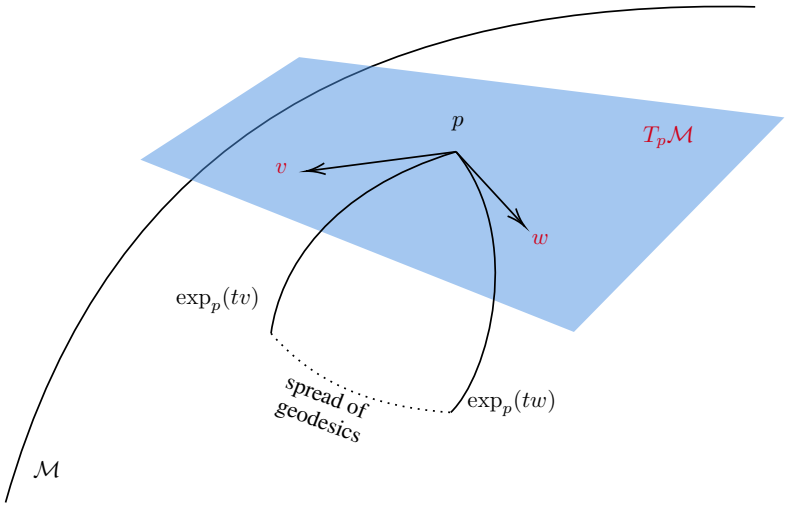


Figure C.1: A visualization for the spread of geodesics as in Proposition 3.1. A tangent space (blue plane) and tangent vectors are annotated in red.

C.1 PROOFS

Proof of Proposition C.1. Similar results can be found at Proposition 2.7 of do Carmo (1992), Proposition 5.4, of (Lang, 1999, IX, §5). We use the form presented by Meyer (1989). In the following, we reproduce the proof to the equation (9) of Meyer (1989) with some additional clarification.

Let $\gamma_0(s) = \exp_p(sv)$ and $\gamma_1(s) = \exp_p(sw)$, define a family of curves

$$V(s,t) = \exp_{\gamma_0(s)}\left(t\exp^{-1}_{\gamma_0(s)}\gamma_1(s)\right),$$

so that the curves $V_s : t \mapsto V(s,t)$ are geodesics connecting $\gamma_0(s)$ and $\gamma_1(s)$ (c.f., e.g., proposition 5.19 and equation (10.2) of Lee (2018)), and that V is a variation through geodesics V_s. Further, let $T = \partial_t V$, which is a tangent field of velocities. Let $E = \partial_s V$, which is a Jacobi field through geodesics V_s that vanishes at p. Denote $H(s) = dist\left(\gamma_0(s), \gamma_1(s)\right)^2 = \|T\| \,|_{s,t}$ for any $t \in [0,1]$, where the "$|_{s,t}$" means to take value at point $V(s,t)$. Then by the chain rules for covariant derivatives (see, e.g. Lee, 2018, chapter 4), we have

$$\begin{aligned}
\frac{d}{ds} H(s) &= \frac{d}{ds} \langle T, T\rangle \,|_{s,t} = 2 \langle D_s T, T\rangle \,|_{s,t}, \\
\left(\frac{d}{ds}\right)^2 H(s) &= 2 \left(\left\langle D_s^2 T, T\right\rangle + \|D_s T\|^2\right) |_{s,t}, \\
\left(\frac{d}{ds}\right)^3 H(s) &= 2 \left(\left\langle D_s^3 T, T\right\rangle + 3 \left\langle D_s^2 T, D_s T\right\rangle\right) |_{s,t}, \\
\left(\frac{d}{ds}\right)^4 H(s) &= 2 \left(\left\langle D_s^4 T, T\right\rangle + 3 \left\|D_s^2 T\right\| + 4 \left\langle D_s^3 T, D_s T\right\rangle\right) |_{s,t}.
\end{aligned}$$

Note that $V_0 = p$ for all t, so that $T|_{s=0,t} = 0$ for all t, hence $H'(0) = 0$.

Note that $V_s : t \mapsto V(s,t)$, $s \mapsto V(s,0)$ and $s \mapsto V(s,1)$ are geodesics; thus $D_t T = 0$ for all t, $D_s E|_{s,t=0} = 0$ and $D_s E|_{s,t=1} = 0$ for all s. In addition, by lemma 6.2 of Lee (2018), $D_s T = D_t E$.

By Jacobi equation, $D_t^2 E + R(E,T)T = 0$ for all s, which implies $D_t^2 E|_{s=0} = 0$ since $T|_{s=0} = 0$. This means the vector field $t \mapsto E|_{s=0,t}$ at p is linear in t, together with $E|_{s=0,t=0} = v$ and $E|_{s=0,t=1} = w$, we can write

$$E|_{s=0,t} = v + t(w - v)$$

for $t \in [0,1]$. Therefore $D_s T|_{s=0,t} = D_t E|_{s=0,t} = w - v$, which implies $H''(0) = 2\|v - w\|^2$.

Proceeding to the third order derivatives, observe that

$$H'''(0) = 6 \left\langle D_s^2 T, D_s T\right\rangle |_{s=0,t},$$

and by proposition 7.5 of Lee (2018), $D_s^2 T = D_s D_t E = D_t D_s E + R(E,T)E$, thus it suffices to show

$$D_s E|_{s=0,t} = 0, \text{ for all } t, \tag{C.2}$$

in order to argue $H'''(0) = 0$. Since it is known that $D_s E|_{s=0,t=0} = 0 = D_s E|_{s=0,t=1}$, it suffices to consider its derivative for (C.2). Use proposition 7.5 of Lee (2018) repeatedly, we have

$$\begin{aligned}
&D_t^2 D_s E|_{s=0,t} \\
&= D_t D_s D_t E|_{s=0,t} + D_t \left(R(T,E)E\right) |_{s=0,t} \\
&= D_t D_s^2 T|_{s=0,t} \\
&= \left(D_s D_t D_s T - R(E,T)(D_s T)\right) |_{s=0,t} \\
&= D_s D_t D_s T|_{s=0,t} \\
&= D_s \left(D_s D_t T - R(E,T)T\right) |_{s=0,t} \\
&= D_s \left(R(T,E)T\right) |_{s=0,t},
\end{aligned}$$

where the last equation is due to $D_t T = 0$. Further, by chain rule of covariant derivative (c.f. e.g. proposition 4.15 of Lee (2018)),

$$D_s \left(R(T,E)T\right) = \left(\nabla_E R\right)(T,E)T + R(D_s T, E)T + R(T, D_s E)T + R(T,E)D_s T,$$

which equals to zero at $s = 0, t$ since $T|_{s=0,t} = 0$ for all t. Hence $t \mapsto D_s E|_{s=0,t}$ is also a linear vector field, implying (C.2) and subsequently $H'''(0) = 0$.

For the fourth order derivative, note that (C.2) also implies that $D_t D_s E|_{s=0,t} = 0$ and that $D_s^2 T|_{s=0,t} = 0$ for all t. Therefore,

$$H^{(4)}(0) = 8 \left\langle D_s^3 T, D_s T\right\rangle |_{s=0,t}.$$

Further,

$$\begin{aligned}
D_s^3 T &= D_s^2 D_t E \\
&= D_s \left(D_t D_s E + R(E,T)E\right) \\
&= D_s D_t D_s E + \left(\nabla_E R\right)(E,T)E + R(D_s E, T)E + R(E, D_s T)E + R(E,T)(D_s E),
\end{aligned}$$

so $D_s^3 T|_{s=0,t} = (D_s D_t D_s E + R(E, D_s T)E)|_{s=0,t}$. Thus,

$$\left\langle D_s^3 T, D_s T \right\rangle|_{s=0,t} = (\langle D_s D_t D_s E, D_s T\rangle + \langle R(E, D_s T)E, D_s T\rangle)|_{s=0,t}.$$

Recall at $s = 0$, $D_s T|_{s=0,t} = D_t E|_{s=0,t} = w - v$, therefore

$$\begin{aligned}
&\langle R(E, D_s T)E, D_s T\rangle|_{s=0,t} \\
&= Rm(E, D_s T, E, D_s T)|_{s=0,t} \\
&= Rm(v, w-v, v, w-v) + tRm(v, w-v, w-v, w-v) \\
&= Rm(v, w, v, w).
\end{aligned}$$

Further,

$$\begin{aligned}
&\langle D_s D_t D_s E, D_s T\rangle|_{s=0,t} \\
&= \left(D_s \langle D_t D_s E, D_s T\rangle - \left\langle D_t D_s E, D_s^2 T\right\rangle\right)|_{s=0,t} \\
&= D_s \langle D_t D_s E, D_s T\rangle|_{s=0,t} \\
&= D_s D_t \langle D_s E, D_s T\rangle|_{s=0,t} - D_s \left\langle D_s E, D_t^2 E\right\rangle|_{s=0,t},
\end{aligned}$$

where the second term in the last line vanishes since $D_t^2 E|_{s=0,t} = 0$ and $D_s E|_{s=0,t} = 0$. Moreover, since the Levi–Civita connection is torsion free, we have

$$\langle D_s D_t D_s E, D_s T\rangle|_{s=0,t} = D_s D_t \langle D_s E, D_s T\rangle|_{s=0,t} = D_t D_s \langle D_s E, D_s T\rangle|_{s=0,t},$$

which should be irrelevant to t, so that $D_s \langle D_s E, D_s T\rangle|_{s=0,t}$ is linear in t. Yet

$$D_s \langle D_s E, D_s T\rangle = \left\langle D_s^2 E, D_s T\right\rangle + \left\langle D_s E, D_s^2 T\right\rangle,$$

which vanishes at $s = 0$ and $t = 0, 1$. Hence $D_s \langle D_s E, D_s T\rangle|_{s=0,t} = 0$ for all $t \in [0, 1]$. Combining those with the symmetries of Riemann curvature tensor leads to the desired expansion. □

Proof of Proposition C.2. Under the coordinate chart, we can write the geodesic curve as $\gamma : t \mapsto \left(\gamma^1(t), \ldots, \gamma^d(t)\right)$ for some smooth function $\gamma^1, \ldots, \gamma^d$. Then for any $i = 1, \ldots, d$, univariate Taylor expansion provides

$$\gamma^i(t) = \gamma^i(0) + \dot{\gamma}^i(0)t + \frac{1}{2}t^2\ddot{\gamma}^i(0) + \frac{1}{6}t^3\dddot{\gamma}^i(0) + O(t^4)$$

as $t \to 0$, where $\dot{\gamma}^i$, $\ddot{\gamma}^i$, and $\dddot{\gamma}^i$ are the first, second, and third order derivative of γ^i w.r.t. t. Note that the first derivative $\dot{\gamma}^i(0) = v^i$, and the geodesic equation and its derivative give

$$\begin{aligned}
\ddot{\gamma}^i(0) &= -v^m v^n \Gamma^i_{mn}, \\
\dddot{\gamma}^i(0) &= v^m v^n v^l \left(2\Gamma^k_{mn}\Gamma^i_{kl} - \partial_l \Gamma^i_{mn}\right).
\end{aligned}$$

Plugging into the initial Taylor expansion gives the desired result. □

Proof of Proposition 3.1 in the maintext. By Proposition C.1 and Proposition C.2, as $t \to 0$, we have

$$\begin{aligned}
&t^2 \|v - w\|^2 \\
&= \left(\delta^i_{0-1} + t\mathscr{R}^i_0(t) - t\mathscr{R}^i_1(t)\right) G_{ij} \left(\delta^j_{0-1} + t\mathscr{R}^j_0(t) - t\mathscr{R}^j_1(t)\right) \\
&= \delta^i_{0-1}\delta^j_{0-1}G_{ij} + 2t\delta^i_{0-1}\left(\mathscr{R}^j_0(t) - \mathscr{R}^j_1(t)\right) G_{ij} + O(t^4) \\
&= \delta^i_{0-1}\delta^j_{0-1}G_{ij} + \delta^i_{0-1}\left(\delta^k_0\delta^l_0 - \delta^k_1\delta^l_1\right)\left(\Gamma^j_{kl}G_{ij}\right) + O(t^4),
\end{aligned}$$

where

$$\mathscr{R}^i_a(t) = \frac{1}{2t}\delta^m_a\delta^n_a\Gamma^i_{mn} + \frac{1}{6t}\delta^m_a\delta^n_a\delta^l_a\left(\Gamma^k_{mn}\Gamma^i_{kl} + \partial_l\Gamma^i_{mn}\right) + O(t^3)$$

for $a = 0, 1$, similar to (C.1). Note that $\delta^i_0 = \delta^i_0(t) = O(t)$, it suffices to keep only the first term in the $\mathscr{R}^j_a(t)$, which is $O(t)$. □

D ASYMPTOTIC OF THE ESTIMATED METRIC TENSOR

Now we discuss the variance and bias of the estimated metric tensors. For simplicity, use the squared loss $Q(\mu, y) = (\mu - y)^2$, the identity link $g(\mu) = \mu$, and exclude the intercept $\beta^{(0)}$ and the terms $\beta^{(2)}_{ijk}$ for derivative. Given a suitable order of the indices i, j, we rewrite (3.6) into matrix form. Denote

$$\mathbf{D}_u = \left(\delta^1_{u,0-1}\delta^1_{u,0-1}, \ldots, \delta^i_{u,0-1}\delta^j_{u,0-1}, \ldots, \delta^d_{u,0-1}\delta^d_{u,0-1}\right)^T,$$
$$\boldsymbol{\beta} = \left(\beta^{(1)}_{11}, \ldots, \beta^{(1)}_{ij}, \ldots, \beta^{(1)}_{dd}\right)^T,$$

then the linear predictor $\eta_n = \mathbf{D}_u^T\boldsymbol{\beta}$. Further, write

$$\mathbf{D} = (\mathbf{D}_1, \ldots, \mathbf{D}_N)^T, \quad \boldsymbol{Y} = (Y_1, \ldots, Y_N)^T, \quad \mathbf{W} = \operatorname{diag}(w_1, \ldots, w_N),$$

so the loss (3.7) becomes

$$(\boldsymbol{Y} - \mathbf{D}\boldsymbol{\beta})^T \mathbf{W} (\boldsymbol{Y} - \mathbf{D}\boldsymbol{\beta}), \tag{D.1}$$

whose minimizer is $\hat{\boldsymbol{\beta}} = \left(\mathbf{D}^T\mathbf{W}\mathbf{D}\right)^{-1}\mathbf{D}^T\mathbf{W}\boldsymbol{Y}$. Therefore

$$\operatorname{bias}\left(\hat{\boldsymbol{\beta}}|\mathbf{D}\right) = \left(\mathbf{D}^T\mathbf{W}\mathbf{D}\right)^{-1}\mathbf{D}^T\mathbf{W}\boldsymbol{r},$$
$$\operatorname{var}\left(\hat{\boldsymbol{\beta}}|\mathbf{D}\right) = \left(\mathbf{D}^T\mathbf{W}\mathbf{D}\right)^{-1}\mathbf{D}^T\boldsymbol{\Sigma}\mathbf{D}\left(\mathbf{D}^T\mathbf{W}\mathbf{D}\right)^{-1},$$

where

$$\boldsymbol{r} = \left(\mathbb{E}(Y_u|\mathbf{D}) - \delta^i_{u,0-1}\delta^j_{u,0-1}G_{ij}\right)_{1\le n\le N},$$
$$\boldsymbol{\Sigma} = \operatorname{diag}\left(w_u^2 \operatorname{var}(Y_u|X_{u0}, X_{u1})\right)_{1\le n\le N}.$$

The assumptions (A1)–(A4) in the main text are reiterated here.

(A1) The joint density of endpoints X_{u0}, X_{u1} is positive and continuously differentiable.
(A2) The functions G_{ij}, Γ^k_{ij} are C^2-smooth for $i, j, k = 1, \ldots, d$.
(A3) The kernel K in weights (3.10) is symmetric, continuous, and has bounded support.
(A4) $\sup_u \operatorname{var}(Y_u|X_{u0}, X_{u1}) < \infty$.

Proposition D.1. Denote

$$S_{1N,i_1i_2i_3i_4} = \sum_{u=1}^N w_u\delta^{i_1}_{u,0-1}\delta^{i_2}_{u,0-1}\delta^{i_3}_{u,0-1}\delta^{i_4}_{u,0-1},$$
$$S_{2N,i_1i_2i_3i_4} = \sum_{u=1}^N w_u^2 \operatorname{var}(Y_u|X_{u0}, X_{u1})\,\delta^{i_1}_{u,0-1}\delta^{i_2}_{u,0-1}\delta^{i_3}_{u,0-1}\delta^{i_4}_{u,0-1},$$
$$S_{3N,i_1i_2} = \sum_{u=1}^N w_u\delta^{i_1}_{u,0-1}\delta^{i_2}_{u,0-1}R_u,$$

where

$$R_u = \sum_{1\le k,l,m,r\le d} \delta^m_{u,0-1}\left(\delta^k_{n0}\delta^l_{n0} - \delta^k_{n1}\delta^l_{n1}\right)\Gamma^r_{kl}G_{mr}.$$

Under (A1), (A2), (A3), and (A4), and suppose that $h \to 0$ and $Nh^{2d} \to \infty$, then

$$\mathbb{E}S_{1N,i_1i_2i_3i_4} = O\left(Nh^4\right), \quad \operatorname{var} S_{1N,i_1i_2i_3i_4} = O\left(Nh^{8-2d}\right),$$
$$\mathbb{E}S_{2N,i_1i_2i_3i_4} = O\left(Nh^{4-2d}\right), \quad \operatorname{var} S_{2N,i_1i_2i_3i_4} = O\left(Nh^{8-6d}\right),$$
$$\mathbb{E}S_{3N,i_1i_2i_3i_4} = O\left(Nh^6\right), \quad \operatorname{var} S_{3N,i_1i_2i_3i_4} = O\left(Nh^{10-2d}\right),$$

as $h \to 0$ and $N \to \infty$. So

$$S_{1N,i_1i_2i_3i_4} = O_p\left(Nh^4\right), \quad S_{2N,i_1i_2i_3i_4} = O_p\left(Nh^{4-2d}\right).$$

If further $Nh^{2+2d} \to \infty$, then

$$S_{3N,i_1i_2i_3i_4} = O_p\left(Nh^6\right),$$

as $h \to 0$ and $N \to \infty$.

Proof. Write

$$U_{u;i_1i_2i_3i_4} = w_u \delta_{u,0-1}^{i_1} \delta_{u,0-1}^{i_2} \delta_{u,0-1}^{i_3} \delta_{u,0-1}^{i_4},$$

so

$$\begin{aligned}
&\mathbb{E}U_{u;i_1i_2i_3i_4} \\
&= \int h^{-2d} \prod_{i=1}^{d} \left(K\left(\delta_{n0}^i/h\right) K\left(\delta_{n1}^i/h\right)\right) \delta_{n,0-1}^{i_1} \delta_{n,0-1}^{i_2} \delta_{n,0-1}^{i_3} \delta_{n,0-1}^{i_4} \cdot \\
&\quad f(p^1 + \delta_{n0}^1, \ldots, p^d + \delta_{n1}^d) d\delta_{n0}^1 \ldots d\delta_{n1}^d \\
&= h^4 \int K\left(s_{u0}^1\right) \cdot \cdots \cdot K\left(s_{u1}^d\right) \left(s_{u0}^{i_1} - s_{u1}^{i_1}\right) \left(s_{u0}^{i_2} - s_{u1}^{i_2}\right) \left(s_{u0}^{i_3} - s_{u1}^{i_3}\right) \left(s_{u0}^{i_4} - s_{u1}^{i_4}\right) \cdot \\
&\quad \left(f(p^1, \ldots p^d) + o(1)\right) ds_{u0}^1 \ldots ds_{u1}^d \\
&= O(h^4),
\end{aligned}$$

where f is the joint density of endpoints X_{u0}, X_{u1}, the second last equality is due to change of variables, and the last due to (A1). Similar argument implies

$$\begin{aligned}
\operatorname{var} U_{u;i_1i_2i_3i_4} &= O\left(h^{8-2d}\right), \\
\mathbb{E} w_u U_{u;i_1i_2i_3i_4} &= O\left(h^{4-2d}\right), \\
\operatorname{var} w_u U_{u;i_1i_2i_3i_4} &= O\left(h^{8-6d}\right).
\end{aligned}$$

These rates apply uniformly over n, therefore by i.i.d. and that $\operatorname{var} Y_u | X_{u0}, X_{u1}$ is uniformly bounded,

$$\begin{aligned}
\mathbb{E}S_{1N,i_1i_2i_3i_4} &= O\left(Nh^4\right), \quad \operatorname{var} S_{1N,i_1i_2i_3i_4} = O\left(Nh^{8-2d}\right), \\
\mathbb{E}S_{2N,i_1i_2i_3i_4} &= O\left(Nh^{4-2d}\right), \quad \operatorname{var} S_{2N,i_1i_2i_3i_4} = O\left(Nh^{8-6d}\right).
\end{aligned}$$

Hence

$$S_{1N,i_1i_2i_3i_4} = \mathbb{E}S_{1N,i_1i_2i_3i_4} + O_p\left(\sqrt{\operatorname{var} S_{1N,i_1i_2i_3i_4}}\right) = O_p\left(Nh^4\right),$$

under $h \to 0$ and $Nh^{2d} \to \infty$. Similarly we have results for $S_{2N,i_1i_2i_3i_4}$.

Next, write

$$V_{u;i_1i_2} = w_u \delta_{u,0-1}^{i_1} \delta_{u,0-1}^{i_2} R_u.$$

Note that

$$\begin{aligned}
&\mathbb{E}V_{u;i_1i_2} \\
&= \int h^{-2d} \prod_{i=1}^{d} \left(K\left(\delta_{n0}^i/h\right) K\left(\delta_{n1}^i/h\right)\right) \delta_{n,0-1}^{i_1} \delta_{n,0-1}^{i_2} \times \\
&\quad \sum_{1 \le k,l,m,r \le d} \delta_{u,0-1}^m \left(\delta_{n0}^k \delta_{n0}^l - \delta_{n1}^k \delta_{n1}^l\right) F_{klm} \times \\
&\quad f(p^1 + \delta_{n0}^1, \ldots, p^d + \delta_{n1}^d) d\delta_{n0}^1 \ldots d\delta_{n1}^d \\
&= h^5 \int K\left(s_{u0}^1\right) \cdot \cdots \cdot K\left(s_{u1}^d\right) \left(s_{u0}^{i_1} - s_{u1}^{i_1}\right) \left(s_{u0}^{i_2} - s_{u1}^{i_2}\right) \times \\
&\quad \sum_{1 \le k,l,m,r \le d} \left(s_{u0}^{i_2} - s_{u1}^{i_2}\right) \left(s_{u0}^k s_{u0}^l - s_{u1}^k s_{u1}^l\right) F_{klm} \times \\
&\quad \left(f(p^1, \ldots p^d) + h \sum_{r=1}^{d} \frac{\partial f}{\partial p^r}(p) \left(s_{u0}^r + s_{u1}^r\right) + o(h)\right) ds_{u0}^1 \ldots ds_{u1}^d \\
&= O(h^6),
\end{aligned}$$

where $F_{klm} = \Gamma^r_{kl} G_{mr}$. Indeed, since the kernel K is symmetric, and the leading terms in the integrant is of fifth power of s, thus with some abuse of notation, $\mathbb{E}V_{u;i_1i_2} = h^5 F \int K(s)s^5 \left(f + hO(s)\right) ds = O(h^6)$. Similarly

$$\operatorname{var} V_{u;i_1i_2} = O(h^{10-2d}).$$

The rest of the results for $S_{3N,i_1i_2i_3i_4}$ proceeds analogously to that of $S_{1N,i_1i_2i_3i_4}$ and $S_{2N,i_1i_2i_3i_4}$. □

Proposition D.2. Under the conditions of Proposition D.1,

$$\operatorname{bias}\left(\hat{\boldsymbol{\beta}}|\mathbf{X}\right) = O_p\left(h^2\right), \quad \operatorname{var}\left(\hat{\boldsymbol{\beta}}|\mathbf{X}\right) = O_p\left(\frac{1}{Nh^{4+2d}}\right),$$

as $h \to 0$ and $Nh^{2+2d} \to \infty$, where $\mathbf{X}$ are the observed endpoints.

Proof. Note that $S_{1N;i_1i_2i_3i_4}$ are elements of $\mathbf{D}^T\mathbf{W}\mathbf{D}$, where one pair of (i_1, i_2) index a row while one pair of (i_3, i_4) index a column for $i_1, i_2, i_3, i_4 = 1, \ldots, d$. Similarly $S_{2N;i_1i_2i_3i_4}$ are elements of $\mathbf{D}^T\boldsymbol{\Sigma}\mathbf{D}$, and $S_{3N;i_1i_2}$ are elements of $\mathbf{D}^T\mathbf{W}\boldsymbol{r}$ by Proposition 3.1. Applying Proposition D.1 leads to the result. □