The Annals of Statistics
2021, Vol. 49, No. 6, 3510–3533
https://doi.org/10.1214/21-AOS2095

TOTAL VARIATION REGULARIZED FRÉCHET REGRESSION FOR METRIC-SPACE VALUED DATA

BY ZHENHUA LIN[1] AND HANS-GEORG MÜLLER[2]

[1] *Department of Statistics and Data Science, National University of Singapore, linz@nus.edu.sg*
[2] *Department of Statistics, University of California, Davis, hgmueller@ucdavis.edu*

Non-Euclidean data that are indexed with a scalar predictor such as time are increasingly encountered in data applications, while statistical methodology and theory for such random objects are not well developed yet. To address the need for new methodology in this area, we develop a total variation regularization technique for nonparametric Fréchet regression, which refers to a regression setting where a response residing in a metric space is paired with a scalar predictor and the target is a conditional Fréchet mean. Specifically, we seek to approximate an unknown metric-space valued function by an estimator that minimizes the Fréchet version of least squares and at the same time has small total variation, appropriately defined for metric-space valued objects. We show that the resulting estimator is representable by a piece-wise constant function and establish the minimax convergence rate of the proposed estimator for metric data objects that reside in Hadamard spaces. We illustrate the numerical performance of the proposed method for both simulated and real data, including metric spaces of symmetric positive-definite matrices with the affine-invariant distance, of probability distributions on the real line with the Wasserstein distance, and of phylogenetic trees with the Billera–Holmes–Vogtmann metric.

1. Introduction. Regression analysis is a foundational technique in statistics aiming to model the relationship between response variables and covariates or predictor variables. Conventional regression models are designed for Euclidean responses Y and predictors X and include parametric models such as linear or polynomial regression and generalized linear models as well as various nonparametric approaches, such as kernel and spline smoothing. All of these models target the conditional expectation $\mathbb{E}(Y|X)$.

In response to the emergence of new types of data, the basic Euclidean regression models have been extended to the case of non-Euclidean data, where a relatively well-studied scenario concerns manifold-valued responses. For instance, Chang (1989), Fisher (1995) studied regression models for spherical and circular data, while Shi et al. (2009), Steinke, Hein and Schölkopf (2010), Davis et al. (2010), Fletcher (2013), Cornea et al. (2017) investigated such models for the case of more general Riemannian manifolds. Also classical local regression techniques, such as Nadaraya–Watson smoothing and local polynomial smoothing, have been generalized to cover responses that lie on manifolds (Pelletier (2006), Yuan et al. (2012), Hinkle, Fletcher and Joshi (2014)). In this paper, we extend the scope of these previous approaches and study the regression problem for response variables that are situated on a metric space, more specifically, a Hadamard or Alexandrov space. Due to the absence of rich geometric and algebraic structure in these metric spaces, this problem poses new challenges that go beyond the regression problem for the Euclidean or manifold case.

While regression with metric-space valued responses covers a wide range of random objects and therefore is of intrinsic interest, the literature on this topic so far is quite limited.

Received June 2020; revised May 2021.
MSC2020 subject classifications. Primary 62R20; secondary 62R30.
Key words and phrases. Brain imaging, Fréchet mean, Hadamard space, Alexandrov space, non-Euclidean data, symmetric positive-definite matrix, random objects, Wasserstein metric, phylogenetic tree.

Existing work includes Faraway (2014), who considered regression for non-Euclidean data by a Euclidean embedding using distance matrices, similar to multidimensional scaling, as well as intrinsic approaches by Hein (2009), who studied Nadaraya–Watson kernel regression for general metric spaces, and by Petersen and Müller (2019), who introduced linear and local linear regression for metric-space valued response variables and approached the regression problem within the framework of conditional Fréchet means.

In this paper, we propose a novel regularization approach for nonparametric regression with metric-space valued response variables and a scalar predictor variable. We utilize a total variation based penalty, introducing in Section 3 an appropriate modification of the definition of total variation that covers metric-space valued functions. Specifically, the inclusion of a total variation penalty term in the estimating equation for Fréchet regression leads to a penalized M-estimation approach for metric-space valued data. We refer to the proposed method as total variation regularized Fréchet regression or simply *regularized Fréchet regression*. While regularized Fréchet regression can be developed for any geodesic metric space, we focus here primarily on the family of Hadamard spaces. This family includes the Euclidean space and forms a rich class of metric spaces that have important practical applications; see Examples 1–3 and Section 6 for more details.

Total variation regularization was introduced by Rudin, Osher and Fatemi (1992) for image recovery/denoising. There is a vast literature on this regularization technique from the perspective of image denoising and signal processing; see Chambolle et al. (2010) for a brief introduction and review. From a statistical perspective and for Euclidean data, this method was studied by Mammen and van de Geer (1997) from the viewpoint of locally adaptive regression splines and by Tibshirani et al. (2005), who connected it to the lasso. Recent developments along this line include optimal rates (Hütter and Rigollet (2016)), trend filtering (Kim et al. (2009), Tibshirani (2014)) and total variation regularized regression when predictors are on a tree or graph (Wang et al. (2016), Ortelli and van de Geer (2018)). Extensions to manifold-valued data were first investigated by Pennec, Fillard and Ayache (2006) with a robust variant of the total variation regularization, then by Lellmann et al. (2013), Weinmann, Demaret and Storath (2014) with the first-order total variation, and further by Bergmann et al. (2014), Bergmann and Weinmann (2016) with the second-order total variation, although without asymptotic analysis. Total variation penalties were also shown to confer advantages for regression models in brain imaging (Wang, Zhu and ADNI (2017)). We generalize these approaches to the case of data in a Hadamard space and provide a detailed asymptotic analysis for total variation regularized Fréchet regression for the first time. The generalization of total variation regularization to Hadamard spaces, and especially the theoretical analysis, are challenged by the lack of rich differential structures that are not available in generic Hadamard spaces.

We tackle these challenges by leveraging the convexity of the Hadamard space, taking advantage of the convexity of the distance function and the strong convexity of the squared distance function; see Section 4. Moreover, to overcome the technical difficulties arising from the lack of vector and analytic structures of Hadamard spaces, we develop new geometric ideas that are relevant for statistical analysis in these spaces, such as Alexandrov inner product, geometric interpolation of metric-space valued functions, and geometric center of functions (Fréchet integrals); see Appendix B for details. Combined with convexity, these new constructions enable us to obtain minimax rates of convergence for the proposed estimator for a family of Hadamard spaces and functions of bounded variation. In addition, as these geometric constructions apply to general metric spaces and convexity extends to certain subspaces of Alexandrov spaces, the theory also applies for certain non-Hadamard spaces.

The structure of the paper is as follows. A brief introduction to metric geometry is given in Section 2. Total variation regularized Fréchet regression is introduced in Section 3, and asymptotic results are presented in Section 4. Numerical studies for synthetic data are provided in Section 5. In Section 6, we illustrate the application of the proposed method to analyze data on the evolution of human mortality profiles using the Wasserstein distance on the space of probability distributions and to study the dynamics of brain connectivity using task-related functional magnetic resonance imaging (fMRI) signals and the affine-invariant distance on the space of symmetric positive-definite matrices.

2. Concepts and tools from metric geometry.

To state the estimation method and theory in Sections 3 and 4, we need to make use of various concepts from metric geometry that are briefly reviewed here; a more comprehensive treatment can be found in Chapters 2, 4 and 9 of Burago, Burago and Ivanov (2001).

Geodesics. For a generic metric space $(\mathcal{M}, d)$ and a closed interval $\mathcal{T} = [a, b] \subset \mathbb{R}$, given a curve γ parameterized by $\mathcal{T}$ on $\mathcal{M}$, that is, $\gamma : \mathcal{T} \to \mathcal{M}$, and a set $P = \{t_0 \le t_1 \le \cdots \le t_k\} \subset \mathcal{T}$ consisting of $k + 1$ points in $\mathcal{T}$, we use the quantity $R_d(\gamma, P) = \sum_{j=1}^{k} d(\gamma(t_j), \gamma(t_{j-1}))$ to define the length of γ, denoted by $|\gamma|$, which is given by

$$|\gamma| = \sup_{P\in\mathcal{P}} R_d(\gamma, P); \tag{2.1}$$

here $\mathcal{P}$ is the collection of subsets of $\mathcal{T}$ whose cardinality is finite. The metric space $(\mathcal{M}, d)$ is a length space if $d(p, q) = \inf_\gamma |\gamma|$, where the infimum ranges over all continuous curves $\gamma : \mathcal{T} \to \mathcal{M}$ connecting p and q, that is, $\gamma(a) = p$ and $\gamma(b) = q$. A geodesic on $\mathcal{M}$ is a curve $\gamma : \mathcal{T} \to \mathcal{M}$ such that $d(\gamma(s), \gamma(t)) = |t - s|$ for $s, t \in \mathcal{T}$. The metric space $(\mathcal{M}, d)$ is a geodesic space if any pair of points can be connected by a geodesic, and is a uniquely geodesic space if this geodesic is unique. The geodesic connecting p and q in a uniquely geodesic space is denoted by $\overline{pq}$. Geodesics in a metric space are the counterpart of straight lines in a Euclidean space. They have been explored for statistical regression of non-Euclidean data, such as geodesic regression (Fletcher (2013)).

Curvature. Unlike Euclidean spaces, a general metric space is often not flat, and curvature is used to measure the amount of deviation from being flat. A standard approach to classifying curvature is to compare geodesic triangles on the metric space to those on the following reference spaces M_κ^2:

- When $\kappa = 0$, $M_\kappa^2 = \mathbb{R}^2$ with the standard Euclidean distance;
- When $\kappa < 0$, M_κ^2 is the hyperbolic space $\mathbb{H}^2 = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 - z^2 = -1 \text{ and } z > 0\}$ with the hyperbolic distance function $d(p, q) = \cosh^{-1}(z_p z_q - x_p x_q - y_p y_q)/\sqrt{-\kappa}$, where $p = (x_p, y_p, z_p)$ and $q = (x_q, y_q, z_q)$;
- When $\kappa > 0$, M_κ^2 is the sphere $\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$ with the angular distance function $d(p, q) = \cos^{-1}(x_p x_q + y_p y_q + z_p z_q)/\sqrt{\kappa}$.

A geodesic triangle with vertices p, q, r in a uniquely geodesic space $\mathcal{M}$, denoted by $\triangle(p, q, r)$, consists of three geodesic segments that connect p to q, p to r and q to r, respectively. A comparison triangle of $\triangle(p, q, r)$ in the reference space M_κ^2 is a geodesic triangle on M_κ^2 formed by vertices $\bar{p}$, $\bar{q}$, $\bar{r}$ such that $d(p, q) = \bar{d}_\kappa(\bar{p}, \bar{q})$, $d(p, r) = \bar{d}_\kappa(\bar{p}, \bar{r})$, and $d(q, r) = \bar{d}_\kappa(\bar{q}, \bar{r})$, where $\bar{d}_\kappa$ denotes the distance function on M_κ^2. In addition, every point x on the geodesic $\overline{pq}$ ($\overline{pr}$, respectively) has a counterpart $\bar{x}$ on the geodesic segment $\overline{\bar{p}\bar{q}}$ ($\overline{\bar{p}\bar{r}}$, respectively) of the comparison triangle such that $d(p, x) = \bar{d}_\kappa(\bar{p}, \bar{x})$. We say the (global) curvature of $\mathcal{M}$ is lower (upper, respectively) bounded by κ if every geodesic triangle with perimeter less than $2D_\kappa$, where $D_\kappa = \pi/\sqrt{\kappa}$ if $\kappa > 0$ and $D_\kappa = \infty$ otherwise, satisfies the following property: There exists a comparison triangle $\triangle(\bar{p}, \bar{q}, \bar{r})$ in M_κ such that $d(x, y) \ge \bar{d}_\kappa(\bar{x}, \bar{y})$ ($d(x, y) \le \bar{d}_\kappa(\bar{x}, \bar{y})$, respectively) for all $x \in \overline{pq}$ and $y \in \overline{pr}$ and their comparison points $\bar{x}$ and $\bar{y}$ on $\triangle(\bar{p}, \bar{q}, \bar{r})$.

Angles. The comparison angle $\bar{\angle}_p(q,r)$ between q and r at p is defined by

$$\bar{\angle}_p(q,r) = \arccos \frac{d^2(p,q) + d^2(p,r) - d^2(q,r)}{2d(p,q)d(p,r)}. \tag{2.2}$$

This is utilized to introduce the concept of an (Alexandrov) angle between two geodesics γ and η emanating from p in a uniquely geodesic space, which is denoted by $\angle_p(\gamma,\eta)$ and defined by

$$\angle_p(\gamma,\eta) = \limsup_{s,t\to 0} \bar{\angle}_p\big(\gamma(s),\eta(t)\big).$$

Note that $\angle_p(\gamma,\eta)$ does not depend on the length of γ or η. For three distinct points p, q, r in a uniquely geodesic subset of $\mathcal{M}$, we define the angle $\angle_p(q,r) = \angle_p(\overline{pq}, \overline{pr})$.

Alexandrov spaces and Hadamard spaces. A geodesic space with lower or upper bounded curvature is called an Alexandrov space, and a complete geodesic space with curvature upper bounded by 0 is called a Hadamard space. Every geodesic triangle $\triangle(p,q,r)$ in a Hadamard space then satisfies the CAT(0) inequality, that is, $d(x,y) \le \bar{d}_0(\bar{x},\bar{y})$ for all $x \in \overline{pq}$ and $y \in \overline{pr}$ and their comparison points $\bar{x}, \bar{y} \in \mathbb{R}^2$. A geodesic space in which every geodesic triangle satisfies the CAT(0) inequality is called a CAT(0) space; a Hadamard space is a complete CAT(0) space. Moreover, every CAT(0) space is uniquely geodesic. Every Euclidean space is a Hadamard space, while non-Euclidean Hadamard spaces include symmetric positive definite matrices, some Wasserstein spaces and Billera–Holmes–Vogtmann phylogenetic tree spaces and more; see Examples 1–3. These spaces have broad applications in science and statistics.

Riemannian manifolds. A Riemannian manifold is a smooth manifold with a smooth metric tensor $\langle\cdot,\cdot\rangle$ (Lang (1995), page 170), such that for each $p \in \mathcal{M}$, the tensor $\langle\cdot,\cdot\rangle_p$ defines an inner product on the tangent space $T_p\mathcal{M}$ at p. The metric tensor induces a distance function that turns the Riemannian manifold into a metric space (Lang (1995), page 184). The sectional curvature at p is defined for two linearly independent tangent vectors u and v at $p \in \mathcal{M}$ and is given by $\frac{\langle \mathfrak{R}(u,v)v,u\rangle_p}{\langle u,u\rangle_p \langle v,v\rangle_p - \langle u,v\rangle_p^2} \in \mathbb{R}$, where $\mathfrak{R}$ is the Riemannian curvature tensor (Lang (1995), page 227). A complete Riemannian manifold is a Hadamard manifold if it is simply connected and has everywhere nonpositive sectional curvature.

3. Regularized Fréchet regression with total variation.

Let $(\mathcal{M}, d)$ be a metric space and Y a random element in $\mathcal{M}$, where d denotes the distance function on $\mathcal{M}$. When $\mathcal{M}$ is a Euclidean space, which is a special metric space, the expectation or mean of Y is an important concept to characterize the average location of Y. For a non-Euclidean metric space, we replace the mean with the Fréchet mean, which is an element of $\mathcal{M}$ that minimizes the Fréchet function $F(\cdot) = \mathbb{E}d^2(\cdot, Y)$; in the Euclidean case it coincides with the usual mean for random vectors with finite second moments. In a general metric space with a given probability measure, the Fréchet mean might not exist, and even when it exists it might not be unique. We shall assume that Fréchet means exist and are unique for the random objects we consider in the following. This is the case for Hadamard spaces when $F(p) < \infty$ for some $p \in \mathcal{M}$ (Bhattacharya and Patrangenaru (2003), Sturm (2003), Afsari (2011), Patrangenaru and Ellingson (2015)) and Alexandrov spaces with sufficient concentration assumption and/or additional convexity conditions (Lin and Müller (2021), Lemma S.7).

We consider a curve $\mu : \mathcal{T} \to \mathcal{M}$ that potentially varies with the sample size n and is parameterized by the interval $\mathcal{T} = [a,b]$. For $n > 0$ independent observations Y_i at the designated time point $t_i \in \mathcal{T}$ for $i = 1, \dots, n$, we assume the following model:

$$\mathbb{E}d^2(y, Y_i) < \infty \quad \text{for some } y \in \mathcal{M} \quad \text{and} \quad \mu(t_i) = \arg\min_{y\in\mathcal{M}} \mathbb{E}d^2(y, Y_i), \tag{3.1}$$

and assume that $a \le t_1 \le \cdots \le t_n \le b$ are equally spaced; the assumption of equal spacing that we adopt here for simplicity is not essential, and the results can be easily extended to the non-equally spaced case, by applying the concept of design densities (Sacks and Ylvisaker (1970)).

Our goal is to obtain a mean curve estimate $\hat{\mu}$ from the given data pairs (t_i, Y_i) by minimizing the loss function

$$L_\lambda(\gamma) = \frac{1}{n}\sum_{i=1}^{n} d^2\big(\gamma(t_i), Y_i\big) + \lambda\,\mathrm{TV}(\gamma),$$

where $\mathrm{TV}(\gamma) = |\gamma|$ is the total variation of the curve γ, measured by its length as defined by equation (2.1), and $\lambda \ge 0$ is a regularization parameter depending on n. The curve estimate is then

$$\hat{\mu} \in \operatorname*{arg\,min}_{\mathrm{TV}(\gamma)<\infty} L_\lambda(\gamma), \tag{3.2}$$

and its deviation from the target μ is quantified by the pseudo-metrics

$$d_n(\hat{\mu}, \mu) = \left\{ n^{-1}\sum_{i=1}^{n} d^2\big(\hat{\mu}(t_i), \mu(t_i)\big)\right\}^{1/2}, \tag{3.3}$$

where $\tilde{d}$ is a pseudo-metric if $\tilde{d}(f, g) = \tilde{d}(g, f) \ge 0$ and $\tilde{d}(f, h) \le \tilde{d}(f, g) + \tilde{d}(g, h)$ for all f, g, h. In the above, both L_λ and d_n are empirical, in the sense that they compare γ and $\hat{\mu}$ with their respective targets only at the design points $t_1, \dots, t_n$. Nevertheless, the theory developed in the next section implies that with probability tending to one $\hat{\mu}$ converges to μ, in the sense that $\int_{\mathcal{T}} d^2(\hat{\mu}(t), \mu(t))\,\mathrm{d}t \to 0$, under the assumption $\mathrm{TV}(\mu) \le C$ for a fixed constant $C \ge 0$ and a suitable asymptotic assumption on the spacing of the design points t_i that will be satisfied for example if these points are equidistantly distributed over an interval.

The estimator $\hat{\mu}$, although not unique, has the property that $\hat{\mu}(t) = \hat{\mu}(t_1)$ for $t \in [a, t_1]$ and $\hat{\mu}(t) = \hat{\mu}(t_n)$ for $t \in [t_n, b]$. Otherwise, the following function

$$\check{\mu}(t) = \begin{cases} \hat{\mu}(t_1) & \text{for } t \in [a, t_1), \\ \hat{\mu}(t) & \text{for } t \in [t_1, t_n], \\ \hat{\mu}(t_n) & \text{for } t \in (t_n, b], \end{cases}$$

satisfies $n^{-1}\sum_{i=1}^{n} d^2(\check{\mu}(t_i), Y_i) = n^{-1}\sum_{i=1}^{n} d^2(\hat{\mu}(t_i), Y_i)$ and $\mathrm{TV}(\check{\mu}) < \mathrm{TV}(\hat{\mu})$, which implies $L_\lambda(\check{\mu}) < L_\lambda(\hat{\mu})$ and thus contradicts the optimality of $\hat{\mu}$. Indeed, the following result shows that $\hat{\mu}$ can be chosen to have a simple structure.

PROPOSITION 1. *For any $\tilde{\mu}$ that minimizes $L_\lambda(\cdot)$, there is a step function $\hat{\mu}$ such that $\hat{\mu}(t_i) = \tilde{\mu}(t_i)$ for all $i = 1, \dots, n$ and* $\mathrm{TV}(\hat{\mu}) \le \mathrm{TV}(\tilde{\mu})$.

PROOF. It is clear that $\mathrm{TV}(\tilde{\mu}) \ge \sum_{i=0}^{n} d(\tilde{\mu}(t_{i+1}), \tilde{\mu}(t_i))$, where $t_0 = a$ and $t_{n+1} = b$. Define

$$\hat{\mu}(t) = \begin{cases} \tilde{\mu}(t_i), & t \in [a, b) \text{ and } t \in [t_i, t_{i+1}), \\ \tilde{\mu}(t_n), & t = b. \end{cases}$$

Then $\hat{\mu}(t_i) = \tilde{\mu}(t_i)$ for $i = 1, \dots, n$. Also, from the definition, $\hat{\mu}(t)$ is constant over $[t_i, t_{i+1})$. One thus finds $\mathrm{TV}(\hat{\mu}) = \sum_{i=0}^{n} d(\hat{\mu}(t_{i+1}), \hat{\mu}(t_i)) = \sum_{i=0}^{n} d(\tilde{\mu}(t_{i+1}), \tilde{\mu}(t_i)) \le \mathrm{TV}(\tilde{\mu})$. □

The above proposition shows that one can always choose a step function to minimize the loss function L_λ. In the following, we may therefore assume that $\hat{\mu}$ is a step function. The class of step functions is not only sufficiently powerful to approximate any function of finite total variation, but also advantageous in modeling functions that are discontinuous since it incorporates jumps of the function estimates, in contrast to classical smoothing methods that usually assume a smooth underlying regression function. Incorporating jumps or discontinuities is of interest in many applications (Kolar and Xing (2012), Zhu, Fan and Kong (2014), Dubey and Müller (2020a)). Our approach makes it possible to go beyond Euclidean spaces and to fit metric-space valued functions with jumps, as demonstrated in Section 6.2.

The tuning parameter λ controls the number of constant pieces of the estimate $\hat{\mu}$ and the magnitude of the distance between the pieces. For instance, a large value of λ leads to a small number of constant pieces. In the next section we will show that the choice $\lambda \asymp n^{-2/3}$ will optimize the asymptotic performance, where the notation $\lambda \asymp n^{-2/3}$ denotes that there are constants $c_2 \geq c_1 > 0$ such that $c_1 n^{-2/3} \leq \lambda \leq c_2 n^{-2/3}$. In practice, λ can be chosen via cross-validation. In some situations it is useful to choose it as the minimal number that yields a desired number of pieces of $\hat{\mu}$; see Section 6.2. For computation of $\hat{\mu}$, we adopt the iterative proximal point algorithm of Weinmann, Demaret and Storath (2014), who showed that this algorithm is convergent for Hadamard spaces; further details are in Appendix A.

4. Theory.

4.1. *Hadamard manifolds and spaces.* To study the asymptotic properties of the estimate $\hat{\mu}$ given in (3.2), we assume uniform sub-Gaussianity of the random quantities $d(\mu(t_i), Y_i)$, as follows. A random variable X is sub-Gaussian if $\mathbb{E}\exp(\beta X^2) < \infty$ for a constant $\beta > 0$, and a collection $\mathcal{X}$ of random variables is uniformly sub-Gaussian, if there are constants $\beta, \zeta > 0$ such that $\mathbb{E}\exp(\beta X^2) \leq \zeta < \infty$ for all $X \in \mathcal{X}$. The following condition states that the distances of random objects Y_i to their Fréchet means are uniformly sub-Gaussian. This is guaranteed and thus the condition is not needed whenever the diameter of the space $\mathcal{M}$ is bounded.

(H1) There exist constants $\beta > 0$ and $\zeta > 0$ such that for the data Y_i in model (3.1)

$$\sup_{1\leq i\leq n} \mathbb{E}\left[\exp\{\beta d^2(\mu(t_i), Y_i)\}\right] \leq \zeta < \infty,$$

that is, the random variables $d(\mu(t_i), Y_i)$ are uniformly sub-Gaussian.

Let $\mathscr{V}_{\mathcal{M}}$ be the collection of all $\mathcal{M}$-valued curves of bounded total variation. We focus on a subcollection $\mathscr{G}_{\mathcal{M}} \subset \mathscr{V}_{\mathcal{M}}$, which could correspond to the entire collection $\mathscr{V}_{\mathcal{M}}$ or a proper subcollection of $\mathscr{V}_{\mathcal{M}}$ such as the class of Lipschitz continuous curves. Then the pseudo-metric function d_n in (3.3) turns $\mathscr{G}_{\mathcal{M}}$ into a pseudo-metric space. Let $\mathscr{G}_{\mathcal{M}}^R(C) \subset \mathscr{G}_{\mathcal{M}}$ be a collection of functions $\gamma \in \mathscr{G}_{\mathcal{M}}$ with $\mathrm{TV}(\gamma) \leq C$, such that there exists a ball $\mathcal{B} \subset \mathcal{M}$ of radius $R > 0$ with $\gamma(t) \in \mathcal{B}$ for all γ and t; we write $\mathscr{G}_{\mathcal{M}}(C) = \mathscr{G}_{\mathcal{M}}^{\infty}(C)$. The following result, valid for any (non-unique) minimizer $\hat{\mu}$ in (3.2), establishes the convergence rate of the estimator $\hat{\mu}$ for μ, where μ is allowed to vary with the sample size n.

THEOREM 1. *For a family $\mathscr{R}(p,\kappa)$ of complete and simply connected Riemannian manifolds of dimension no larger than p and with sectional curvature bounded between $\kappa \leq 0$ and 0, choosing $\lambda \asymp n^{-2/3}$ implies that*

$$\lim_{D\to\infty} \limsup_{n\to\infty} \sup_{F\in\mathscr{F}_n} \mathbb{P}_F\{d_n(\hat{\mu}, \mu) > Dn^{-1/3}\} = 0,$$

where μ is defined in (3.1), $\hat{\mu}$ is given in (3.2), $\mathbb{P}_F$ is the probability measure induced by F, and $\mathscr{F}_n = \mathscr{F}_n(p, \kappa, C, \beta, \zeta)$ for constants $p, C, \beta, \zeta > 0$ and $\kappa \leq 0$ is the collection of joint probability distributions of $Y_1, \dots, Y_n$ on $\mathcal{M}$ for which $\mathcal{M} \in \mathscr{R}(p,\kappa)$, $\mathrm{TV}(\mu) \leq C$ and (H1) holds for $\beta, \zeta > 0$.

The manifold in the above theorem is a Hadamard manifold which is also a Hadamard space according to Theorem 1A.6 of Bridson and Haefliger (1999). This motivates us to generalize the above result to general Hadamard spaces that are not a manifold. To this end, we first observe that Riemannian manifold-valued functions of bounded total variation satisfy an entropy condition, as follows. For a subset $\mathscr{B}$ of $\mathscr{G}_{\mathcal{M}}$, the minimal number of balls of radius δ in $(\mathscr{G}_{\mathcal{M}}, d_n)$ to cover $\mathscr{B}$ is denoted by $N(\delta, \mathscr{B}, d_n)$. The covering number $N(\delta, \mathscr{B}, d_n)$ depends on d_n, which in turn depends on the metric d as per (3.3). Proposition 4 in Appendix D shows that manifolds $\mathcal{M}$ in the family $\mathscr{R}(p, \kappa)$ of Theorem 1 satisfy the following condition:

(H2) For a fixed $R > 0$, there exists a constant $K > 0$ that may depend on R, such that $\log N(\delta, \mathscr{G}^r_{\mathcal{M}}(r), d_n) \leq K\delta^{-1}$ for all $\delta > 0$, $n \geq 1$ and $0 < r \leq R$.

This condition essentially controls the (local) complexity of the underlying space $\mathcal{M}$, and is key for the asymptotic analysis based on empirical process theory, such as Mammen and van de Geer (1997). For those Hadamard spaces and classes $\mathscr{G}_{\mathcal{M}}$ of functions that satisfy the condition, we have the following result that generalizes Theorem 1.

THEOREM 2. *For $C > 0$, for a family $\mathscr{H}(K)$ of Hadamard spaces such that for each $\mathcal{M} \in \mathscr{H}(K)$ the class of functions $\mathscr{G}_{\mathcal{M}}$ satisfies the condition (H2) for $R = 15C$, with $\lambda \asymp n^{-2/3}$, one has*

$$\lim_{D\to\infty} \limsup_{n\to\infty} \sup_{F\in\mathscr{F}_n} \mathbb{P}_F\{d_n(\hat{\mu}, \mu) > Dn^{-1/3}\} = 0,$$

where μ is defined in (3.1), $\hat{\mu}$ is given in (3.2), $\mathbb{P}_F$ is the probability measure induced by F, and $\mathscr{F}_n = \mathscr{F}_n(K, C, \beta, \zeta)$ for constants $K, C, \beta, \zeta > 0$ is the collection of joint probability distributions of $Y_1, \dots, Y_n$ on $\mathcal{M}$ for which $\mathcal{M} \in \mathscr{H}(K)$, $\mu \in \mathscr{G}_{\mathcal{M}}(C)$, and (H1) holds for $\beta, \zeta > 0$.

When $\mathcal{M}$ is the one-dimensional Euclidean space $\mathbb{R}$, Donoho and Johnstone (1998) showed that the minimax rate is $n^{-1/3}$ for the class of uniformly bounded variation; see also Sadhanala, Wang and Tibshirani (2016). Since $\mathscr{H}(K)$ contains the one-dimensional Euclidean space for the same class of functions, the rate in the above theorem is also the minimax rate for the family $\mathscr{H}(K)$; our result is thus a generalization of the minimax result of Donoho and Johnstone (1998) to Hadamard spaces. In addition, if the entropy condition of (H2) is replaced with $\log N(\delta, \mathscr{G}^r_{\mathcal{M}}(r), d_n) \leq K\delta^{-\alpha}$ for some constant $\alpha \in (0, 2)$, then the proof of Theorem 2 can be modified to show that $d(\hat{\mu}, \mu) = O_P(n^{-1/(2+\alpha)})$.

There are various geometric properties of Hadamard spaces that enable the extension in Theorem 2; the most important among these is the convexity outlined in the following proposition.

PROPOSITION 2. *For $C > 0$, let $\mathscr{M}(K)$ be a family of metric spaces such that for each $\mathcal{M} \in \mathscr{M}(K)$ the class $\mathscr{G}_{\mathcal{M}}(C)$ of functions satisfies (H2) with $R = 15C$. In addition, the following conditions hold for a universal constant $C_1 > 0$. For each $\mathcal{M} \in \mathscr{M}(K)$:*

(a) *$d^2(q, r) \geq d^2(p, r) - 2d(p, q)d(p, r)\cos\angle_p(q, r) + d^2(p, q)$ for all $p, q, r \in \mathcal{M}$;*

(b) *the function $f(r) = d(p, r)\cos\angle_p(q, r)$ is Lipschitz continuous with a Lipschitz constant no larger than C_1 for all $p, q \in \mathcal{M}$;*

(c) *$\mathbb{E}\{d(\mu(t_i), Y_i)\cos\angle_{\mu(t_i)}(Y_i, q)\} \leq 0$ for all $q \in \mathcal{M}$, $n \geq 1$ and $1 \leq i \leq n$.*

For $\lambda \asymp n^{-2/3}$, it then holds that

$$\lim_{D\to\infty} \limsup_{n\to\infty} \sup_{F\in\mathscr{F}_n} \mathbb{P}_F\{d_n(\hat{\mu}, \mu) > Dn^{-1/3}\} = 0, \tag{4.1}$$

where μ is defined in (3.1), *$\hat{\mu}$ is given in* (3.2), *$\mathbb{P}_F$ is the probability measure induced by F, and $\mathscr{F}_n = \mathscr{F}_n(K, C, \beta, \zeta)$ is a collection of joint probability distributions of $Y_1, \dots, Y_n$ on $\mathcal{M} \in \mathscr{M}(K)$ such that $\mu \in \mathscr{G}_{\mathcal{M}}(C)$, and the conditions* (c) *and* (H1) *hold for $\beta, \zeta > 0$.*

The first two conditions of the above proposition emerge as properties of Hadamard space. In fact, condition (a) is an alternative characterization of the CAT(0) space, which has nonpositive curvature (also known as NPC space). To see this, by Proposition 1.7 in Chapter II.1 of Bridson and Haefliger (1999), $\mathcal{M}$ is a CAT(0) space if and only if for all $p, q, r \in \mathcal{M}$, $d(q,r) \geq \bar{d}_0(\bar{q}, \bar{r})$, where $\bar{p}$, $\bar{q}$, $\bar{r}$ form a triangle in the reference space $M_0^2 = \mathbb{R}^2$ such that $d(p,q) = \bar{d}_0(\bar{p}, \bar{q})$, $d(p,r) = \bar{d}_0(\bar{p}, \bar{r})$ and $\angle_{\bar{p}}(\bar{q}, \bar{r}) = \angle_p(q,r)$. Then, by the law of cosines one further has $d^2(q,r) \geq \bar{d}_0^2(\bar{q}, \bar{r}) = \bar{d}_0^2(\bar{p}, \bar{q}) - 2\bar{d}_0(\bar{p}, \bar{q})\bar{d}_0(\bar{p}, \bar{r}) \cos \angle_{\bar{p}}(\bar{q}, \bar{r}) + \bar{d}_0^2(\bar{p}, \bar{r}) = d^2(p,q) - 2d(p,q)d(p,r) \cos \angle_p(q,r) + d^2(p,r)$. As the condition (a) implies that $\mathcal{M}$ is a CAT(0) space which is uniquely geodesic, the angles $\angle_p(q,r)$ and $\angle_{\mu(t_i)}(Y_i, q)$ in Proposition 2 are well defined. Verification of the Lipschitz condition (b) is nontrivial for a general Hadamard space. Using various properties of the Hadamard space, we show in Lemma S.1 (Lin and Müller (2021)) that condition (b) holds for all Hadamard spaces with the universal constant $C_1 = 5$. Finally, Lemma S.7 (Lin and Müller (2021)) shows that condition (c) also holds for Hadamard spaces. Consequently, Theorem 2 follows directly from Proposition 2, and Theorem 1 follows as a special case of Theorem 2.

The CAT(0) inequality, which holds for Hadamard spaces, implies the convexity of the distance function, that is,

$$d(\llbracket p, q \rrbracket_\theta, \llbracket p, r \rrbracket_\theta) \leq \theta d(q,r) \quad \text{for all } \theta \in [0, 1] \text{ and all } p, q, r \in \mathcal{M}, \tag{4.2}$$

where $\llbracket p, q \rrbracket_\theta$ denotes the point that sits on the geodesic segment connecting p to q and satisfies $d(p, \llbracket p, q \rrbracket_\theta) = \theta d(p,q)$. This convexity is used to bound the total variation of the geodesically interpolated functions $\tilde{\gamma}_\theta(t) = \llbracket \mu(t), \gamma(t) \rrbracket_\theta$ by the total variation of the functions μ and γ; see Section S.1 of the Supplementary Material (Lin and Müller (2021)). We provide an overview of the main steps of the proof of Proposition 2 demonstrating how it relies on new geometric ideas that are introduced here to establish this key result in Appendix B, while the detailed steps of the proof are provided in Section S.1 of the Supplementary Material.

In the following, we discuss three pertinent examples which will also be further investigated in simulations and data applications.

EXAMPLE 1 (Symmetric positive-definite matrices). Symmetric positive-definite (SPD) matrices as random objects arise in many applications that include computer vision (Rathi, Tannenbaum and Michailovich (2007)), medical imaging (Fillard et al. (2005), Arsigny et al. (2006), Pennec, Fillard and Ayache (2006), Fletcher and Joshi (2007), Dryden, Koloydenko and Zhou (2009)) and neuroscience (Friston (2011)). For example, diffusion tensor imaging, which is commonly used to obtain brain connectivity maps based on magnetic resonance imaging (MRI), produces 3 × 3 SPD matrices that characterize the local diffusion (Zhou et al. (2016)). For the space of $m \times m$ SPD matrices, denoted by $\mathrm{Sym}_\star^+(m)$, the Euclidean distance function $d_E(A, B) = \|A - B\|_F$ that is based on the Frobenius norm $\|\cdot\|_F$ suffers from the so-called swelling effect: The determinant of the average SPD matrix is larger than any of the individual determinants (Arsigny et al. (2007)). Rectifying this issue motivates the use of more sophisticated distance functions, such as the Log-Euclidean distance $d_{\mathrm{LE}}(A, B) = \|\log A - \log B\|_F$ (Arsigny et al. (2007)), the affine-invariant distance $d_{\mathrm{AI}}(A, B) = \|\log(A^{-1/2} B A^{-1/2})\|_F$ (Moakher (2005), Pennec, Fillard and Ayache (2006)) or the Log-Cholesky distance (Lin (2019)), where $\log A$ is the matrix logarithm of A. Either of the above distance functions is indeed induced by a Riemannian metric tensor that turns

$\mathrm{Sym}^+_\star(m)$ into a complete and simply connected Riemannian manifold of nonpositive and bounded sectional curvature. Therefore, Theorem 1 applies to this case.

EXAMPLE 2 (Wasserstein space $\mathcal{W}_2(\mathbb{R})$). Let $\mathcal{W}_2(\mathbb{R})$ be the space of probability distributions on the real line $\mathbb{R}$ and with finite second moments, equipped with the Wasserstein distance $d_W(G_1, G_2) = [\int_0^1 \{G_1^{-1}(s) - G_2^{-1}(s)\}^2 \,\mathrm{d}s]^{1/2}$, where G_1^{-1} and G_2^{-1} are the (left continuous) quantile functions corresponding to distribution functions G_1 and G_2. According to Proposition 4.1 of Kloeckner (2010), $\mathcal{W}_2(\mathbb{R})$ is a CAT(0) space. As $\mathcal{W}_2(\mathbb{R})$ inherits the completeness of $\mathbb{R}$, $\mathcal{W}_2(\mathbb{R})$ is also a Hadamard space. We illustrate the utility of $\mathcal{W}_2(\mathbb{R})$ for data analysis in a study of mortality profiles in Section 6.1. As in the proof of Proposition 1 of Petersen and Müller (2019), one can show that $\sup_{G\in\mathcal{W}_2(\mathbb{R})} \log N(\epsilon\delta, B_G(\delta), d_W) \le K\epsilon^{-1}$ for a constant K and all $\delta, \epsilon > 0$, where $B_G(\delta) = \{\tilde{G} \in \mathcal{W}_2(\mathbb{R}) : d_W(G, \tilde{G})) \le \delta\}$. Then, for the function class $\mathscr{G}$ of Lipschitz continuous $\mathcal{W}_2(\mathbb{R})$-valued functions defined on $\mathcal{T}$, using Proposition 3 in Appendix D, we can establish condition (H2), and therefore the rate in Theorem 2 applies. It is worth noting that $\mathcal{W}_2(\mathbb{R}^m)$ is not a Hadamard space for $m \ge 2$ (Kloeckner (2010), Section 4), so that Theorem 2 does not apply.

EXAMPLE 3 (Phylogenetic trees). Phylogenetic trees are central data objects in the field of evolutionary biology, where they are used to represent the evolutionary history of a set of organisms. In a seminal paper by Billera, Holmes and Vogtmann (2001), phylogenetic trees with m leaves are modeled by metric m-trees endowed with a metric that turns the space of phylogenetic m-trees into a metric space, as follows. A leaf is a vertex that is connected by only one edge, and a metric m-tree is a tree with m uniquely labeled leaves and positive lengths on all interior edges, where an edge is called an interior edge if it does not connect to a leaf. A collection of m-trees that have the same tree structure (taking leaf labels into account) but different edge lengths can be identified with the orthant $(0, \infty)^r$, where r (determined by the tree structure) is the number of interior edges of each tree in the collection. Collections of different tree structures, identified by different orthants, can be glued together along the common faces of the orthants. With this identification between points and metric m-trees, a natural distance function d_T on the space $\mathscr{T}_m$ of all metric m-trees is defined in the following way: For two trees in the same orthant, their distance is the Euclidean distance, while for two trees from different orthants, their distance is the minimum length over all paths that connect them and consist of only connected segments, where a segment is a straight line within an orthant. According to Lemma 4.1 of Billera, Holmes and Vogtmann (2001), the space $(\mathscr{T}_m, d_T)$ is a CAT(0) space. In addition, as a cubical complex, by Theorem 1.1 of Bridson (1991) it is also a complete metric space and thus a Hadamard space. For a fixed m, from the construction of $\mathscr{T}_m$, one can see that the covering number $N(\epsilon\delta, B_x(\delta), d_T)$ for the ball $B_x(\delta)$ centered at $x \in \mathscr{T}_m$ and with radius δ is of the same order as the covering number of the unit ball of a finite-dimensional Euclidean space, which is $O(\epsilon^{-k})$ for a $k = k(m) \ge 1$. For the function class $\mathscr{G}$ of $\mathscr{T}_m$-valued Lipschitz continuous functions, using Proposition 3 in Appendix D, one finds that the condition (H2) holds for $\mathscr{T}_m$ and $\mathscr{G}$. Therefore, Theorem 2 applies to this case.

4.2. *Extension to Alexandrov spaces.* The development of our main results crucially depends on the convexity of the Hadamard space, characterized by condition (a) of Proposition 2, which is shown to be equivalent to the CAT(0) inequality and implies the convexity (4.2) of the distance function of the Hadamard space. By examining the proofs of Proposition 2 and Lemma S.5 in the Supplementary Material (Lin and Müller (2021)), one finds that condition (a) can be relaxed to

$$d^2(q, r) \ge d^2(p, q) - 2d(p, r)d(p, q)\cos\angle_p(q, r) + cd^2(p, r) \tag{4.3}$$

for a universal constant $c > 0$, where we note that $c = 1$ for Hadamard spaces. It turns out that inequality (4.3) holds for some subspaces of Alexandrov spaces with positive lower and upper bounded curvature, and thus our main results potentially carry over to such subspaces.

Another key ingredient is the strong convexity of the squared distance function of a Hadamard space. A real-valued function f defined on a convex subset of $\mathbb{R}^k$ is strongly convex with parameter $\eta > 0$ if $f((1-\theta)p + \theta q) \le (1-\theta)f(p) + \theta f(q) - \eta\theta(1-\theta)\|p-q\|^2$ for all p, q in the convex subset and $\theta \in [0, 1]$. To generalize this concept to functions with geodesic-metric-space valued arguments, we observe that the convex combination $(1-\theta)u + \theta v$ lies on the straight line connecting u and v, and is conveniently replaced with a point on the geodesic connecting p and q. Specifically, we refer to a function f defined on a geodesically convex subset $\mathcal{C}$ of a geodesic space as a strongly convex function on $\mathcal{C}$ with parameter $\eta > 0$ if $f([\![p, q]\!]_\theta) \le (1-\theta)f(p) + \theta f(q) - \eta\theta(1-\theta)d^2(p, q)$ for all $p, q \in \mathcal{C}$ and $\theta \in [0, 1]$, where a subset in a geodesic space is geodesically convex if for any two points in the subset there exists a unique geodesic contained within the subset that connects those two points. One of the nice properties of strongly convex functions is the existence and uniqueness of a minimizer on a geodesically convex closed subspace when the function is continuous (Sturm (2003), Proposition 1.7). For any fixed element q of a Hadamard space, the function $f(\cdot) = d^2(\cdot, q)$ that is defined on this space is continuous and strongly convex with parameter $\eta = 1$ (Bačák (2015), equation (2)). This implies the strong convexity of the Fréchet function $F(\cdot) = \mathbb{E}d^2(\cdot, Y)$, whence the Fréchet mean of a random object on a Hadamard space always exists and is unique provided that the Fréchet function is finite. For specific Alexandrov spaces, the squared distance function shares the property of being strongly convex over some geodesically convex subspaces; see Example 4 below.

Utilizing strong convexity and the relaxed condition (4.3) makes it possible to extend the main results in Section 4.1 to certain Alexandrov spaces. Let $\mathcal{M}$ be an Alexandrov space with positive lower and upper bound on curvature, where the upper bound is denoted by κ. The space $\mathcal{M}$ generally has a finite diameter, according to Theorem 1.9 of Petrunin and Tuschmann (1999). Consequently, the sub-Gaussianity condition (H1) is automatically satisfied for all random objects in $\mathcal{M}$. We need the following additional assumptions.

(A1) There exists $Q > 0$ such that $\log N(\delta, \mathscr{G}^r_{\mathcal{M}}(r), d_n) \le rQ\delta^{-1}$ for all $\delta > 0$ and $r > 0$.

(A2) There exists a geodesically convex closed subset $\mathcal{C} \subset \mathcal{M}$ of diameter less than $\pi/(2\sqrt{\kappa})$ such that:

(A2a) $Y_i \in \mathcal{C}$ for all $n \ge 1$ and $1 \le i \le n$,

(A2b) the function $h(x) = d^2(x, y)$ is strongly convex with a universal constant $C_2 > 0$ for all $y \in \mathcal{C}$, and

(A2c) $d^2(q, r) \ge d^2(p, q) - 2d(p, r)d(p, q)\cos\angle_p(q, r) + C_3 d^2(p, r) \ge 0$ for a universal constant $C_3 > 0$ and all $p, q, r \in \mathcal{C}$.

The entropy condition (A1) is a simplified version of the condition (H2), as now the space $\mathcal{M}$ is of bounded diameter. The bound on the diameter of the subset $\mathcal{C}$ implies that $\mathcal{C}$ is a uniquely geodesic subset of $\mathcal{M}$ and thus ensures that the angle $\angle_p(q, r)$ in (A2c) is well defined. As previously mentioned, the strong convexity condition (A2b) implies the existence and uniqueness of the Fréchet mean, and (A2c) is a relaxation of condition (a) of Proposition 2. Then, with an argument similar to the proof of Proposition 2, the following holds.

THEOREM 3. *For a family $\mathscr{A}(R, \kappa)$ of positively curved Alexandrov spaces, all of which have a diameter bounded by R and a curvature upper bounded by $\kappa > 0$, with $\lambda \asymp n^{-2/3}$, one has*

$$\lim_{D\to\infty} \limsup_{n\to\infty} \sup_{F\in\mathscr{F}_n} \mathbb{P}_F\{d_n(\hat{\mu}, \mu) > Dn^{-1/3}\} = 0,$$

where μ is defined in (3.1), $\hat{\mu}$ *is given in* (3.2), $\mathbb{P}_F$ *is the probability measure induced by a probability distribution* F, *and* $\mathscr{F}_n = \mathscr{F}_n(R, \kappa, Q, C, C_2, C_3)$ *for constants* R, κ, Q, C, C_2, C_3 *is the collection of joint probability distributions of* $Y_1, \dots, Y_n$ *on* $\mathcal{M}$ *for which* $\mathcal{M} \in \mathscr{A}(R, \kappa)$, $\mu \in \mathscr{G}_{\mathcal{M}}(C)$, *and conditions* (A1)–(A2) *hold.*

EXAMPLE 4 (Time-indexed compositional data). Such data arise in various settings that include longitudinal compositional data (Dai and Müller (2018)). Specifically, for compositional data $Y_i = (z_{i,1}, \dots, z_{i,k+1})$ such that $z_{i,j} \ge 0$ and $\sum_{j=1}^{k+1} z_{i,j} = 1$, $i = 1, \dots, n$, one may apply the square root transformation on each $z_{i,j}$ and view $(\sqrt{z_{i,1}}, \dots, \sqrt{z_{i,k+1}})$ as elements of the quadrant $\mathcal{C} = \{(x_1, \dots, x_{k+1}) \in \mathbb{S}^k : x_j \ge 0 \text{ for } j = 1, \dots, k+1\}$. Compositional data can thus be viewed as sampled from the convex subset $\mathcal{C}$, where the diameter of this quadrant is $\pi/2$. Then, for all $p, q, r \in \mathcal{C}$, whenever $d(q, r) \le c_1 < \pi/2$ for a universal constant $c_1 > 0$, according to the Taylor expansion of the function $h(\cdot) = d^2(\cdot, r)$ at p and its gradient and Hessian (Pennec (2018), Supplement A), we find that (4.3) holds for some universal constant $c = c_2 > 0$ (depending on c_1). In addition, the Hessian of h is positive on $\mathcal{C}$ uniformly for all $r \in \mathcal{C}$, which implies the strong convexity of h. Then condition (A2) is satisfied if $\Pr\{d(\partial\mathcal{C}, Y_i) \ge c_3 \text{ for all } 1 \le i \le n\} = 1$ for a universal constant $c_3 > 0$, where $\partial\mathcal{C}$ is the boundary of $\mathcal{C}$ and $d(\partial\mathcal{C}, p)$ is the distance of p to the set $\partial\mathcal{C}$. This requirement corresponds to points being not too close to the boundary of $\mathcal{C}$. This is a mild condition, as c_3 can be arbitrarily small. For the class $\mathscr{G}$ of $\mathbb{S}^k$-valued functions of bounded variation defined on $\mathcal{T}$, applying Proposition 4 in Appendix D, we find that (A1) is also satisfied, and thus Theorem 3 applies.

In the above example, all data are located in a subset that has a diameter less than $\pi/2$ and is thus strictly smaller than a hemisphere. If we allow data to be arbitrarily close to the equator, then the constant c_2 approaches to zero, and thus the convexity conditions in (A2) might be violated. As pointed out by a reviewer, the minimal distance to the equator will play a non-ignorable role, and the convergence rate of Theorem 3 is expected to change in dependence on this minimal distance, along with changing constants C_2 and C_3 in (A2). In the extreme case that all data points are located on the equator, the population Fréchet mean μ may not be uniquely defined and thus the total variation regularized estimator might not converge. Another extreme case is that the expected Hessian vanishes at the Fréchet mean. For this case Eltzner and Huckemann (2019) show that the empirical Fréchet mean may still converge to the population Fréchet mean, but at a slower rate. Whether the regularized estimator proposed here exhibits a similar behavior is of theoretical interest and could be a topic for future research.

5. Simulation studies.

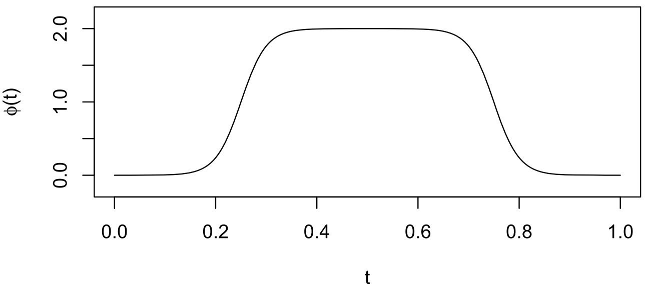
We consider three metric spaces, namely, the SPD matrix space $\mathrm{Sym}_\star^+(m)$ endowed with the affine-invariant distance in Example 1 with $m = 3$, the Wasserstein space $\mathcal{W}_2(\mathbb{R})$ in Example 2, and the Billera–Holmes–Vogtmann space of phylogenetic trees in Example 3. For each of these metric spaces, two settings are examined with $\mathcal{T} = [a, b] = [0, 1]$. In the first setting, the underlying mean functions $\mu(t)$, $t \in \mathcal{T}$, are locally constant, while in the second setting they smoothly vary with $t \in \mathcal{T}$. Further details are given in Table 1. The first setting represents a favorable scenario for total variation regularized Fréchet regression, since the estimator is also locally constant, while the second setting is more challenging.

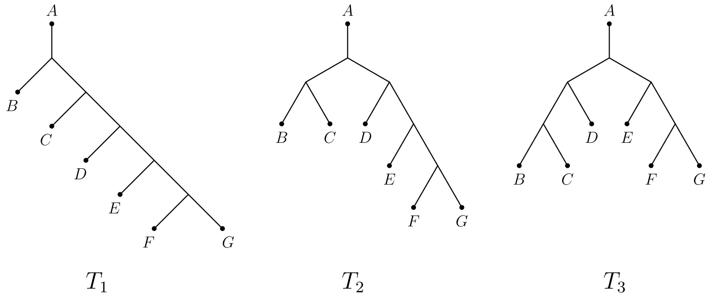
For each setting, we investigated two sample sizes, $n = 50$ and $n = 150$ for the design points $t_i = (i-1)/(n-1)$ with $i = 1, \dots, n$. For the SPD matrix space, data Y_i were generated as $Y_i = \mu(t_i)^{1/2} \exp\{\mu(t_i)^{-1/2} S_i \mu(t_i)^{-1/2}\} \mu(t_i)^{1/2}$ with $\mathrm{vec}(S_i) \overset{i.i.d}{\sim} N(0, 0.25^2 I_6)$,

TABLE 1

The mean functions for the metric spaces and settings considered in the simulation study, where I_3 is the 3×3 identity matrix, $N(\nu, \sigma^2)$ denotes the Gaussian distribution with mean ν and variance σ^2, $\phi(t) = 2(1 + e^{-40(t-0.25)})^{-1}$ if $t \in [0, 0.5)$ and $\phi(t) = 2(1 + e^{40(t-0.75)})^{-1}$ if $t \in [0.5, 1]$, where the included figures depict the function ϕ that is continuous with rapid changes and is used in Setting II, and phylogenetic trees T_1, T_2 and T_3. The length of each edge of these trees is one

	Setting I	Setting II
SPD	$\mu(t) = \begin{cases} I_3 & t \in [0, \frac{1}{3}), \\ 2I_3 & t \in [\frac{1}{3}, \frac{2}{3}), \\ 3I_3 & t \in [\frac{2}{3}, 1] \end{cases}$	$\mu(t) = \{1 + \phi(t)\}I_3$
Wasserstein	$\mu(t) = \begin{cases} N(0, 1) & t \in [0, \frac{1}{3}), \\ N(1, 1.5^2) & t \in [\frac{1}{3}, \frac{2}{3}), \\ N(2, 2^2) & t \in [\frac{2}{3}, 1] \end{cases}$	$\mu(t) = N(\phi(t), \{1 + \phi(t)\}^2)$
Tree	$\mu(t) = \begin{cases} T_1 & t \in [0, \frac{1}{3}), \\ T_2 & t \in [\frac{1}{3}, \frac{2}{3}), \\ T_3 & t \in [\frac{2}{3}, 1] \end{cases}$	$\mu(t) = [\![T_1, T_3]\!]_{\phi(t)/2}$





where $\mu(t)$ is as in Table 1, S_i is a 3×3 symmetric matrix and $\text{vec}(S)$ is its vector representation, that is, the 6-dimensional vector obtained by stacking elements in the lower triangular part of S, and I_6 denotes the 6×6 identity matrix.

For the Wasserstein space, we adopted the method in Petersen and Müller (2019) to generate observations Y_i, as follows. Let $a_i = \mathbb{E}Z$ and $b_i = \{\mathbb{E}(Z - \mathbb{E}Z)^2\}^{1/2}$ for $Z \sim \mu(t_i)$, where again the distributions $\mu(t)$ are as listed in Table 1 for the Wasserstein case. We then first sample $\nu_i \sim N(a_i, 1)$ and $\sigma_i \sim \text{Gamma}(\alpha_1, \alpha_2)$, with shape parameter $\alpha_1 = 0.5b_i^2$ and rate parameter $\alpha_2 = 0.5b_i$. Note that $\mathbb{E}\nu_i = a_i$ and $\mathbb{E}\sigma_i = b_i$. Then Y_i is obtained by transporting the distribution $N(\nu_i, \sigma_i^2)$ by a transport map $\mathscr{T}$ that is uniformly sampled from the collection of maps $\mathscr{T}_k(x) = x - \sin(kx)/|k|$ for $k \in \{\pm 2, \pm 1\}$. Note that Y_i is not a Gaussian distribution due to the transportation. Nevertheless, one can show that the Fréchet mean of Y_i is exactly $\mu(t_i)$.

TABLE 2

Simulation results for average Root Integrated Squared Error (*RISE*) *of the total variation regularized estimators for the fitted versus true functions for the two settings considered and random objects corresponding to symmetric positive definite* (*SPD*) *matrices*, *probability distributions with the Wasserstein metric*, *and phylogenetic trees*. *The standard errors based on* 100 *Monte Carlo replicates are given in parentheses*

	SPD		Wasserstein		Trees	
Setting	$n=50$	$n=150$	$n=50$	$n=150$	$n=50$	$n=150$
I	0.210 (0.057)	0.124 (0.042)	0.516 (0.127)	0.321 (0.064)	0.294 (0.116)	0.209 (0.083)
II	0.256 (0.054)	0.164 (0.041)	0.604 (0.141)	0.372 (0.073)	0.368 (0.131)	0.235 (0.097)

For the case of phylogenetic trees, we generated each Y_i by translating $\mu(t_i)$ along a random geodesic emanating from $\mu(t_i)$ for a random distance that follows the uniform distribution on $[0, 0.5]$. This requires identification and computation of geodesics in the Billera–Holmes–Vogtmann phylogenetic tree space $\mathscr{T}_m$ ($m = 7$ in our setting), for which we employed the algorithm by Owen and Provan (2011).

The regularization parameter λ was chosen by fivefold cross-validation. Specifically, we treated the design points as if they were random, and randomly split the data $\mathcal{D} := \{(t_1, Y_1), \dots, (t_n, Y_n)\}$ into five even partitions $\mathcal{D}_1, \dots, \mathcal{D}_5$. For a given value of λ, for each $k = 1, \dots, 5$, the proposed estimation procedure was applied to $\mathcal{D}\backslash\mathcal{D}_k$ to obtain an estimator $\hat{\mu}_{-k}$. The cross-validation error for the given λ was calculated by $\sum_{k=1}^{5} \sum_{(t,Y)\in\mathcal{D}_k} d^2(\hat{\mu}_{-k}(t), Y)$, and the value of λ minimizing the cross-validation error was selected. The results are based on 100 Monte Carlo runs. The estimation quality of $\hat{\mu}$ is quantified by the root integrated squared error (RISE)

$$\mathrm{RISE}(\hat{\mu}) = \left\{\int_{\mathcal{T}} d_{\mathcal{M}}^2(\hat{\mu}(t), \mu(t))\,\mathrm{d}t\right\}^{1/2}.$$

The results in Table 2 indicate that as sample size grows, the estimation error decreases in both the favorable setting and the challenging setting. Moreover, we observe that the decay rate of the empirical RISE in the table, defined as the ratio of the RISE with $n = 150$ and the RISE with $n = 50$, is approximately 0.62. This seems to agree quite well with our theory in Section 4 that suggests a rate of $(50/150)^{1/3} \approx 0.69$.

6. Applications.

6.1. *Mortality.* We applied the proposed method to analyze the evolution of the distributions of age-at-death using mortality data from the Human Mortality Database at www.mortality.org. The database contains yearly mortality for 37 countries, grouped by age from 0 to 110+. Specifically, the data provide a lifetable with a discretization by year, which can be easily converted into a histogram of age-at-death, one for each country and calendar year. Starting from these fine-grained histograms, a simple smoothing step then leads to the density function of age-at-death for a given country and calendar year. We focus on the adult (age 18 or more) mortality densities of Russia and the calendar years from 1959 to 2014. The time-indexed densities of age-at-death are shown in the form of a heat map in Figure 1(a) for males and for females in Figure 2(a). The patterns of mortality for males and females are seen to differ substantially.

Applying the proposed total variation regularized Fréchet regression for distributions as random objects with the Wasserstein distance to these data, we employ a fine grid on the interval $[10^{-2.5}, 10^{-0.1}]$ and use the aforementioned fivefold cross validation to select the

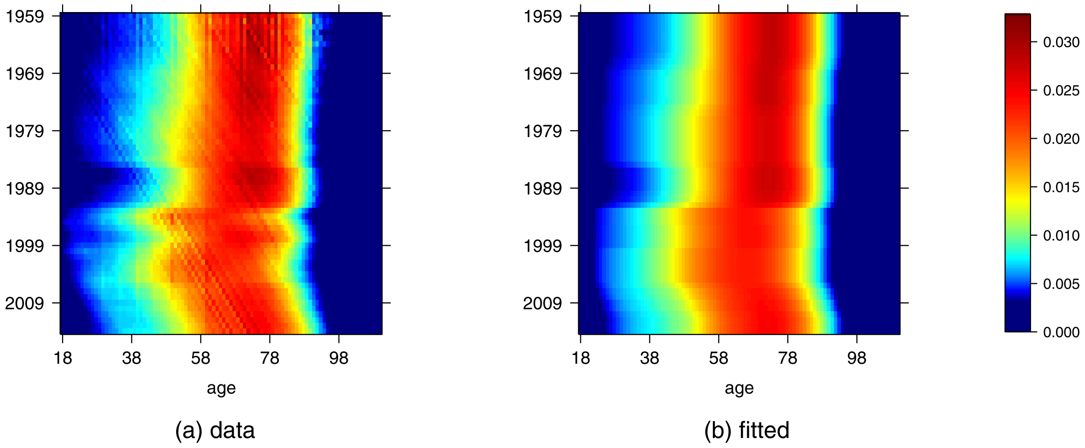


FIG. 1. *Total variation regularized Fréchet regression for time-indexed mortality distributions of males in Russia, where panel* (*a*) *displays the raw yearly mortality density functions, and panel* (*b*) *the fitted densities obtained with total variation regularization.*

regularization parameter λ. The selected values are $\lambda = 10^{-1.5}$ and $\lambda = 10^{-1.7}$ for males and females, and the resulting estimates are shown in Figure 1(b) and Figure 2(b), respectively.

This suggests that the proposed total variation regularized Fréchet estimator adapts well to the smoothness of the target function. For example, the female mortality dynamics is seen to be relatively smooth, and the estimator accordingly is also quite smooth. In contrast, male age-at-death distributions exhibit sharp shifts; the proposed estimator reflects this well and preserves the discontinuities in the mortality dynamics. This demonstrates desirable flexibility of total variation regularized Fréchet regression, as it appropriately reflects relatively smooth trajectories, while at the same time preserving edges/boundaries when present. This flexibility has been documented previously for the Euclidean case (Strong and Chan (2003)), and is shown here to extend to the much more complex case of metric-space valued data.

Specifically, a major shift in mortality distributions occurred around 1992 and is well represented in the estimates for both males and females, with a much larger shift for males. The direction of the shift was towards increased mortality for both males and females, as the age-at-death distributions moved left, implying increased mortality at younger ages. A weaker

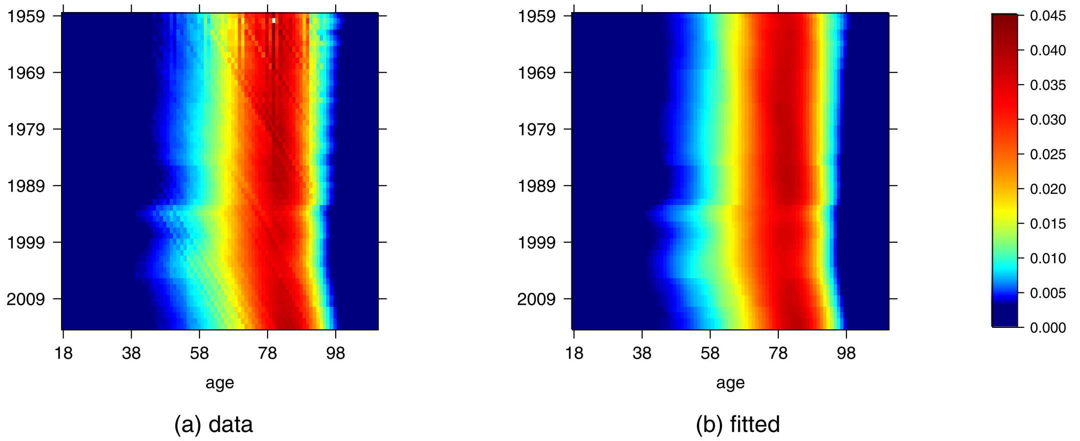


FIG. 2. *Total variation regularized Fréchet regression for time-indexed mortality distributions of females in Russia, where panels* (*a*) *and* (*b*) *are as in Figure* 1.

shift that occurred in 2008 is also captured by the estimator for both males and females, and again is more expressed for males. This latter shift was towards decreased mortality.

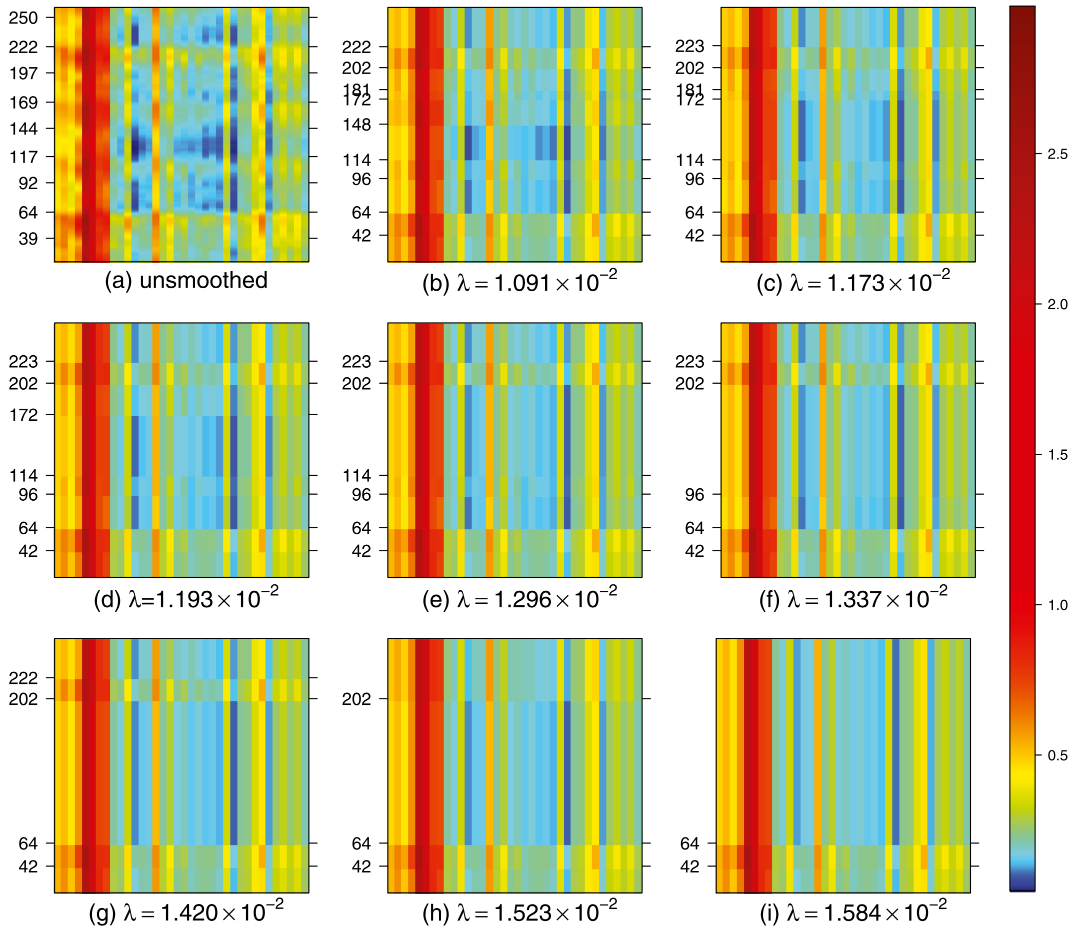
These findings pinpoint a period from 1992–2008, during which the turmoil following the collapse of the Soviet Union 1988–1991 appears to have had devastating impacts on mortality. The strong shift in 1992 is relatively easy to explain with social ills such as increased alcoholism and joblessness that followed the collapse of the Soviet Union; it affected males more than females.

6.2. *Functional connectivity.* We applied the proposed total variation regularization method for random objects also to data on functional connectivity in the human brain from the Human Connectome Project (Essen et al. (2013)) that were collected between 2012 and 2015. Out of 970 subjects in the study, for 850 subjects social cognition task related fMRI data are available. In this study, each participant was sequentially presented with five short video clips while in a brain scanner, which recorded a fMRI signal. Each clip showed squares, circles and triangles that either interacted in a certain way or moved randomly. The fMRI signals were recorded at 274 time points spaced 0.72 seconds apart. The starting times for the five video clips are approximately at time points 11, 64, 117, 169 and 222, respectively, with ending times approximately at time points 39, 92, 144, 197 and 250, respectively, so there are overall 10 time points where the nature of the visual input is changing. A natural question is then whether changes in brain connectivity, as quantified by fMRI signals, are associated with the above time points that indicate changes in visual input. To address this question, we estimated the changes through total variation regularized Fréchet regression without using knowledge about the video clip switch times. As described in Appendix C, we selected 8 brain regions and applied a preprocessing pipeline to obtain the observations $Y_i \in \mathrm{Sym}_\star^+(8)$ at each time point t_i, for $i = 1, \dots, n = 243$, which are depicted in Figure 3(a), where for illustration purposes each SPD matrix has been vectorized into an $8(8+1)/2 = 36$ (taking symmetry into account) dimensional vector represented by a row in the heat map, indicating the relative values of the vector elements.

This SPD sequence is quite noisy and does not clearly indicate whether the mean brain connectivity changes in accordance with the transition points of the visual input as described above. Thus, to gain insight whether the pattern of brain connectivity follows the pattern of visual inputs, it is necessary to denoise these data. Assuming constant brain connectivity while the visual input is constant (video on or off), this motivates the fitting of locally constant functions with a few knots for SPD random objects and thus the application of the proposed total variation regularized Fréchet regression. This is due to the fact that the proposed estimator $\hat{\mu}$ can be viewed as a locally constant function in time with adaptive knot placement, mapping time into metric space, in our case the space of SPD matrices.

When applying total variation regularized Fréchet regression, one has to select the regularization parameter λ. Generally, we recommend to use the aforementioned cross-validation procedure. However, in the particular application at hand, since we may assume that the number of jumps (the discontinuous points of μ) is known to be $J = 10$, we can simply choose the smallest value of λ that yields $\hat{J} = 10$ jumps of $\hat{\mu}$. Due to the choice of $P = 16 > 11$ for computing Y_i in Appendix C, the sequence does not contain sufficient information about the start time point of the first video clip, which is $t = 11$. Therefore, we target $J = 9$ and choose the smallest value of λ that yields $\hat{J} = 9$.

Practically, we performed the proposed total variation regularization for the SPD case on the sequence Y_i for different choices of the regularization parameter λ on a fine grid within the interval $[0.01, 0.02]$. Panels (b)–(i) in Figure 3 display the resulting estimates by using



(a) unsmoothed

(b) $\lambda = 1.091 \times 10^{-2}$

(c) $\lambda = 1.173 \times 10^{-2}$

(d) λ=1.193×10^{-2}

(e) $\lambda = 1.296 \times 10^{-2}$

(f) $\lambda = 1.337 \times 10^{-2}$

(g) $\lambda = 1.420 \times 10^{-2}$

(h) $\lambda = 1.523 \times 10^{-2}$

(i) $\lambda = 1.584 \times 10^{-2}$

FIG. 3. *Total variation regularized Fréchet regression for dynamic functional connectivity derived from fMRI data. Time is on the vertical axis and the times where visual input changes are explicitly indicated by the tick labels in Panel* (*a*). *The lower triangular portions of the SPD covariance matrices of brain connectivity are shown in vectorized form along the horizontal axis. Panel* (*a*) *depicts the raw empirical functional connectivity and panels* (*b*)–(*i*) *depict fitted connectivity, obtained by applying the proposed total variation regularized Fréchet regression. From each panel to the next, the regularization parameter is successively increased such that the number of jump points decreases by one. For each of the panels* (*b*)–(*i*), *the tick labels on the left side indicate the locations of the jump points of the fitted step function. In* (*b*) *and* (*c*), *labels for time points* 172 *and* 181 *are overlapping.*

the affine-invariant distance (Moakher (2005), Pennec, Fillard and Ayache (2006)); results by using the Log-Euclidean distance (Arsigny et al. (2007)) are similar. For each panel, the minimal value of the regularization parameter λ was chosen so that the number of jump points ranged from 9 (smaller λ) to 2 (larger λ), respectively. From Figure 3(b), where one has 9 jump points of $\hat{\mu}$, we find that the detected jump points closely match the times when the videos clips started and ended, with the exception of time points 11 and 250, which is due to insufficient data between these first and last events and the respective boundaries, and the event at time point 197, which is split into two jump points, at time points 181 and 202. As λ increases, the number of jump points of the estimates decreases. Further discussion can be found in Appendix C.

7. Concluding remarks. The theoretical developments of the paper are rooted in convexity properties of Hadamard spaces, which provide key ingredients for establishing the minimax convergence rate for the class of Hadamard spaces. The minimax convergence rate is achieved for the one-dimensional Euclidean space, which is a special case of a Hadamard space. As suggested by a reviewer, in light of the work Hotz et al. (2013) which shows that the sample Fréchet mean converges to its population counterpart at a rate faster than $n^{-1/2}$ in some negatively curved spaces, an interesting future topic is to investigate whether a convergence rate faster than $n^{-1/3}$ is possible for our estimator in some Hadamard spaces of strictly negative curvature. The convexity also entails an extension to some subspaces of positively curved Alexandrov spaces, where distance functions are strongly convex over the subspaces as per condition (A2b). However, a comprehensive treatment for the case of Alexandrov spaces is substantially more challenging, as seen in Example 4 and the related discussion. This requires a theory beyond convexity and thus falls outside of the scope of this paper.

Extensions to multivariate or manifold-valued domains are also interesting and nontrivial. For multivariate domains, one promising direction is to extend the Hardy–Krause total variation that is utilized by Fang, Guntuboyina and Sen (2021) for multidimensional total variation regularization, since it extends to the multidimensional case most of the features of the one-dimensional case, for example, given by the supremum over partitions. Other interesting and important topics to explore in the future include finite risk bounds and sharp oracle inequalities for the estimated regression function under the setting of Section 4, complementing the asymptotic theory developed in this paper; sharp bounds are very challenging in this setting due to the limited geometric and analytic structure that is available in general metric spaces.

Reviewers have pointed out that the entropy condition (H2) is local in nature, in the sense that the constant K might depend on R and it holds only for all $r \le R$ for an arbitrary but fixed $R > 0$ and that if the entropy condition were global, that is, $\log N(r\delta, \mathscr{G}_{\mathcal{M}}^{r}(r), d_n) \le K\delta^{-1}$ for all $r, \delta > 0$, the proof of Proposition 2 could be simplified by using a strategy of van de Geer (2001), where the constant C might also vary with sample size n. It remains however unclear how such a global entropy condition can be verified for the class of general metric-space valued functions of bounded variation. Even for more specific metric spaces such as Riemannian manifolds, Proposition 4 in the Supplementary Material (Lin and Müller (2021)) suggests that the curvature effect plays an important role in the metric entropy bound. More precise results are left for future study.

An alternative way to allow C to vary with n in Proposition 2, suggested by a reviewer, is to exploit convexity as in Chinot, Lecué and Lerasle (2020), where one does not require a metric entropy condition. However, Chinot, Lecué and Lerasle (2020) and the related work Alquier, Cottet and Lecué (2019) require the concept of Gaussian mean width to characterize complexity of the class of functions under consideration, which is indirectly connected to metric entropy, for example, via Sudakov's inequality (Ledoux and Talagrand (2011), Theorem 3.18) and Dudley's inequality (Ledoux and Talagrand (2011), Theorem 11.17). Generalization of Gaussian mean width to metric-space valued functions and determining its precise relation with metric entropy is another challenging and interesting topic for future exploration.

APPENDIX A: COMPUTATIONAL DETAILS

To compute the total variation regularized estimator defined in (3.2), we adopt a simplified version of the cyclic proximal point algorithm proposed by Weinmann, Demaret and Storath (2014). To find the step function estimator according to Proposition 1, noting that

$\mathrm{TV}(\hat{\mu}) = \sum_{i=1}^{n-1} d(\hat{\mu}(t_i), \hat{\mu}(t_{i+1}))$, it is sufficient to compute $\hat{\mu}_i \equiv \hat{\mu}(t_i)$ for $i = 1, \dots, n$. This is achieved by minimizing the function

$$\tilde{L}_\lambda(p_1, \dots, p_n) = \frac{1}{2}\sum_{i=1}^{n} d^2(p_i, Y_i) + \frac{n\lambda}{2}\sum_{j=1}^{n-1} d(p_j, p_{j+1})$$

over the product space $\mathcal{M}^n$. For $\mathbf{p} = (p_1, \dots, p_n)$ and $G(\mathbf{p}) = \sum_{i=1}^{n} d^2(p_i, Y_i)$, the family of proximal mappings of G is defined by

$$\operatorname{prox}_{\alpha G}\mathbf{p} = \operatorname*{arg\,min}_{\mathbf{q}\in\mathcal{M}^n}\left(\alpha G(\mathbf{q}) + \frac{n}{2}d_n^2(\mathbf{p}, \mathbf{q})\right),$$

where $\alpha > 0$ is a parameter and $d_n^2(\mathbf{p}, \mathbf{q}) = n^{-1}\sum_{i=1}^{n} d^2(p_i, q_i)$. It is easy to check that the kth component of $\operatorname{prox}_{\alpha G}\mathbf{p}$ is $[\![p_k, Y_k]\!]_\theta$ with $\theta = \alpha(1 + \alpha)^{-1}$, where we recall that $[\![p, q]\!]_\theta$ denotes the point sitting on the geodesic segment connecting p and q that satisfies $d(p, [\![p, q]\!]_\theta) = \theta d(p, q)$.

For the proximal mappings of the function $H_j(\mathbf{p}) = d(p_j, p_{j+1})$, given by

$$\operatorname{prox}_{\alpha H_j}\mathbf{p} = \operatorname*{arg\,min}_{\mathbf{q}\in\mathcal{M}^n}\left(\alpha H_j(\mathbf{q}) + \frac{n}{2}d_n^2(\mathbf{p}, \mathbf{q})\right),$$

one finds that if $k \neq j, j + 1$, then the kth component of $\operatorname{prox}_{\alpha H_j}\mathbf{p}$ is equal to p_k. It is shown in Weinmann, Demaret and Storath (2014) that the jth component of $\operatorname{prox}_{\alpha H_j}\mathbf{p}$ is given by $[\![p_j, p_{j+1}]\!]_\theta$, while the $(j + 1)$th component is $[\![p_{j+1}, p_j]\!]_\theta$, where $\theta = \min\{\alpha/d(p_j, p_{j+1}), 1/2\}$ and that the algorithm converges to the minimizer of $\tilde{L}_\lambda$ for Hadamard spaces.

The computational details are summarized in Algorithm 1, where the symbol := denotes the assignment or update operator, evaluating the expression on the right-hand side and then assigning the value to the variable on the left-hand side.

Algorithm 1 Cyclic Proximal Point Algorithm for Total Variation Regularized Fréchet Regression

```
Require: $\alpha_1, \alpha_2, \dots$ such that $\sum_{k=1}^{\infty}\alpha_k^2 < \infty$ and $\sum_{k=1}^{\infty}\alpha_k = \infty$
for $i = 1, \dots, n$ do
    $\hat{\mu}_i := Y_i$
end for
for $r = 1, 2, \dots$ do
    for $i = 1, \dots, n - 1$ do
        $\theta := \frac{\alpha_r}{1+\alpha_r}$
        $\hat{\mu}_i := [\![\hat{\mu}_i, Y_i]\!]_\theta$
    end for
    for $j = 1, \dots, n - 1$ do
        $\theta := \min\{\alpha_r\lambda n/\{2d(\hat{\mu}_j, \hat{\mu}_{j+1})\}, 1/2\}$
        $\hat{\mu}'_j := [\![\hat{\mu}_j, \hat{\mu}_{j+1}]\!]_\theta$ and $\hat{\mu}'_{j+1} := [\![\hat{\mu}_{j+1}, \hat{\mu}_j]\!]_\theta$
        $\hat{\mu}_j := \hat{\mu}'_j$ and $\hat{\mu}_{j+1} := \hat{\mu}'_j$
    end for
end for
Output: $\hat{\mu}(t_i) = \hat{\mu}_1, \dots, \hat{\mu}(t_n) = \hat{\mu}$
```

APPENDIX B: KEY STEPS AND GEOMETRY IN THE PROOFS

To prove Proposition 2, we develop novel geometric arguments that make it feasible to extend arguments of Mammen and van de Geer (1997) to metric-space valued random objects. We first outline how key arguments used in Mammen and van de Geer (1997) (rephrased in our context in terms of language and notations) can be modified to connect them to the core ideas of our geometric constructions, and then provide a detailed proof of Proposition 2 in the Supplementary Material (Lin and Müller (2021)). In this section, the referenced lemmas and equations with labels prefixed by "S" are described in Lin and Müller (2021).

There are three key steps in the proofs of Theorems 9 and 10 of Mammen and van de Geer (1997) that were deployed to study the total variation regularized regression for the traditional situation where $\mathcal{M} = \mathbb{R}$. Once these steps have been identified and established, the rest of the proof of Mammen and van de Geer (1997) is standard. However, these key steps were geared to the linear structure and analytic properties of $\mathbb{R}$, and there is no possibility to modify them for situations without Euclidean structure. To provide versions for general Hadamard spaces is a serious challenge that we tackle in this paper. To overcome the technical hurdles, we need to leverage the convexity of Hadamard spaces to obtain geometric versions of these key steps, as follows.

The first key ingredient is the decomposition of μ into two orthogonal parts by projecting into a space of polynomials and its orthogonal complement. These two parts are handled separately. The complement part is uniformly bounded whenever $\mathrm{TV}(\mu) < C$. The problem is then transformed into estimating a uniformly bounded $\mathbb{R}$-valued function μ (here we reuse the symbol μ to conform to the notation used in our proof) with $\mathrm{TV}(\mu) < C$ via total variation regularization. For Hadamard spaces, such projections and the space of polynomials do not exist. To circumvent this difficulty, we introduce the concept of *center* of an $\mathcal{M}$-valued function γ, which can be characterized as a Fréchet integral (Petersen and Müller (2016), Dubey and Müller (2020b)) and is defined to be the minimizer of the function $F_\gamma(p) = \int_{\mathcal{T}} d^2(p, \gamma(t))\,\mathrm{d}t$ over $\mathcal{M}$, if $F_\gamma(p) < \infty$ for some $p \in \mathcal{M}$. Its discrete version, the center of γ at $t_1, \dots, t_n$, is the minimizer of $F_{\gamma,n}(p) = n^{-1}\sum_{i=1}^n d^2(p, \gamma(t_i))$. Instead of projection, we show that the centers of $\hat{\mu}$ and μ at $t_1, \dots, t_n$ are close to each other in Lemma S.6. Consequently, we can restrict our focus on functions whose center is close to the center of μ. This makes it possible to bypass the decomposition of μ and $\hat{\mu}$. Note that $\mu(t_1), \dots, \mu(t_n)$ themselves are the *centers* of the observed data. The center of μ is then the center of these centers.

A second key ingredient in Mammen and van de Geer (1997) is the inequality $d_n^2(\mu, \hat{\mu}) \le \lambda\{\mathrm{TV}(\mu) - \mathrm{TV}(\hat{\mu})\} + 2n^{-1}\sum_{i=1}^n \varepsilon_i\{\hat{\mu}(t_i) - \mu(t_i)\}$, where $\varepsilon_i = Y_i - \mu(t_i)$. In Hadamard spaces, neither the ε_i nor the differences $\hat{\mu}(t_i) - \mu(t_i)$ or the products $\varepsilon_i\{\hat{\mu}(t_i) - \mu(t_i)\}$ exist, as these notions are all intimately tied to an underlying Euclidean structure that is not present in metric spaces. To address this challenge, we first use the convexity condition (a) in Proposition 2 to obtain a similar inequality. A key step is then to replace the products $\varepsilon_i\{\hat{\mu}(t_i) - \mu(t_i)\}$ with $d(\mu(t_i), \hat{\mu}(t_i))d(\mu(t_i), Y_i)\cos\angle_{\mu(t_i)}(Y_i, \hat{\mu}(t_i))$, which we refer to as Alexandrov inner product in this paper, and to replace the assumption of zero mean errors $\mathbb{E}\varepsilon_i = 0$ with the characterization of Fréchet means in (S.13). These concepts have not been studied previously to the knowledge of the authors and are likely of more general interest.

The third key ingredient is the observation that the function $\hat{\mu} - \mu$, after being scaled by $\mathrm{TV}(\hat{\mu}) + C$, has total variation bounded by a constant, that is, $\mathrm{TV}((\hat{\mu} - \mu)/(\mathrm{TV}(\hat{\mu}) + C)) \le 1$. This eventually enables one to use Lemma 3.5 of van de Geer (1990) for the function $(\hat{\mu} - \mu)/(\mathrm{TV}(\hat{\mu}) + C)$ in order to bound the term $n^{-1}\sum_{i=1}^n \varepsilon_i\{\hat{\mu}(t_i) - \mu(t_i)\}$ by $d^{1/2}(\hat{\mu}, \mu)(\mathrm{TV}(\hat{\mu}) + C)^{1/2}O_P(n^{-1/2})$. Then the rate of $\hat{\mu}$ can be derived by a standard argument that combines this with the inequality obtained for the second key ingredient. In our context, it is difficult to find a geometric counterpart of $\mathrm{TV}((\hat{\mu} - \mu)/(\mathrm{TV}(\hat{\mu}) + C))$ as this involves subtraction and scaling of functions, which are not available in non-Euclidean spaces. To overcome this hurdle, we propose the new idea of geodesic interpolation $\tilde{\gamma}_\theta$ between two functions μ and γ, defined by $\tilde{\gamma}_\theta(t) = [\![\mu(t), \gamma(t)]\!]_\theta$ for $\theta \in [0, 1]$. Then the convexity (S.8) suggests $d(\tilde{\gamma}_\theta(s), \tilde{\gamma}_\theta(t)) \le \theta d(\gamma(s), \gamma(t)) + (1 - \theta)d(\mu(s), \mu(t))$ and further $\mathrm{TV}(\tilde{\gamma}_\theta) \le \theta\,\mathrm{TV}(\gamma) + (1 - \theta)\,\mathrm{TV}(\mu)$. In other words, the total variation of the interpolated function $\tilde{\gamma}_\theta$ is bounded by the convex combination of the total variations of μ and γ. If we set $\theta = C/\{\mathrm{TV}(\gamma) + C\}$, then $\mathrm{TV}(\tilde{\gamma}_\theta) \le 2C$ when $\mathrm{TV}(\mu) \le C$. In particular, this interpolation preserves the closeness of the centers, that is, according to Lemma S.4, if the center of γ is close to the center of μ, then the center of $\tilde{\gamma}_\theta$ is also close to the center of μ. Thus the interpolation simultaneously mimics the subtraction and scaling of $\mathbb{R}$-valued functions. This is again a general principle that we expect to be useful for other investigations where one requires a metric-space counterpart of a standardization procedure that involves function subtraction and scaling.

We note that in order to establish the closeness of the centers in Lemma S.6, we first establish a suboptimal rate for $\hat{\mu}$ in Lemma S.5 using last two ideas in the above described key ingredients. This is made possible by Lemma S.3, where we use the sub-Gaussianity condition (H1) and convexity of the Hadamard space to show that, with probability tending to one, the image of $\hat{\mu}$ is encompassed by a ball centered at the center of μ with radius of the order $\log n$. As the proof of Proposition 2 depends on Lemma S.5 and several other proofs are similar, to avoid repetition, we provide details about the implementation of the above described ideas mainly in the proof of Lemma S.5, and in the proof of Proposition 2 those additional details that are genuinely different from those developed for the proof of Lemma S.5.

APPENDIX C: FURTHER DETAILS ON THE APPLICATION TO BRAIN CONNECTIVITY

Data preprocessing. We divided the brain into 68 regions of interest based on the "Desikan–Killiany" atlas (Desikan et al. (2006)) and picked eight possible regions that are related to social skills, that is, the left and right part of superior temporal, inferior parietal, temporal pole and precuneus (Green, Horan and Lee (2015)). The dynamics of functional connectivity for each subject is represented by the changing nature of the cross-covariance between these eight regions, computed by a moving local window that includes $2P$ time points. Specifically, denoting by V_{ij} the vector of the BOLD (blood-oxygen-level dependent) fMRI signals of the jth subject at the ith time point, the connectivity at $i = P + 1, 18, \dots, 274 - P + 1$ is computed by

$$\Sigma_{ij} = \frac{1}{P} \sum_{k=i-P}^{i+P-1} (V_{kj} - \bar{V}_{ij})(V_{kj} - \bar{V}_{ij})^T \quad \text{with } \bar{V}_{ij} = \frac{1}{P} \sum_{k=i-P}^{i+P-1} V_{kj}.$$

In a last preprocessing step, we aggregated the information at the same time point across all subjects by computing

$$Y_i = \operatorname*{arg\,min}_{\Sigma \in \mathrm{Sym}_\star^+(8)} \frac{1}{850} \sum_{j=1}^{850} d^2(\Sigma, \Sigma_{ij}),$$

where d is the affine-invariant distance (Moakher (2005), Pennec, Fillard and Ayache (2006)) on $\mathrm{Sym}_\star^+(8)$. The sequence $Y_1, \dots, Y_n$ then constituted the observed time-indexed random objects to be analyzed by the proposed regularized Fréchet regression.

We set $P = 16$ and found that the results were not sensitive to the choice of P within the reasonable range $[12, 20]$. This led to a sequence of $n = 243$ time-indexed 8×8 covariance (symmetric positive definite, SPD) matrices. For better numerical stability, each matrix was scaled by the constant 10^{-3}.

Further discussion of the results. In panel (i) of Figure 3, there are only two jump points left, at time points 42 and 64. This suggests that changes in the fMRI signal caused by early events are more pervasive than those at later events, which is also in line with the fact that the video transition at time point 197 gave rise to two estimated jump points, located slightly before and after. These findings might be due to a stronger brain reaction to the stimulus when the video clip is presented early on in the recording sequence, with subsequent attenuation.

This example demonstrates that changing the penalty can be used as a tool to determine a hierarchy of jump points with the more pronounced jump points persisting even when large penalties are applied. Remarkably, the location of the estimated jump points is hardly affected by the size of the penalty in this example.

APPENDIX D: AUXILIARY RESULTS

PROPOSITION 3. *Let $(\mathcal{X}, d)$ be a metric space that has a finite diameter and satisfies $\sup_{x\in\mathcal{X}} \log N(\epsilon\delta, B_x(\delta), d) \le K\epsilon^{-\alpha}$ for constants $\alpha, K > 0$ and for all $\epsilon, \delta > 0$, where $B_x(\delta)$ denotes the ball in $\mathcal{X}$ centered at x and with radius δ. For a collection $\mathscr{B}(L)$ of Lipschitz continuous $\mathcal{X}$-valued functions defined on $\mathcal{T} = [a, b]$ with a common Lipschitz constant $L < \infty$, it holds that*

$$\log N\big(\delta, \mathscr{B}(L), d_n\big) \le 6^{\alpha} K\big\{2L(b-a)\delta^{-1} + 1\big\} + 4^{\alpha} K R^{\alpha} \delta^{-\alpha},$$

where R denotes the diameter of $\mathcal{X}$.

PROPOSITION 4. *Let $(\mathcal{M}, d)$ be a connected smooth Riemannian manifold, and $\Omega \subset \mathcal{M}$ a closed uniquely geodesic subspace of diameter $R > 0$. Suppose that $\mathscr{B} \equiv \mathscr{B}(p, D_1, D_2)$ is a collection of Ω-valued functions defined on $\mathcal{T}$ such that $\sup_t d(\gamma(t), p) \le D_1$ and $\mathrm{TV}(\gamma) \le D_2$ for some $p \in \Omega$ and all $\gamma \in \mathscr{B}$, where D_1, D_2 are constants. Let $\kappa \ge 0$ be a constant such that the sectional curvature of Ω falls into the interval $[-\kappa, \kappa]$. Then for all $D_1 \in (0, R]$ and $D_2 > 0$,*

$$\log N(\delta, \mathscr{B}, d_n) \le k\big\{c_0 k^{1/2}\big(1 + c_\kappa R^2\big)^2 D_2 \delta^{-1} + \log\big(D_1\big(1 + c_\kappa R^2\big)k^{1/2}\delta^{-1}\big)\big\},$$

where k is the dimension of $\mathcal{M}$, c_0 is an absolute constant and c_κ is a constant depending only on κ.

Acknowledgments. We extend our sincere thanks to the Editor, Associate Editor and several referees for their constructive comments that lead to numerous improvements over a previous version. Data were provided in part by the Human Connectome Project, WU-Minn Consortium (PI: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Funding. The first author was supported by NUS Startup Grant R-155-000-217-133.
The second author was supported by NSF Grants DMS-1712864 and DMS-2014626.

SUPPLEMENTARY MATERIAL

Supplement to "Total variation regularized Fréchet regression for metric-space valued data" (DOI: 10.1214/21-AOS2095SUPP; .pdf). We provide proofs for propositions and theorems in Section 4 and Appendix D.

REFERENCES

AFSARI, B. (2011). Riemannian L^p center of mass: Existence, uniqueness, and convexity. *Proc. Amer. Math. Soc.* **139** 655–673. MR2736346 https://doi.org/10.1090/S0002-9939-2010-10541-5

ALQUIER, P., COTTET, V. and LECUÉ, G. (2019). Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *Ann. Statist.* **47** 2117–2144. MR3953446 https://doi.org/10.1214/18-AOS1742

ARSIGNY, V., FILLARD, P., PENNEC, X. and AYACHE, N. (2006). Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magn. Reson. Med.* **56** 411–421.

ARSIGNY, V., FILLARD, P., PENNEC, X. and AYACHE, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* **29** 328–347. MR2288028 https://doi.org/10.1137/050637996

BAČÁK, M. (2015). Convergence of nonlinear semigroups under nonpositive curvature. *Trans. Amer. Math. Soc.* **367** 3929–3953. MR3324915 https://doi.org/10.1090/S0002-9947-2015-06087-5

BERGMANN, R. and WEINMANN, A. (2016). A second-order TV-type approach for inpainting and denoising higher dimensional combined cyclic and vector space data. *J. Math. Imaging Vision* **55** 401–427. MR3489791 https://doi.org/10.1007/s10851-015-0627-3

BERGMANN, R., LAUS, F., STEIDL, G. and WEINMANN, A. (2014). Second order differences of cyclic data and applications in variational denoising. *SIAM J. Imaging Sci.* **7** 2916–2953. MR3293451 https://doi.org/10.1137/140969993

BHATTACHARYA, R. and PATRANGENARU, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.* **31** 1–29. MR1962498 https://doi.org/10.1214/aos/1046294456

BILLERA, L. J., HOLMES, S. P. and VOGTMANN, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** 733–767. MR1867931 https://doi.org/10.1006/aama.2001.0759

BRIDSON, M. R. (1991). Geodesics and curvature in metric simplicial complexes. Ph.D. thesis, Cornell University. MR2686786

BRIDSON, M. R. and HAEFLIGER, A. (1999). *Metric Spaces of Non-positive Curvature. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **319**. Springer, Berlin. MR1744486 https://doi.org/10.1007/978-3-662-12494-9

BURAGO, D., BURAGO, Y. and IVANOV, S. (2001). *A Course in Metric Geometry. Graduate Studies in Mathematics* **33**. Amer. Math. Soc., Providence, RI. MR1835418 https://doi.org/10.1090/gsm/033

CHAMBOLLE, A., CASELLES, V., CREMERS, D., NOVAGA, M. and POCK, T. (2010). An introduction to total variation for image analysis. In *Theoretical Foundations and Numerical Methods for Sparse Recovery. Radon Ser. Comput. Appl. Math.* **9** 263–340. de Gruyter, Berlin. MR2731599 https://doi.org/10.1515/9783110226157.263

CHANG, T. (1989). Spherical regression with errors in variables. *Ann. Statist.* **17** 293–306. MR0981451 https://doi.org/10.1214/aos/1176347017

CHINOT, G., LECUÉ, G. and LERASLE, M. (2020). Robust statistical learning with Lipschitz and convex loss functions. *Probab. Theory Related Fields* **176** 897–940. MR4087486 https://doi.org/10.1007/s00440-019-00931-3

CORNEA, E., ZHU, H., KIM, P. and IBRAHIM, J. G. (2017). Regression models on Riemannian symmetric spaces. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 463–482. MR3611755 https://doi.org/10.1111/rssb.12169

DAI, X. and MÜLLER, H.-G. (2018). Principal component analysis for functional data on Riemannian manifolds and spheres. *Ann. Statist.* **46** 3334–3361. MR3852654 https://doi.org/10.1214/17-AOS1660

DAVIS, B. C., FLETCHER, P. T., BULLITT, E. and JOSHI, S. (2010). Population shape regression from random design data. *Int. J. Comput. Vis.* **90** 255–266.

DESIKAN, R. S., SÉGONNE, F., FISCHL, B., QUINN, B. T., DICKERSON, B. C., BLACKER, D., BUCKNER, R. L., DALE, A. M., MAGUIRE, R. P. et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31** 968–980.

DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921. MR1635414 https://doi.org/10.1214/aos/1024691081

DRYDEN, I. L., KOLOYDENKO, A. and ZHOU, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann. Appl. Stat.* **3** 1102–1123. MR2750388 https://doi.org/10.1214/09-AOAS249

DUBEY, P. and MÜLLER, H.-G. (2020a). Fréchet change-point detection. *Ann. Statist.* **48** 3312–3335. MR4185810 https://doi.org/10.1214/19-AOS1930

DUBEY, P. and MÜLLER, H.-G. (2020b). Functional models for time-varying random objects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 275–327. MR4084166

ELTZNER, B. and HUCKEMANN, S. F. (2019). A smeary central limit theorem for manifolds with application to high-dimensional spheres. *Ann. Statist.* **47** 3360–3381. MR4025745 https://doi.org/10.1214/18-AOS1781

ESSEN, D. C. V., SMITH, S. M., BARCH, D. M., BEHRENS, T. E. J., YACOUB, E., UGURBIL, K. and WU-MINN HCP CONSORTIUM (2013). The Wu–Minn human connectome project: An overview. *NeuroImage* **80** 62–79.

FANG, B., GUNTUBOYINA, A. and SEN, B. (2021). Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. *Ann. Statist.* **49** 769–792. MR4255107 https://doi.org/10.1214/20-aos1977

FARAWAY, J. J. (2014). Regression for non-Euclidean data using distance matrices. *J. Appl. Stat.* **41** 2342–2357. MR3256391 https://doi.org/10.1080/02664763.2014.909794

FILLARD, P., ARSIGNY, V., AYACHE, N. and PENNEC, X. (2005). A Riemannian framework for the processing of tensor-valued images. In *International Workshop on Deep Structure, Singularities, and Computer Vision* 112–123.

FISHER, N. I. (1995). *Statistical Analysis of Circular Data*. Cambridge Univ. Press, Cambridge. MR1251957 https://doi.org/10.1017/CBO9780511564345

FLETCHER, P. T. (2013). Geodesic regression and the theory of least squares on Riemannian manifolds. *Int. J. Comput. Vis.* **105** 171–185. MR3104017 https://doi.org/10.1007/s11263-012-0591-y

FLETCHER, T. and JOSHI, S. (2007). Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Process.* **87** 250–262.

FRISTON, K. J. (2011). Functional and effective connectivity: A review. *Brain Connect.* **1** 13–36.

GREEN, M. F., HORAN, W. P. and LEE, J. (2015). Social cognition in schizophrenia. *Nat. Rev. Neurosci.* **16** 620–631.

HEIN, M. (2009). Robust nonparametric regression with metric-space valued output. In *Advances in Neural Information Processing Systems* 718–726.

HINKLE, J., FLETCHER, P. T. and JOSHI, S. (2014). Intrinsic polynomials for regression on Riemannian manifolds. *J. Math. Imaging Vision* **50** 32–52. MR3233133 https://doi.org/10.1007/s10851-013-0489-5

HOTZ, T., HUCKEMANN, S., LE, H., MARRON, J. S., MATTINGLY, J. C., MILLER, E., NOLEN, J., OWEN, M., PATRANGENARU, V. et al. (2013). Sticky central limit theorems on open books. *Ann. Appl. Probab.* **23** 2238–2258. MR3127934 https://doi.org/10.1214/12-AAP899

HÜTTER, J.-C. and RIGOLLET, P. (2016). Optimal rates for total variation denoising. In 29*th Annual Conference on Learning Theory* (V. Feldman, A. Rakhlin and O. Shamir, eds.). *Proceedings of Machine Learning Research* **49** 1115–1146. PMLR, Columbia University, New York, New York, USA.

KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. MR2505584 https://doi.org/10.1137/070690274

KLOECKNER, B. (2010). A geometric study of Wasserstein spaces: Euclidean spaces. *Ann. Sc. Norm. Super. Pisa Cl. Sci.* (5) **9** 297–323. MR2731158

KOLAR, M. and XING, E. P. (2012). Estimating networks with jumps. *Electron. J. Stat.* **6** 2069–2106. MR3020257 https://doi.org/10.1214/12-EJS739

LEDOUX, M. and TALAGRAND, M. (2011). *Probability in Banach Spaces: Isoperimetry and Processes*. *Classics in Mathematics*. Springer, Berlin. MR2814399

LANG, S. (1995). *Differential and Riemannian Manifolds*. Springer, New York.

LELLMANN, J., STREKALOVSKIY, E., KOETTER, S. and CREMERS, D. (2013). Total variation regularization for functions with values in a manifold. In 2013 *IEEE International Conference on Computer Vision* 2944–2951. IEEE, New York.

LIN, Z. (2019). Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition. *SIAM J. Matrix Anal. Appl.* **40** 1353–1370. MR4032859 https://doi.org/10.1137/18M1221084

LIN, Z. and MÜLLER, H.-G. (2021). Supplement to "Total variation regularized Fréchet regression for metric-space valued data." https://doi.org/10.1214/21-AOS2095SUPP

MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. MR1429931 https://doi.org/10.1214/aos/1034276635

MOAKHER, M. (2005). A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* **26** 735–747. MR2137480 https://doi.org/10.1137/S0895479803436937

ORTELLI, F. and VAN DE GEER, S. (2018). On the total variation regularized estimator over a class of tree graphs. *Electron. J. Stat.* **12** 4517–4570. MR3892703 https://doi.org/10.1214/18-ejs1519

OWEN, M. and PROVAN, J. S. (2011). A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8** 2–13.

PATRANGENARU, V. and ELLINGSON, L. (2015). *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*. CRC Press, Boca Raton, FL. MR3444169

PELLETIER, B. (2006). Non-parametric regression estimation on closed Riemannian manifolds. *J. Nonparametr. Stat.* **18** 57–67. MR2214065 https://doi.org/10.1080/10485250500504828

PENNEC, X. (2018). Barycentric subspace analysis on manifolds. *Ann. Statist.* **46** 2711–2746. MR3851753 https://doi.org/10.1214/17-AOS1636

PENNEC, X., FILLARD, P. and AYACHE, N. (2006). A Riemannian framework for tensor computing. *Int. J. Comput. Vis.* **66** 41–66.

PETERSEN, A. and MÜLLER, H.-G. (2016). Fréchet integration and adaptive metric selection for interpretable covariances of multivariate functional data. *Biometrika* **103** 103–120. MR3465824 https://doi.org/10.1093/biomet/asv054

PETERSEN, A. and MÜLLER, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *Ann. Statist.* **47** 691–719. MR3909947 https://doi.org/10.1214/17-AOS1624

PETRUNIN, A. and TUSCHMANN, W. (1999). Diffeomorphism finiteness, positive pinching, and second homotopy. *Geom. Funct. Anal.* **9** 736–774. MR1719602 https://doi.org/10.1007/s000390050101

RATHI, Y., TANNENBAUM, A. and MICHAILOVICH, O. (2007). Segmenting images on the tensor manifold. In 2007 *IEEE Conference on Computer Vision and Pattern Recognition* 1–8. IEEE, New York.

RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D, Nonlinear Phenom.* **60** 259–268. MR3363401 https://doi.org/10.1016/0167-2789(92)90242-F

SACKS, J. and YLVISAKER, D. (1970). Designs for regression problems with correlated errors. III. *Ann. Math. Stat.* **41** 2057–2074. MR0270530 https://doi.org/10.1214/aoms/1177696705

SADHANALA, V., WANG, Y.-X. and TIBSHIRANI, R. J. (2016). Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Neural Information Processing Systems* 3521–3529.

SHI, X., STYNER, M., LIEBERMAN, J., IBRAHIM, J. G., LIN, W. and ZHU, H. (2009). Intrinsic regression models for manifold-valued data. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI* **12** 192–199.

STEINKE, F., HEIN, M. and SCHÖLKOPF, B. (2010). Nonparametric regression between general Riemannian manifolds. *SIAM J. Imaging Sci.* **3** 527–563. MR2736019 https://doi.org/10.1137/080744189

STRONG, D. and CHAN, T. (2003). Edge-preserving and scale-dependent properties of total variation regularization. *Inverse Probl.* **19** S165–S187. MR2036526 https://doi.org/10.1088/0266-5611/19/6/059

STURM, K.-T. (2003). Probability measures on metric spaces of nonpositive curvature. In *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces* (*Paris*, 2002). *Contemp. Math.* **338** 357–390. Amer. Math. Soc., Providence, RI. MR2039961 https://doi.org/10.1090/conm/338/06080

TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323. MR3189487 https://doi.org/10.1214/13-AOS1189

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641 https://doi.org/10.1111/j.1467-9868.2005.00490.x

VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. MR1056343 https://doi.org/10.1214/aos/1176347632

VAN DE GEER, S. (2001). Least squares estimation with complexity penalties. *Math. Methods Statist.* **10** 355–374. MR1867165

WANG, X., ZHU, H. and ADNI (2017). Generalized scalar-on-image regression models via total variation. *J. Amer. Statist. Assoc.* **112** 1156–1168. MR3735367 https://doi.org/10.1080/01621459.2016.1194846

WANG, Y.-X., SHARPNACK, J., SMOLA, A. J. and TIBSHIRANI, R. J. (2016). Trend filtering on graphs. *J. Mach. Learn. Res.* **17** Paper No. 105, 41 pp. MR3543511

WEINMANN, A., DEMARET, L. and STORATH, M. (2014). Total variation regularization for manifold-valued data. *SIAM J. Imaging Sci.* **7** 2226–2257. MR3278838 https://doi.org/10.1137/130951075

YUAN, Y., ZHU, H., LIN, W. and MARRON, J. S. (2012). Local polynomial regression for symmetric positive definite matrices. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 697–719. MR2965956 https://doi.org/10.1111/j.1467-9868.2011.01022.x

ZHOU, D., DRYDEN, I. L., KOLOYDENKO, A. A., AUDENAERT, K. M. R. and BAI, L. (2016). Regularisation, interpolation and visualisation of diffusion tensor images using non-Euclidean statistics. *J. Appl. Stat.* **43** 943–978. MR3457097 https://doi.org/10.1080/02664763.2015.1080671

ZHU, H., FAN, J. and KONG, L. (2014). Spatially varying coefficient model for neuroimaging data with jump discontinuities. *J. Amer. Statist. Assoc.* **109** 1084–1098. MR3265682 https://doi.org/10.1080/01621459.2014.881742