The Annals of Statistics
2022, Vol. 50, No. 6, 3647–3676
https://doi.org/10.1214/22-AOS2245

THE STEIN EFFECT FOR FRÉCHET MEANS

BY ANDREW MCCORMACK[a] AND PETER HOFF[b]

Department of Statistical Science, Duke University, [a]*andrew.mccormack@duke.edu,* [b]*peter.hoff@duke.edu*

The Fréchet mean is a useful description of location for a probability distribution on a metric space that is not necessarily a vector space. This article considers simultaneous estimation of multiple Fréchet means from a decision-theoretic perspective, and in particular, the extent to which the unbiased estimator of a Fréchet mean can be dominated by a generalization of the James–Stein shrinkage estimator. It is shown that if the metric space satisfies a nonpositive curvature condition, then this generalized James–Stein estimator asymptotically dominates the unbiased estimator as the dimension of the space grows. These results hold for a large class of distributions on a variety of spaces, including Hilbert spaces and, therefore, partially extend known results on the applicability of the James–Stein estimator to nonnormal distributions on Euclidean spaces. Simulation studies on phylogenetic trees and symmetric positive definite matrices are presented, numerically demonstrating the efficacy of this generalized James–Stein estimator.

1. Introduction. In his seminal 1948 article, Fréchet generalized the notion of the mean of a real-valued random variable to a metric space-valued random object [30]. Like the usual mean, the Fréchet mean provides a summary of the location of a distribution, from which a notion of Fréchet variance may also be defined. Fréchet means and variances have been used for statistical analysis of data from nonstandard sample spaces, such as spaces of phylogenetic trees, symmetric positive definite matrices in diffusion tensor imaging and functional data analysis on Wasserstein spaces, to name a few [11, 55, 57, 58]. In terms of methodological development, [21, 59] use Fréchet means to develop extensions of linear regression and ANOVA that are applicable for metric space-valued data. Additionally, substantial effort has gone into studying the convergence properties of sample Fréchet means and variances [10, 32, 75].

This article primarily considers the simultaneous estimation of multiple Fréchet means, and conditions under which a generalized James–Stein shrinkage estimator dominates the natural estimator, the unbiased estimator of the Frechét mean. As shown in [39, 64], if $X \sim N_n(\theta, \sigma^2 I)$ with σ^2 known and $n \geq 3$, X is dominated by the James–Stein shrinkage estimator $\delta_{JS}(X)$, given by

$$\delta_{JS}(X) = \left(\frac{\sigma^2(n-2)}{\|X-\psi\|^2}\right)\psi + \left(1 - \frac{\sigma^2(n-2)}{\|X-\psi\|^2}\right)X, \tag{1}$$

where ψ is a known shrinkage point. Intuitively, this estimator is obtained by starting from X and "shrinking" toward ψ by an amount that is adaptively estimated from the data X. Under the average squared-error loss function $L(\theta, \delta) = \frac{1}{n}\|\theta - \delta\|^2$, the risk of the James–Stein estimator is

$$R(\theta, \delta_{JS}) = \sigma^2 - \sigma^4\left(1 - \frac{2}{n}\right)E\left(\frac{n-2}{\|X-\psi\|^2}\right), \tag{2}$$

Received July 2021; revised October 2022.
MSC2020 subject classifications. Primary 62R20; secondary 62C15.
Key words and phrases. Admissibility, empirical Bayes, Hadamard space, nonparametric, shrinkage.

while the risk of the unbiased estimator X is σ^2 [29]. For large n, the relative improvement of the Stein estimator over X approximately depends on the ratio $\sigma^2/(\sigma^2 + \|\theta - \psi\|^2)$. If the shrinkage point is aptly chosen so that $\|\theta - \psi\|^2$ is small relative to variance of the components of X, the James–Stein estimator will significantly outperform X. The fact that δ_{JS} dominates X is often interpreted as an indication of how sharing information across seemingly unrelated populations can lead to an improved estimator of $\theta_1, \ldots, \theta_n$ with respect to squared error loss summed across all populations. Indeed, the James–Stein estimator may be derived as an empirical Bayes estimator in which $\|X - \psi\|^2$ provides information about the likely magnitude of $\|\theta - \psi\|^2$ [25]. Multivariate generalizations of δ_{JS} have been developed in the setting where X_i and θ_i are vectors with $X_i \sim N_p(\theta_i, \Sigma)$ [24, 48, 65, 71]. When $\Sigma = \sigma^2 I$ these multivariate generalizations can improve over (1) if it is assumed that the θ_i have some shared structure, such as $\theta_i \sim N_p(0, A)$.

In this article, we study a generalization of the Stein estimation problem where we are interested in estimating the Fréchet means $\theta_1, \ldots, \theta_n$ of n different populations given a single metric space valued observation X_i from each population. The estimator (1) can be extended to sample spaces that are uniquely geodesic metric spaces, which are metric spaces where there is a unique path of minimum length, or geodesic, between any two points. The estimator of $\theta_1, \ldots, \theta_n$ that we consider is obtained by traveling from $X = (X_1, \ldots, X_n)$ to a shrinkage point ψ along a geodesic by an amount that is adaptively estimated from X. If the geodesics in the metric space have tractable, known forms, then this estimator is simple to compute in practice.

We develop theoretical results in two different settings that demonstrate that under some mild conditions, the proposed geodesic James–Stein estimator dominates the unbiased estimator asymptotically as the number of populations n increases. The first setting in Section 4 corresponds to the classical James–Stein estimation problem, where the Fréchet means $\theta_1, \ldots, \theta_n$ are fixed. The second setting in Section 5 assumes that the θ_i's are i.i.d. and evaluates the marginal (Bayes) risk of the geodesic James–Stein estimator. In the latter setting, it is reasonable to shrink the observations toward their sample Fréchet mean. Of note is that the domination results obtained are nonparametric; only moment bounds are placed on the family of distributions under consideration. As a consequence, the geodesic James–Stein estimator is robust, having reasonable performance across a wide range of distributions.

As explained by Beran [7] and detailed in Stein's groundbreaking paper [64], the construction of Stein's estimator was motivated in part by the observation that for a large n, by the law of large number the triangle with vertices $n^{-1/2}\psi$, $n^{-1/2}X$ and $n^{-1/2}\theta$ is approximately a right triangle with hypotenuse given by the edge between $n^{-1/2}X$ and $n^{-1/2}\psi$. After rescaling this triangle, Stein's estimator seeks to estimate the point found by projecting $n^{-1/2}\theta$ onto the line spanned by the hypotenuse. By using the knowledge that the given triangle is approximately a right triangle, it is possible to find this projected point since the side lengths $\|n^{-1/2}X - n^{-1/2}\psi\|^2$ and $\|n^{-1/2}X - n^{-1/2}\theta\|^2 \approx \sigma^2$ are known. These asymptotic considerations, which do not rely on the assumption of normality, motivate many of the results to follow.

In particular, the possibility of a Stein effect in a metric space, that is, domination of X by a shrinkage estimator, will partly depend on the geometry of metric space generalizations of triangles, or equivalently the curvature of the metric space. Roughly, the Stein effect is absent in spaces with positive curvature, and generally present in flat spaces or spaces with negative curvature. These latter two spaces are known as Hadamard spaces [67], and encompass a wide variety of metric spaces such as the aforementioned spaces of trees, symmetric positive-definite matrices and Wasserstein space on $\mathbb{R}$. We emphasize here that since any Hilbert space is a Hadamard space, all of the results we develop apply in the setting where the sample space is $\mathbb{R}^n$. Moreover, any closed, convex subset of a Hilbert space, for which the Wasserstein

space on $\mathbb{R}$ is an example, is also a Hadamard space. These results are of interest in this setting as they hold under mild, nonparametric assumptions.

Recent related work [73, 74] examines shrinkage estimators for generalizations of the Gaussian distribution on Lie groups and the manifold of symmetric positive definite matrices. Previous work generalizing the Stein estimator in Euclidean space has involved extending domination results to nonnormal distributions or to larger classes of loss functions [15, 42]. Typically, such distributions are assumed to have some sort of spherical symmetry or exponential family structure [13, 37]. A related focus of research on Stein estimators has been finding estimators that dominate the positive part James–Stein estimator [6], which is known to be inadmissible [16, 63]. For a comprehensive account of shrinkage estimation, see [29].

An outline of the remainder of this article is as follows: In Section 2, the concepts of Fréchet means, variances and Hadamard spaces are reviewed. Section 3 applies these concepts to the problem of estimating a Fréchet mean, and considers conditional Fréchet means and randomized and unbiased estimators. Section 4 provides the core theoretical results of the article, where the geodesic James–Stein estimator is introduced and its risk function for the multigroup estimation problem is investigated. A natural extension of this problem is to place a prior distribution on the Fréchet means of each group. This is done in Section 5 where we introduce the possibility of adaptively estimating a shrinkage point. Asymptotic optimality properties of the geodesic James–Stein estimator and the relationship to empirical Bayes estimators are also discussed in this section. In Section 6, motivated by the problem of estimating gene trees in phylogenetics, we demonstrate numerically how the geodesic James–Stein estimator exhibits favorable performance relative to X in a simulation study on the space of trees. The geodesic James–Stein estimator is then applied in the space of symmetric positive definite matrices as a method of smoothing diffusion tensor MRI data. Proofs of all the results in this article are provided in the Supplementary Material [52].

2. Preliminaries.

2.1. *Metric space valued random objects.* Let $(\mathcal{X}, d)$ be a metric measure space equipped with the Borel σ-algebra $\mathcal{B}$, induced from the metric topology on $\mathcal{X}$. A metric space valued random object X is a $\mathcal{B}$-measurable function from a probability space $(\mathcal{Y}, \mathcal{G}, Q)$ into $\mathcal{X}$. The probability distribution P of X on $(\mathcal{X}, \mathcal{B})$ is defined as the standard pushforward measure, $P(A) := Q(X \in B) = Q(X^{-1}(B))$, $\forall B \in \mathcal{B}$. The existence of the underlying probability space $(\mathcal{Y}, \mathcal{G}, Q)$ is implicitly assumed throughout this article.

Statistical inference for a distribution P is often focused on the estimation of a location of the distribution, and measures of variability about this location. In Euclidean space, the mean of a random variable provides one of the most basic notions of average location or central tendency. In $\mathbb{R}^n$, the integral $\int X\,dP$ that defines the mean of X depends heavily on the vector space structure of $\mathbb{R}^n$. Fréchet [30] proposed a generalization of the Euclidean mean, referred to as the Fréchet mean or the alternatively the barycenter, that applies to arbitrary metric spaces. The idea is that a mean of a random object X should be the collection of points in $\mathcal{X}$ that are on average the closest to X.

DEFINITION 2.1. The Fréchet mean of X, denoted by E_2X, is the solution set to the following variational problem:

$$E_2X := \operatorname*{argmin}_{x \in \mathcal{X}} E\big(d(x, X)^2\big). \tag{3}$$

When $\mathcal{X} = \mathbb{R}^n$ with the Euclidean metric, E_2X coincides with the usual Euclidean mean. The above definition can be further generalized by changing the exponent of the distance

function in (3). In $\mathbb{R}^n$, if $d(x, X)$ is instead raised to the first power in (3), the resulting generalized Fréchet mean is the set of componentwise L^1 medians of X. Unlike the Euclidean case, the existence and uniqueness of the solutions to (3) is not guaranteed, so that E_2X is set-valued in general and can even be the empty set. A simple example of the nonexistence of a Fréchet mean is when $X \sim N(0, 1)$ on the space $\mathbb{R} - \{0\}$.

If E_2X is to be meaningful, we require that $E(d(x, X)^2) < \infty$ for at least one $x \in \mathcal{X}$. By the triangle inequality, $d(x, X) \leq d(x, x_0) + d(x_0, X)$, which implies that $E(d(x, X)^2) < \infty$ for all $x \in \mathcal{X}$. We say that $X \in \mathcal{L}^2(\mathcal{X})$ if $E(d(x, X)^2) < \infty$ for all $x \in \mathcal{X}$. It should be remarked that this is slightly different than the situation in Euclidean space since a Euclidean mean $E(X)$ exists and is finite as long as $E(|X|) < \infty$ or equivalently $E(|X - x|) < \infty$, $\forall x \in \mathbb{R}^n$. There is a more general definition of a Fréchet mean that accounts for this minor discrepancy, although we do not have any need for this extra generality [67].

Having defined a mean, it is useful to have a measure describing the spread of X about this mean. The Fréchet variance captures the average squared distance of X from its corresponding Fréchet mean.

DEFINITION 2.2. The Fréchet variance of $X \in \mathcal{L}^2(\mathcal{X})$, denoted by V_2X, is the number

$$V_2X := \inf_{x \in \mathcal{X}} E\big(d(x, X)^2\big). \tag{4}$$

If $X \in \mathbb{R}^n$ with covariance matrix Σ, then the 2-Fréchet variance of X is the total variance $\operatorname{tr}(\Sigma)$, which is the sum of the variances of each component of X. As seen from this example, Fréchet variances do not capture any information about how the spread of X varies in different "directions" in the metric space. Fréchet variances only summarize the average squared distance of a random object from its Fréchet mean set.

Throughout the remainder of this article if X is distributed according to P, then the notation $E_2P := E_2X$ and $V_2P := V_2X$ will be used.

2.2. *Hadamard spaces.* A geodesic curve in a metric space $(\mathcal{X}, d)$ is a generalization of a straight line segment in $\mathbb{R}^n$.

DEFINITION 2.3. The curve $\gamma : [a, b] \to \mathcal{X}$, where $-\infty < a < b < \infty$, is a speed v geodesic if $d(\gamma(t_1), \gamma(t_0)) = v|t_1 - t_0|$ for all $a \leq t_1, t_0 \leq b$.

This definition requires that the points on the curve γ look exactly the same as the points on a corresponding interval in $\mathbb{R}$ with respect to the metric. Thus, the map $f : [va, vb] \to \gamma([a, b])$ defined by $f(s) = \gamma(s/v)$ is an isometry. The length of a curve $\sigma : [a, b] \to \mathcal{X}$ is defined by $\ell(\sigma) = \sup_{a=x_0 \leq \cdots \leq x_k = b} \sum_{i=1}^k d(\sigma(x_i), \sigma(x_{i-1}))$ where the supremum is over any finite partition $(x_0, \ldots, x_k)$ of the interval $[a, b]$. A geodesic connecting any two points is a minimizer of the length functional out of all curves between these points.

A metric space $(\mathcal{X}, d)$ is defined to be a geodesic space if for all $x_1, x_0 \in \mathcal{X}$ there exists a geodesic $\gamma : [a, b] \to \mathcal{X}$ with endpoints, $\gamma(a) = x_0$, $\gamma(b) = x_1$ [3, 17]. The metric space $\mathcal{X}$ is uniquely geodesic if it is geodesic and any two geodesics $\gamma, \sigma : [a, b] \to \mathcal{X}$, with $\gamma(a) = \sigma(a)$, $\gamma(b) = \sigma(b)$ are equal [14]. In a uniquely geodesic space where $\gamma : [0, 1] \to \mathcal{X}$ is a geodesic with $\gamma(0) = x$ and $\gamma(1) = y$, the notation $[x, y]_t$ for $t \in [0, 1]$ will be used to represent the point $\gamma(t)$. The interpretation of $[x, y]_t$ is that this is the point obtained by travelling along the geodesic that connects x to y, whose distance from x is a proportion t of the total length of the geodesic. Similarly, the expression $[x, y]$ represents the image in $\mathcal{X}$ of the geodesic between x and y.

In a normed vector space $(V, \|\cdot\|)$, line segments are geodesic in the sense defined above. To see this, if $\gamma : [a, b] \to V$ is the line segment $\gamma(t) = v_1t + v_0$, then $\|\gamma(t_1) - \gamma(t_0)\| =$

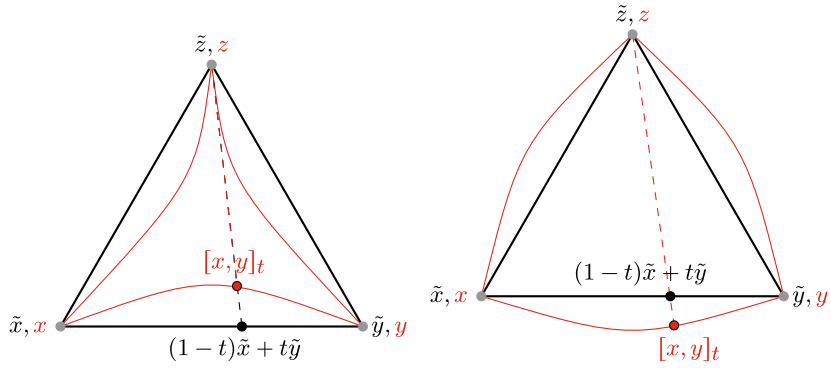

FIG. 1. *Metric space comparison triangles, Δxyz and $\Delta\tilde{x}\tilde{y}\tilde{z}$ in a space with negative Alexandrov curvature (left) and positive Alexandrov curvature (right).*

$\|v_1\||t_1 - t_0|$, implying that γ is a speed $\|v_1\|$ geodesic. Any normed vector space is thus geodesic but may not be uniquely geodesic. In the case where V is an inner product space, V is uniquely geodesic. On a sphere, geodesics are the minor arcs of great circles, which are the shortest paths that connect points on a sphere. The sphere is geodesic but not uniquely geodesic because any two antipodal points can be joined by infinitely many geodesics. It is worth noting that in a Riemannian manifold geodesics are more commonly defined as curves that are local minimizers of the Riemannian length functional. The definition of a geodesic presented here requires that a geodesic be a global minimizer of the length functional and so it is more restrictive than the usual definition if $\mathcal{X}$ is a Riemannian manifold.

The curvature of a uniquely geodesic metric space is primarily described in terms of the geometric properties of generalized triangles in the space. Given three points $x, y, z \in \mathcal{X}$ the triangle $\Delta xyz \subset \mathcal{X}$ is defined as the set of points $[x, y] \cup [y, z] \cup [z, x]$. Due to the triangle inequality, given the numbers $d(x, y)$, $d(y, z)$, $d(z, x)$, there exist points $\tilde{x}$, $\tilde{y}$, $\tilde{z}$ in $\mathbb{R}^2$ such that the triangle $\Delta\tilde{x}\tilde{y}\tilde{z} \subset \mathbb{R}^2$ has side lengths $d(x, y)$, $d(y, z)$ and $d(z, x)$. The Alexandrov curvature [2] of a metric space compares how the distance from $[x, y]_t$ to z in $\mathcal{X}$ differs from the distance from $(1-t)\tilde{x} + t\tilde{y}$ to $\tilde{z}$ in $\mathbb{R}^2$ for $t \in [0, 1]$. A metric space has negative Alexandrov curvature if $d([x, y]_t, z)$ is no greater than $d([\tilde{x}, \tilde{y}]_t, \tilde{z})$ for all $x, y, z \in \mathcal{X}$ while being less than $d([\tilde{x}, \tilde{y}]_t, \tilde{z})$ for at least some triplet of points $x, y, z \in \mathcal{X}$ [14]. Positive Alexandrov curvature is defined similarly, while a space with zero Alexandrov curvature has $d([x, y]_t, z) = d([\tilde{x}, \tilde{y}]_t, \tilde{z})$ for all $x, y, z \in \mathcal{X}$. These requirements can be visualized as positively curved spaces having triangles with edges that bend outwards and negatively curved spaces having triangles with edges that bend inwards, relative to triangles in $\mathbb{R}^2$. See Figure 1 for typical examples of generalized triangles in positively and negatively curved spaces. The generalized triangles in Figure 1 are isometrically embedded in $\mathbb{R}^2$ so that all distances between points are given by Euclidean distance.

For all $x, y, z \in \mathcal{X}$ a metric space with nonpositive curvature satisfies the CAT(0) curvature bound $d([x, y]_t, z) \leq d([\tilde{x}, \tilde{y}]_t, \tilde{z})$. After expanding $d([\tilde{x}, \tilde{y}]_t, \tilde{z})$ in terms of the side lengths of the triangle $\Delta\tilde{x}\tilde{y}\tilde{z}$, the CAT(0) bound is equivalent to

$$d\big([x, y]_t, z\big)^2 \leq (1-t)d(x, z)^2 + td(y, z)^2 - t(1-t)d(x, y)^2 \tag{5}$$

for all $x, y, z \in \mathcal{X}$ and $t \in [0, 1]$.

DEFINITION 2.4. A Hadamard space is a complete, uniquely geodesic metric space that satisfies the CAT(0) curvature bound in (5).

The subset of Hadamard spaces that have zero Alexandrov curvature so that (5) holds with equality are geometrically similar to $\mathbb{R}^n$. In this case, the triangle Δxyz is indistinguishable

from its comparison triangle $\Delta\tilde{x}\tilde{y}\tilde{z}$, and thus Euclidean trigonometry will apply to Δxyz. For example, a version of the Euclidean law of sines or cosines will hold in such a space, and suitably defined interior angles of Δxyz will also sum to π. Any Hilbert space or closed, convex subset thereof is a Hadamard space with vanishing Alexandrov curvature. Consequently, any results that hold for Hadamard spaces will also hold for Hilbert spaces, which is the setting of much of classical statistics.

The definition of Alexandrov curvature is motivated in part as a generalization of the sectional curvature of a Riemannian manifold. As such, any complete Riemannian manifold with nonpositive sectional curvature is a Hadamard space. For example, the saddle surface in $\mathbb{R}^3$ has negative sectional curvature and is a Hadamard space with negative Alexandrov curvature. If one draws a triangle of shortest paths on such a surface, it will look like the comparison triangle in Figure 1. Another easily visualized example of a Hadamard space with nonzero curvature is a metric tree. Metric trees are weighted graphs that are trees endowed with the shortest path metric. Additional details on metric trees can be found in the Supplementary Material [53].

In a Hilbert space, any closed and convex set $\mathcal{C}$ has the property that there exists a unique projection of any point x onto $\mathcal{C}$ that minimizes the squared distance of x from $\mathcal{C}$. If $\mathcal{C}$ is a closed linear subspace, this follows from the Pythagorean theorem. This result can be generalized to Hadamard spaces as follows: A set $\mathcal{C}$ in a geodesic space is said to be convex if for all $x, y \in \mathcal{C}$ we have that $[x, y] \subset \mathcal{C}$. The Hadamard space projection theorem of [5] says that for any point $x \in \mathcal{X}$ and closed, convex subset $\mathcal{C}$ of a Hadamard space there exists a unique point $\Pi(x) \in \mathcal{C}$ that satisfies $d(x, \Pi(x))^2 = \inf_{y \in \mathcal{C}} d(x, y)^2$. In addition, $\Pi(x)$ satisfies the inequality

$$d(x, z)^2 \geq d\big(z, \Pi(x)\big)^2 + d\big(\Pi(x), x\big)^2 \quad \forall z \in \mathcal{C}. \tag{6}$$

The inequality in (6) provides a bound on how close $\Pi(x)$ is to x relative to any other point $z \in \mathcal{C}$. When $\mathcal{C}$ is a closed vector subspace of a Hilbert space, (6) holds with equality and is the Pythagorean theorem.

Analogous to the construction of $L^2(\mathbb{R}^n)$, the set $L^2(\mathcal{X})$ of almost everywhere equal random objects on $\mathcal{X}$ is a Hadamard space under the metric

$$\rho(X, Y) := E\big(d(X, Y)^2\big)^{1/2}.$$

Geodesics are given pointwise by $[X, Y]_t(\omega) = [X(\omega), Y(\omega)]_t$. The CAT(0) bound follows by the linearity of expectations while completeness follows in the same way that completeness of $L^2(\mathbb{R}^n)$ follows from the completeness of $\mathbb{R}^n$ [5].

3. Estimation of Fréchet means in Hadamard spaces. A general Hadamard space point estimation problem can be formulated as follows: Let $\mathcal{P} \subset L^2(\mathcal{X})$ be a family of distributions on the Hadamard space $(\mathcal{X}, d)$ and $g : \mathcal{P} \to \mathcal{X}$ be a functional defined on $\mathcal{P}$ that is an estimand of interest. For example, g could be the Fréchet mean functional $g(P) = E_2 P$. Given a single observation of $X \sim P \in \mathcal{P}$, we seek to estimate $g(P)$ under the squared distance loss

$$L\big(P, \delta(X)\big) := d\big(g(P), \delta(X)\big)^2,$$

where the corresponding risk function is $R(P, \delta) := E(L(P, \delta))$ and $\delta : \mathcal{X} \to \mathcal{X}$ is a generic estimator.

When $(\mathcal{X}, d)$ is a Hadamard space, it is possible to obtain a bias-variance type of inequality for this risk function by using (6). First, note that the collection of constant almost everywhere random objects $\mathcal{C} := \{\theta \in \mathcal{X}\} \subset L^2(\mathcal{X})$ is a closed and convex set in the Hadamard space

$L^2(\mathcal{X})$. By definition, $E_2(\delta(X)) = \operatorname{argmin}_{\theta\in\mathcal{C}} \rho(\delta,\theta)^2$ and the projection theorem implies that the Fréchet mean $E_2(\delta(X)) = \Pi(\delta)$ exists and is unique [67]. Therefore, the nonuniqueness of Fréchet means is not a concern when working in Hadamard spaces. The inequality in (6) becomes

$$E\big(d\big(g(P),\delta\big)^2\big) \geq d\big(g(P),E_2\delta\big)^2 + E\big(d(E_2\delta,\delta)^2\big) \quad \forall g(P)\in\mathcal{X}, \tag{7}$$

which can be viewed as a bias-variance inequality. If $\delta(X)$ is used as an estimator for $g(P)$ under the loss function $L(P,\cdot) = d(g(P),\cdot)^2$ then the term $E(d(E_2\delta,\delta)^2)$ is exactly the Fréchet variance $V_2\delta$, while $d(g(P),E_2\delta)^2$ can be regarded as the squared bias of δ. As $E(d(\delta,E_2\delta)^2) \leq E(d(\delta,\theta)^2)$ for all $\theta\in\mathcal{X}$, δ is risk unbiased for $E_2\delta$ under the squared distance loss [46].

Conditional expectations of random objects in a Hadamard space can be defined in a similar manner to Fréchet means. Recall that for a σ-algebra $\mathcal{G}\subset\mathcal{B}$ the conditional expectation of $X\in L^2(\mathbb{R}^n)$ is the projection of X onto the closed vector subspace of $\mathcal{G}$-measurable random variables in $L^2(\mathbb{R}^n)$. Likewise, taking $\mathcal{C} := \{Y\in L^2(\mathcal{X}) : \sigma(Y)\subset\mathcal{G}\}$ to be the $\mathcal{G}$-measurable random objects in $L^2(\mathcal{X})$, the conditional expectation $E_2(X|\mathcal{G})$, as defined in [5], is given by

$$E_2(X|\mathcal{G}) := \operatorname*{argmin}_{Y\in\mathcal{C}} E\big(d(X,Y)^2\big). \tag{8}$$

As the set $\mathcal{C}$ is closed and convex, $E_2(X|\mathcal{G})$ exists, is unique, and satisfies a version of (6). However, the lack of a vector space structure in $\mathcal{C}$ implies that not all of the familiar properties of Euclidean conditional expectations carry over to Hadamard spaces.

The CAT(0) inequality (5) can be interpreted as a statement about the convexity of the loss function $d(g(P),\cdot)^2$. A function $f:\mathcal{X}\to\mathbb{R}$ on a uniquely geodesic space is said to be metrically convex if $f([x,y]_t)$ is convex as a function of $t\in[0,1]$ for any choice of $x,y\in\mathcal{X}$ [5]. By (5), the function $f_z(x) = d(z,x)^2$ is metrically convex for all $z\in\mathcal{X}$. This convexity yields behavior similar to that of convex functions defined on $\mathbb{R}^n$. For instance, a Fréchet mean version of Jensen's inequality, $E(d(x,\delta)^2) \geq d(x,E_2\delta)^2$ for all $x\in\mathcal{X}$, is an immediate consequence of (7). The metric convexity of the squared distance function is also the key property that allows for the favorable use of the shrinkage estimators considered in the next section.

In $\mathbb{R}^n$ the collection of nonrandomized estimators forms an essentially complete class for decision problems with convex loss functions. The analogous result in Hadamard spaces is the following.

LEMMA 3.1. *Let $\delta(X,U)\in\mathcal{X}$ be a randomized estimator where $U\sim\text{Unif}[0,1]$ is independent of $X\sim P$ and $R(P,\delta)<\infty$. If $L(P,\cdot):\mathcal{X}\to[0,\infty)$ is a metrically convex and lower semicontinuous function for each fixed $P\in\mathcal{P}$ then there exists a nonrandomized estimator $\tilde{\delta}(X)$ with $R(P,\tilde{\delta})\leq R(P,\delta)$ for all $P\in\mathcal{P}$.*

PROOF. Fix P and take $\tilde{\delta}(X) = E_2(\delta(X,U)|\sigma(X))$ as defined by (8). By Jensen's inequality for conditional expectations [5], $L(P,\tilde{\delta}) \leq E(L(P,\delta)|\sigma(X))$ almost everywhere P, from which $R(P,\tilde{\delta})\leq R(P,\delta)$ follows. □

A version of the Rao–Blackwell theorem can be extended to this setting by similar reasoning. Suppose that a σ-algebra $\mathcal{G}$ has the property that $E_2(\delta(X)|\mathcal{G}) = E_2(\delta(Y)|\mathcal{G})$ when $X\sim P$ and $Y\sim Q$ for all $P,Q\in\mathcal{P}$, so that the random object $\tilde{\delta} = E_2(\delta(X)|\mathcal{G})$ does not depend on $P\in\mathcal{P}$. Further suppose that a version of $E_2(\delta|\mathcal{G})(\omega)$ can be realized as a function of $X(\omega)$, so that $\tilde{\delta}(X(\omega)) := E_2(\delta(X)|\mathcal{G})(\omega)$ is an estimator. This second assumption holds in the typical scenario where $\mathcal{X}$ is separable and $\mathcal{G} = \sigma(f(X))$ for some measurable function

f of X [23]. Under a convex and lower semicontinuous loss, the risk of $E(\delta|\mathcal{G})$ is less than or equal to the risk of δ for every $P \in \mathcal{P}$. It should be noted that the standard definition of sufficiency of $\mathcal{G}$, requiring that $P(A|\mathcal{G}) = Q(A|\mathcal{G})$ for all $P, Q \in \mathcal{P}$, $A \in \mathcal{B}$, does not immediately imply that $\tilde{\delta} = E_2(\delta|\mathcal{G})$ is independent of the choice of P. The reason is that in the case of a Euclidean valued $\delta(X)$, conditional expectations can be approximated by conditional probabilities using the dominated convergence theorem for conditional expectations. The relationship between conditional expectations and conditional probabilities is more complex in the variational formulation of the metric conditional expectation in (8).

From the Rao–Blackwell theorem, the Lehmann–Scheffé theorem is easily obtained in a Euclidean setting by taking the conditional expectation of an unbiased estimator with respect to a complete sufficient statistic. We define a metric space point estimator $\delta(X)$ of $g(P) \in \mathcal{X}$ to be unbiased for the family $\mathcal{P}$ if $E_2(\delta(X)) = g(P)$ when $X \sim P$ for all $P \in \mathcal{P}$. The main obstacle toward extending Lehmann–Scheffé to a metric space case is that the tower rule does not hold in general for conditional Fréchet means: If $\mathcal{G} \subset \mathcal{H}$, then it will not always be the case that $E_2(E_2(\delta|\mathcal{H})|\mathcal{G}) = E_2(\delta|\mathcal{G})$ [66]. The reason for this is that $\mathcal{L}^2(\mathcal{G})$ and $\mathcal{L}^2(\mathcal{H})$ do not inherit any Hilbert space structure from $\mathcal{X}$ as they do in the Euclidean case. The Pythagorean theorem applied to nested vector subspaces cannot in general be applied to $\mathcal{L}^2(\mathcal{H}) \subset \mathcal{L}^2(\mathcal{G})$. It follows that even if δ is unbiased for $g(P)$, there is no guarantee that $E_2(\delta|\mathcal{G})$ will remain unbiased for $g(P)$. See the Supplementary Material [53] for an explicit example of this phenomenon. The general idea behind this example is to place a uniform distribution P on points x_1, x_2, x_3 that are chosen so that $[[x_1, x_2]_{1/2}, x_3]_{1/3} \neq E_2 P$. If $\mathcal{G}$ is the σ-algebra generated by the indicator function of $\{x_1, x_2\}$, then $E_2(E_2(X|\mathcal{G})) = [[x_1, x_2]_{1/2}, x_3]_{1/3}$ and the tower rule fails to hold.

The problem we will consider for the remainder of this work is the estimation of a Fréchet mean, $g(P) = E_2 P$, under squared distance loss. If multiple independent observations $X_1, \ldots, X_n \sim P$ are available, the plug-in sample Fréchet mean estimator defined by

$$\bar{X} := \operatorname*{argmin}_{x \in \mathcal{X}} \sum_{i=1}^{n} d(x, X_i)^2 \tag{9}$$

is typically used as an estimator of $E_2 P$. As $\bar{X}$ is the Fréchet mean of the empirical distribution it might be expected that $\bar{X}$ is unbiased for $E_2 P$ with $E_2 \bar{X} = E_2 P$. Again, this is not true in general. For a counterexample, see the Supplementary Material [53]. Under some mild regularity conditions, $\bar{X}$ is asymptotically unbiased as $\bar{X}$ converges in $L^2(\mathcal{X})$ to $E_2 P$ as $n \to \infty$ [61].

Due to the generality of Hadamard spaces, we will work with nonparametric families of distributions that only make mild assumptions on the Fréchet means and variances of random objects. Parametric alternatives do exist, notably the Riemannian normal distributions on a Riemannian manifold introduced by Pennec [56]. However, the Riemannian normal distribution can be challenging to work with as its Fréchet variance is in general related in a complex, nonlinear way to the scale parameter of the distribution and may even depend on the Fréchet mean.

4. Shrinkage estimators in Hadamard spaces. Suppose that one wishes to estimate the Fréchet mean of a distribution P on the Hadamard space $(\mathcal{X}, d)$ under the squared distance loss. Given one observation $X \sim P$, if it is suspected that $\theta := E_2 X$ is close to the point ψ in $\mathcal{X}$ then as an alternative to estimating θ with X one can instead estimate θ with the shrinkage estimator $[X, \psi]_t$ for some $t \in [0, 1]$. Even in the absence of strong prior information about θ, the risk of this shrinkage estimator is always smaller than the squared distance risk of the

estimator X for an appropriate choice of t. If $V_2X = \sigma^2 > 0$, applying the CAT(0) bound in (5) to the estimator $[X, \psi]_t$ gives

$$E\big(d\big(\theta, [X, \psi]_t\big)^2\big) \le (1-t)\sigma^2 + td(\theta, \psi)^2 - t(1-t)E\big(d(X, \psi)^2\big). \tag{10}$$

The right-hand side of (10) is a convex, quadratic function of t. It is seen that if t is small enough, the right-hand side of (10) is less than σ^2 and for such a t, $R(P, [X, \psi]_t) < R(P, X)$. Thus, there always exists an oracle estimator $[X, \psi]_t$ that is strictly closer to θ in $L^2(\mathcal{X})$ than X is.

In high-dimensional Euclidean spaces, the triangle $\Delta\theta X\psi$ is approximately a right triangle with hypotenuse $[X, \psi]$ [7, 64]. The James–Stein estimator can be viewed as an estimate of the projection of θ onto the line segment $[X, \psi]$. The same intuition holds in a Hadamard space where we seek an estimator on the segment $[X, \psi]$ that is close to θ. Moreover, as shown in Figure 1, the sides of comparison triangles in Hadamard spaces with negative curvature bend inward. Due to this curvature, the points $[X, \psi]_t$ will be even closer to θ than the corresponding points in a Euclidean space would be. Another consequence of metric convexity that motivates the use of shrinkage estimators in Hadamard spaces is the bias-variance decomposition in (7). As long as the distribution of X is nondegenerate, $E(d(X, \psi)^2) > d(E_2X, \psi)^2$ so that $d(X, \psi)^2$ on average overestimates the squared distance of ψ from E_2X. To correct this, the estimator $[X, \psi]_t$ is closer to ψ than X is.

For a given ψ, the central question is how should one go about choosing t in $[X, \psi]_t$. The optimal value of t that minimizes the upper bound of the risk in (10) is

$$\tilde{t} := \frac{\sigma^2 + \rho(X, \psi)^2 - d(\theta, \psi)^2}{2\rho(X, \psi)^2}, \tag{11}$$

where we use the notation $\rho(X, \psi)^2 = E(d(X, \psi)^2)$ with ρ being the metric on the Hadamard space $L^2(\mathcal{X})$ defined in Section 2.2. We call $\tilde{t}$ the oracle shrinkage weight although it only minimizes the risk upper bound, not the risk function. The Hadamard bias-variance inequality (7) shows that $\rho(X, \psi)^2 - d(\theta, \psi)^2 \ge \sigma^2$ so that $\tilde{t} \ge \sigma^2/\rho(X, \psi)^2$. Using a plug-in estimate for $\rho(X, \psi)^2$, the shrinkage weight

$$w(X) := 1 \wedge \big(\sigma^2/d(X, \psi)^2\big) \tag{12}$$

serves as an estimate of this lower bound for $\tilde{t}$. In order to use this shrinkage weight, the Fréchet variance σ^2 must be a known quantity. As long as $d(X, \psi)^2$ is sufficiently concentrated around $\rho(X, \psi)^2$ the weight $w(X)$ will tend to underestimate $\tilde{t}$. This reduces the possibility of overshrinking X when using the estimator $[X, \psi]_{w(X)}$.

The following theorem compares shrinkage estimators of the form $[Y, \tilde{\psi}]_{\alpha(Y)}$ in $\mathbb{R}^p$ where $\alpha(Y)$ is an arbitrary weight function, to corresponding shrinkage estimators in the p-dimensional hyperbolic space $\mathbb{H}^p$ and demonstrates the notion that shrinkage is especially beneficial in negatively curved spaces. The space $\mathbb{H}^p$ is the "canonical" example of a negatively curved space. We refer the reader to differential geometry texts such as [20, 45] for further information on $\mathbb{H}^p$, tangent spaces and the exponential map.

THEOREM 4.1. *Fix a $\theta \in \mathbb{H}^p$ and let $\exp_\theta : \mathbb{R}^p \to \mathbb{H}^p$ be the diffeomorphic exponential map from $T_\theta\mathbb{H}^p \cong \mathbb{R}^p$ onto $\mathbb{H}^p$, where $T_\theta\mathbb{H}^p$ is identified with $\mathbb{R}^p$ under some isometric isomorphism. Let Y be a random variable taking values in $\mathbb{R}^p$ with mean $\tilde{\theta}$ and take $X := \exp_\theta(Y - \tilde{\theta})$ so that $E_2X = \theta$. If $0 \le \alpha(Y) \le 1$ is an arbitrary weight function, $\tilde{\psi} \in \mathbb{R}^p$ and $\psi := \exp_\theta(\tilde{\psi} - \tilde{\theta})$ then*

$$\frac{E(d_{\mathbb{H}^p}(\theta, [X, \psi]_{\alpha(Y)})^2)}{E(d_{\mathbb{H}^p}(\theta, X)^2)} \le \frac{E(d_{\mathbb{R}^p}(\tilde{\theta}, [Y, \tilde{\psi}]_{\alpha(Y)})^2)}{E(d_{\mathbb{R}^p}(\tilde{\theta}, Y)^2)}.$$

The above inequality is strict if $P(\{0 < \alpha(Y) < 1\} \cap \{Y \notin \tilde{\theta} + \operatorname{span}(\tilde{\psi} - \tilde{\theta})\}) > 0$. *An analogous statement holds if* $\mathbb{H}^p$ *is replaced with any complete, connected, p-dimensional Riemannian manifold with nonpositive sectional curvature.*

A more general extension of this theorem that applies to tangent cones in Hadamard spaces is provided in the Supplementary Material [52]. This theorem roughly states that the relative improvement of a shrinkage estimator over the natural unbiased estimator is larger in hyperbolic spaces than it is in Euclidean spaces. As an example, suppose that X follows the Riemannian normal distribution on $\mathbb{H}^p$, $p \geq 3$ with Fréchet mean θ so that $X = \exp_\theta(Y)$, where $Y \sim N(0, \sigma^2 I)$ [56]. If $\alpha(Y) = 1 \wedge \sigma^2(p-2)/\|Y - \tilde{\psi}\|^2$ is the positive-part James–Stein estimator weight and $p \geq 3$ then Theorem 4.1 implies that $E(d_{\mathbb{H}^p}(\theta, [X, \psi]_{\alpha(Y)})^2) \leq E(d_{\mathbb{H}^p}(\theta, X)^2)$. Observe however that $[X, \psi]_{\alpha(Y)}$ is not an estimator, as knowledge of the basepoint θ of the exponential map is needed to compute $Y = \log_\theta(X)$, which in turn is used to find shrinkage weight $\alpha(Y)$.

4.1. *Geodesic James–Stein estimator.* Shrinkage estimators are typically used in a setting where observations from different groups are available and information is shared between groups to improve the estimation of group-specific parameters. A multigroup Fréchet mean estimation problem is formulated by first supposing that we have random objects $X = (X_1, \ldots, X_n)$ where each X_i lies in the Hadamard space $(\mathcal{X}_i, d_i)$, has Fréchet mean θ_i, a known Fréchet variance σ_i^2, and is independent of the other X_j's. The decision problem we consider for the remainder of this article is the simultaneous estimation of the collection of Fréchet means $\theta = (\theta_1, \ldots, \theta_n)$ under the loss function

$$L(\theta, \delta(X)) = \sum_{i=1}^{n} d_i(\theta_i, \delta_i(X))^2/n.$$

This problem formulation is the same as the classical James–Stein estimation problem in the special case when $\mathcal{X}_i = \mathbb{R}$ for each i and $X_i \sim N(\theta_i, \sigma^2)$ independently for $i = 1, \ldots, n$. Notice that like the classical James–Stein problem, there is no relationship assumed between the various θ_i's and the X_i's are independent and may not even take values in the same Hadamard space.

The simultaneous point estimation problem can be viewed as estimating a single point in a larger Hadamard space.

DEFINITION 4.2. The product Hadamard space of the Hadamard spaces $\{(\mathcal{X}_i, d_i)\}_{i=1}^n$ is the set $\mathcal{X}^{(n)} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ with the metric d given by

$$d(x, y) := \left(\sum_{i=1}^{n} d_i(x_i, y_i)^2/n \right)^{1/2},$$

where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$.

The multiplicative factor $n^{-1/2}$ is included in the product metric to ease notation. Geodesics in $(\mathcal{X}^{(n)}, d)$ are given pointwise by $[x, y]_t = ([x_1, y_1]_t, \ldots, [x_n, y_n]_t)$, and the CAT(0) inequality follows from the form of $d(x, y)$ so that $\mathcal{X}^{(n)}$ is also a Hadamard space. The collection of observations $X = (X_1, \ldots, X_n)$ is a random object in $\mathcal{X}^{(n)}$ with Fréchet mean $\theta = (\theta_1, \ldots, \theta_n)$ and (total) Fréchet variance $\sigma^2 := V_2 X = \sum_i V_2 X_i/n = \sum_i \sigma_i^2/n$. The simultaneous point estimation problem is to estimate $E_2 X = \theta$ under the loss function $L(\theta, \delta(X)) = d(\theta, \delta(X))^2$, which is exactly the Fréchet mean estimation problem introduced

in Section 3. There is an added nuance to this problem as the independence assumption on the X_i's implies that X must follow a product distribution on $\mathcal{X}^{(n)}$.

By viewing X as an element of the product Hadamard space $\mathcal{X}^{(n)}$, we can form the shrinkage estimator $[X, \psi]_{w(X)}$ with shrinkage weight (12) introduced at the beginning of this section.

DEFINITION 4.3. The geodesic James–Stein estimator $\delta_{JS}(X) := [X, \psi]_{w(X)}$ with shrinkage point $\psi = (\psi_1, \ldots, \psi_n)$ is the $\mathcal{X}^{(n)}$ valued estimator with components given by

$$\delta_{JS}(X)_j := [X_j, \psi_j]_{w(X)}, \quad w(X) = \left(1 \wedge \frac{\sum_{i=1}^n \sigma_i^2}{\sum_{i=1}^n d_i(X_i, \psi_i)^2}\right). \tag{13}$$

In Euclidean space, each $\mathcal{X}_i = \mathbb{R}$ and the positive-part James–Stein estimator $\delta_+(X)$ for $X \sim N_n(\theta, \sigma^2 I)$ is closely related to $\delta_{JS}(X)$ since $\delta_+(X) = [X, \psi]_{1 \wedge \frac{n-2}{n}\sigma^2/d(X,\psi)^2}$. The only difference between $\delta_+(X)$ and $\delta_{JS}(X)$ is the factor $(n-2)/n$ appearing in the shrinkage weight of $\delta_+(X)$. This factor is a remnant of tailoring $\delta_+(X)$ to a Gaussian X. If $\mathcal{X}_i = \mathbb{R}^p$ for all i, then δ_{JS} will shrink every component of the vector X_i toward the vector ψ_i by the same proportion $w(X)$. Consequently, δ_{JS} does not directly focus on the features associated the possibly exotic and high-dimensional object $X_i \in \mathcal{X}_i$. Instead, it focuses on the overall metric properties of each X_i.

Applications of the geodesic James–Stein estimator most commonly arise when one observes metric space random objects lying within the same metric space from different populations. This setting can be thought of as a metric space counterpart to the normal hierarchical model, but without any normality assumptions. If it is suspected that these populations are similar but not the same, the estimator δ_{JS} can be employed as a means of sharing information across populations. More precisely, if population j consists of the observations $\{X_{1j}, \ldots, X_{n_j j}\}$ then the X_j appearing in (13) will represent the sample Fréchet mean of these observations. A natural choice of ψ_j in this setting is to set it equal to the Fréchet mean of the X_j's. If σ_i^2 is unknown the sample Fréchet variance can be used as an estimator. Alternatively, as we discuss in the next section, it suffices to find a lower bound for these Fréchet variances. After making these substitutions, the resulting James–Stein estimator is an analogue to the conditional expectation estimator of means in a normal hierarchical model. There are some challenging technical considerations involved in evaluating the performance of this estimator: sample Fréchet means are not unbiased and the statistical behavior of the sample Fréchet variance is complex. In the sections to follow, we focus on the existence of an asymptotic Stein effect for δ_{JS}. Broadly, our results demonstrate that distributions of the X_j's need not even be related for δ_{JS} to have a smaller risk than the unbiased estimator $(X_1, \ldots, X_n)$.

4.2. *James–Stein risk comparison.* The Gaussian James–Stein estimator dominates X in squared error loss as long as the Gaussian distribution takes values in $\mathbb{R}^n$ with $n \geq 3$ [39, 64]. Similarly, we will be primarily interested in the behavior of $R(P, \delta_{JS})$ as the dimension n of the Hadamard space $\mathcal{X}^{(n)}$ increases. In typical applications, each X_i takes values in the same Hadamard space $\mathcal{X}$, so that $\mathcal{X}_i = \mathcal{X}$ for all i and $\mathcal{X}^{(n)} = \mathcal{X}^n$. To emphasize the dimension n of the Hadamard space $\mathcal{X}^{(n)}$ that X, θ and ψ lie in, we denote these objects by $X^{(n)}, \theta^{(n)}$ and $\psi^{(n)}$. Moreover, when examining how n effects the behavior of δ_{JS} it is helpful to assume that we have a sequence of random objects $\{X^{(n)}\}_{n=1}^\infty$ with corresponding Fréchet means $\{\theta^{(n)}\}_{n=1}^\infty$, as well as a sequence of shrinkage points $\{\psi^{(n)}\}_{n=1}^\infty$. Note that $X^{(k)}$ and $X^{(n)}$ for $k < n$ may be completely unrelated and similarly for $\psi^{(k)}$ and $\psi^{(n)}$.

An upper bound for the loss function of $\delta_{JS}(X^{(n)})$ can be found by plugging in the expression for $[X^{(n)}, \psi^{(n)}]_{w(X^{(n)})}$ into the CAT(0) bound, (5). Defining A to be the set $\{X^{(n)} : \sigma^2 < d(X^{(n)}, \psi^{(n)})^2\}$, which is equal to $\{X^{(n)} : w(X^{(n)}) < 1\}$, it is found that

$$\begin{aligned} L\big(\theta^{(n)}, \delta_{JS}(X^{(n)})\big) &\le I_A\big[(1 - w(X^{(n)}))d(X^{(n)}, \theta^{(n)})^2 + w(X^{(n)})d\big(\theta^{(n)}, \psi^{(n)}\big)^2 \\ &\quad - w(X^{(n)})(1 - w(X^{(n)}))d(X^{(n)}, \psi^{(n)})^2\big] + I_{A^c}d\big(\theta^{(n)}, \psi^{(n)}\big)^2 \\ &= \big[I_A(1 - w(X^{(n)}))\big(d(X^{(n)}, \theta^{(n)})^2 - \sigma^2\big)\big] \\ &\quad + \big[I_A w(X^{(n)})d\big(\theta^{(n)}, \psi^{(n)}\big)^2\big] + \big[I_{A^c}d\big(\theta^{(n)}, \psi^{(n)}\big)^2\big] \\ &:= (a) + (b) + (c). \end{aligned} \tag{14}$$

Notice that the denominator of $I_A w(X^{(n)})$ cancels with $d(X^{(n)}, \psi^{(n)})^2$ so that $I_A w(X^{(n)})d(X^{(n)}, \psi^{(n)})^2 = I_A\sigma^2$, which makes (14) take a reasonably simple form. Heuristically, as $n \to \infty$ by the law of large numbers we expect $d(X^{(n)}, \theta^{(n)})^2 - \sigma^2 \to 0$ and $d(X^{(n)}, \psi^{(n)})^2 - \rho(X^{(n)}, \psi^{(n)})^2 \to 0$. As a result, the term (a) should vanish and since $E(d(X^{(n)}, \psi^{(n)})^2) \ge \sigma^2 + d(\theta^{(n)}, \psi^{(n)})^2$ it is expected that $I_A \to 1$ so that (c) vanishes. Furthermore, $w(X^{(n)}) = I_A\sigma^2/d(X^{(n)}, \psi^{(n)})^2 + I_{A^c} \to \sigma^2/\rho(X^{(n)}, \psi^{(n)})^2$, which yields the approximate risk bound

$$R(P, \delta_{JS}) \lessapprox \sigma^2\frac{d(\theta^{(n)}, \psi^{(n)})^2}{\rho(X^{(n)}, \psi^{(n)})^2} \le \sigma^2\frac{d(\theta^{(n)}, \psi^{(n)})^2}{d(\theta^{(n)}, \psi^{(n)})^2 + \sigma^2} < \sigma^2 = R\big(P, X^{(n)}\big), \tag{15}$$

implying that δ_{JS} has a lower risk than $X^{(n)}$ under squared distance loss.

Regularity conditions on $d(X^{(n)}, \theta^{(n)})^2$ and $d(X^{(n)}, \psi^{(n)})^2$ are needed to ensure that these quantities are close enough to their respective means for large n. The main challenge of obtaining a domination result that is uniform over all choices of the shrinkage point $\psi^{(n)}$ is that $d(X^{(n)}, \psi^{(n)})^2$ can be highly variable. The variance of $d(X^{(n)}, \psi^{(n)})^2$ can be bounded below by a term involving $d(\theta^{(n)}, \psi^{(n)})$ and so it is large if the shrinkage point is chosen poorly so that $d(\theta^{(n)}, \psi^{(n)})$ is large. Restrictions are needed that limit how fast the sequence, $\{d(\theta^{(n)}, \psi^{(n)})\}_{n=1}^{\infty}$, can increase. Despite this, if $\psi^{(n)}$ is chosen to be far away from $\theta^{(n)}$ then $E(d(X^{(n)}, \psi^{(n)})^2)$ will be large which implies that almost no shrinkage will be applied and $\delta_{JS}(X^{(n)}) \approx X^{(n)}$.

The behavior of $d(X^{(n)}, \theta^{(n)})$ can be controlled by bounding its moments. Given a sequence $m := \{m_c\}_{c=1}^{\infty}$ of positive real numbers, for each n we define the family of probability distributions

$$\begin{aligned} \mathcal{P}_m^{(n)} := \big\{&P = P_1 \times \cdots \times P_n : V_2 P = \sigma^2, 0 < E\big(d_i(X_i, E_2 X_i)^c\big) \le m_c, \\ &X_i \sim P_i, c \in \mathbb{N}, i \in 1, \ldots, n\big\}. \end{aligned} \tag{16}$$

The family $\mathcal{P}_m^{(n)}$ is the set of product distributions on $\mathcal{X}^{(n)}$ that have a fixed Fréchet variance and have marginal distributions with "central-moments" that are bounded by the sequence m. Recall that the Fréchet variance $V_2 P$ is $\sum_{i=1}^n E(d_i(X_i^{(n)}, \theta_i^{(n)})^2)/n$, and so it is an average of the Fréchet variances of the marginal distributions. In $\mathbb{R}^n$, the family $\mathcal{P}_m^{(n)}$ corresponds to product distributions with $E(|X_i^{(n)} - EX_i^{(n)}|^c) \le m_c$ and $\sum_{i=1}^n \mathrm{Var}(X_i^{(n)})/n = \sigma^2$.

The following theorem generalizes the classical Gaussian James–Stein domination result to the large nonparametric family $\mathcal{P}_m^{(n)}$. A mild assumption is needed that constrains how fast $d(\theta^{(n)}, \psi^{(n)})^2$ can grow relative to the dimension of the Hadamard space $\mathcal{X}^{(n)}$. It will be shown that this assumption is automatically satisfied if the spaces $\mathcal{X}_i$ have uniformly bounded diameters. At the end of this section, we will further prove that δ_{JS} asymptotically dominates $X^{(n)}$ and has a loss function that is less than $\sigma^2 + \epsilon$ with probability tending to one, regardless of how fast $d(\theta^{(n)}, \psi^{(n)})^2$ grows.

THEOREM 4.4. *Let $\{a_n\}$ be a sequence with $a_n \to \infty$ and take $P \in \mathcal{P}_m^{(n)}$ to be any distribution on $\mathcal{X}^{(n)}$ with a Fréchet mean $\theta^{(n)}$ that satisfies $d(\theta^{(n)}, \psi^{(n)})^2 \leq n/a_n$. There exists an $n^*(m, \{a_n\})$ such that if $n \geq n^*$ then $R(P, \delta_{JS}) < R(P, X^{(n)})$.*

The main limitation of Theorem 4.4 is that the distribution of $X^{(n)} \in \mathcal{P}_m^{(n)}$ for $n \geq n^*$ must satisfy $d(\theta^{(n)}, \psi^{(n)})^2/n \leq a_n^{-1} = o(1)$, which is similar to a condition that appears in Brown and Kou [72] for a heteroskedastic normal model. Although more broadly applicable, this condition is most easily interpreted in terms of a sequence of random objects, $X^{(n)} \sim P^{(n)} \in \mathcal{P}_m^{(n)}$, $n \in \mathbb{N}$. For each n, choose a shrinkage point $\psi^{(n)}$ and suppose that $d(\theta^{(n)}, \psi^{(n)})^2/n \leq a_n^{-1}$ for all n. Theorem 4.4 guarantees that there exists an n^* such that $R(P^{(n)}, \delta_{JS}(X^{(n)})) < R(P^{(n)}, X^{(n)})$ for all $n \geq n^*$. In particular, if $\lim_n d(\theta^{(n)}, \psi^{(n)})^2/n \to 0$ then one can take $a_n^{-1} = d(\theta^{(n)}, \psi^{(n)})^2/n$. Recall that $d(\theta^{(n)}, \psi^{(n)})^2$ is an *average* of squared distances, $\sum_{i=1}^n d_i(\theta_i^{(n)}, \psi_i^{(n)})^2/n$. Therefore, $\lim_n d(\theta^{(n)}, \psi^{(n)})^2/n \to 0$ only requires that the average squared distance of the components of $\theta^{(n)}$ and $\psi^{(n)}$ increases at a rate that is slower than linear. Theorem 4.4 also shows that n^* does not depend on the particular sequence of $X^{(n)}$ chosen, rather it only depends on $\{a_n^{-1}\}$ and m.

Instead of starting with a sequence of random objects one can start with a sequence of shrinkage points, $\psi^{(n)}$. A dual way to view Theorem 4.4 is that given a sequence a_n^{-1} and m, δ_{JS} dominates $X^{(n)}$ over the subfamily, $\{P \in \mathcal{P}_m^{(n)} : d(E_2 P, \psi^{(n)})^2 \leq n a_n^{-1}\}$ of $\mathcal{P}_m^{(n)}$ for $n \geq n^*(m, \{a_n\})$. A special case occurs when the metric spaces $\mathcal{X}_i$ have uniformly bounded diameters, as for a large enough n this subfamily consists of all possible distributions on $\mathcal{X}^{(n)}$ with Fréchet variance σ^2. This follows by taking $a_n = \sqrt{n}$ and using the fact that $d(E_2 P, \psi^{(n)})^2 \leq \operatorname{diam}(\mathcal{X}^{(n)})^2 < \infty$. Moreover, the central moments $E(d_i(X_i^{(n)}, \theta_i^{(n)})^c)$ on a space with uniformly bounded diameter cannot be larger than $\operatorname{diam}(\mathcal{X}_i)^c$, which implies the following global domination result.

COROLLARY 4.5. *If the Hadamard spaces $\mathcal{X}_i, i \in \mathbb{N}$ are all bounded with $\operatorname{diam}(\mathcal{X}_i) \leq D$ for all i, then there exists an $n^*(D)$ such that $R(P, \delta_{JS}) < R(P, X^{(n)})$ for any distribution P on $\mathcal{X}^{(n)}$ and any shrinkage point $\psi^{(n)}$, when $n \geq n^*$.*

The estimator $X^{(n)}$ is thus inadmissible for estimating the Fréchet mean under a squared distance loss when the $\mathcal{X}_i$'s have uniformly bounded diameters and n is large enough. Notably, the dimension n^* in Corollary 4.5 is independent of any choices of $\psi^{(n)}$ or m. Intuition for Corollary 4.5 comes from (10) where it is seen that there always exists an amount of shrinkage where the shrinkage estimator has lower risk than $X^{(n)}$. Under the uniform boundedness assumption on the $\mathcal{X}_i$'s, the shrinkage weight $w(X^{(n)})$ concentrates around $\sigma^2/\rho(X^{(n)}, \psi^{(n)})^2$ closely enough for domination to occur independently of the choice of $\psi^{(n)}$.

Theorem 4.4 and Corollary 4.5 are remarkable since very few assumptions are made about the distribution of X, apart from assuming that the marginal distributions of $X^{(n)}$ have central moments bounded by m_c. On Euclidean spaces, the Stein estimator has been considered for a variety of classes of nonnormal distributions [8, 15]. One popular assumption is to take the distributional family to be spherically or elliptically symmetric [12, 13, 28, 43]. At a high level, these assumptions allow generalizations of Stein's lemma to be applied. When the metric is given by an inner product, Stein's lemma is used to control the term $2\langle X^{(n)} - \theta^{(n)}, \delta(X^{(n)}) - X^{(n)}\rangle$ that appears after expanding $R(P, \delta) = \|\delta - \theta^{(n)}\|^2$. In a general Hadamard space, there is no such decomposition of $d(\delta, \theta^{(n)})^2$. The assumption that the distribution of $X^{(n)}$ is spherically symmetric in $\mathbb{R}^n$ can be somewhat restrictive since this implies for example that the marginal distribution of each component X_i is the same.

An example of a subfamily of distributions on $\mathbb{R}^n$ that is contained in $\mathcal{P}_m^{(n)}$ is the following location family [46]: Let $F_i^{(n)}, i = 1, \ldots, n$ be distributions on $\mathbb{R}$ with mean 0, variance σ^2 and central moments bounded by the sequence $\{m_c\}_{c=1}^{\infty}$. The set of all distributions of random variables of the form $X^{(n)} = \theta^{(n)} + \epsilon^{(n)}$ for any $\theta^{(n)} \in \mathbb{R}^n$ and $\epsilon_i^{(n)} \sim F_i^{(n)}$ is contained in $\mathcal{P}_m^{(n)}$, because the location shifts $\epsilon_i^{(n)} \to \theta_i^{(n)} + \epsilon_i^{(n)}$ do not alter any of the central moments. This location family can be restricted further by assuming that for each n, $\theta^{(n)}$ is known to lie is some set $\Theta^{(n)}$ with $\operatorname{diam}(\Theta^{(n)}) \le D$. Theorem 4.4 implies that if $\psi^{(n)} \in \Theta^{(n)}$ for all n, then there exists a dimension $n^*(D, m)$ for which domination of $X^{(n)}$ occurs. Various results similar to this are known for distributions on $\mathbb{R}^n$ with restricted parameter spaces [51]. Immediate generalizations of this location family exist on arbitrary Hadamard spaces by letting the isometry group, instead of the translation group, act on a sequence of fixed distributions with bounded central moments.

Theorem 4.4 provides a domination result that applies to a subfamily of $\mathcal{P}_m^{(n)}$ for a finite number of groups. The geodesic James–Stein estimator also dominates X asymptotically over all of $P_m^{(n)}$ as $n \to \infty$.

THEOREM 4.6. *Let $X^{(n)} \sim P^{(n)} \in \mathcal{P}_m^{(n)}$ for all $n \in \mathbb{N}$. If $d(\theta^{(n)}, \psi^{(n)})^2 \to \infty$ for a sequence of shrinkage points $\{\psi^{(n)}\}_{n=1}^{\infty}$, then $\limsup_n R(P^{(n)}, \delta_{JS}(X^{(n)})) = \sigma^2$. It follows from Theorem 4.4 that $\limsup_n R(P^{(n)}, \delta_{JS}(X^{(n)})) \le \lim_n R(P^{(n)}, X^{(n)})$ for any sequence of $\psi^{(n)}$'s. Additionally, for all $\epsilon > 0$, $\lim_n P^{(n)}(L(\theta^{(n)}, \delta_{JS}(X^{(n)})) > \sigma^2 + \epsilon) = 0$.*

Theorem 4.6 makes explicit the observation that δ_{JS} behaves similar to $X^{(n)}$ when the shrinkage point is chosen to be far away from $E_2 X^{(n)}$. Consequently, in a simultaneous Fréchet mean estimation problem with a large number of groups the geodesic James–Stein estimator has performance that is comparable to, or much better than, the estimator $X^{(n)}$.

The results in this section also apply to estimators of the form $[X^{(n)}, \psi^{(n)}]_{\alpha w(X^{(n)})}$ where $\alpha \in (0, 1]$. Such estimators apply an amount of shrinkage that is proportional to, but less than δ_{JS}. It follows that

$$[X^{(n)}, \psi^{(n)}]_{\alpha w(X^{(n)})} = [X^{(n)}, [X^{(n)}, \psi^{(n)}]_{w(X^{(n)})}]_{\alpha} = [X^{(n)}, \delta_{JS}]_{\alpha},$$

from which the convexity of the squared distance function implies that

$$d\big(\theta^{(n)}, [X^{(n)}, \psi^{(n)}]_{\alpha w(X^{(n)})}\big)^2 \le (1-\alpha) d\big(\theta^{(n)}, X^{(n)}\big)^2 + \alpha d\big(\theta^{(n)}, \delta_{JS}\big)^2. \tag{17}$$

The risk of $[X^{(n)}, \psi^{(n)}]_{\alpha w(X^{(n)})}$ is therefore no larger than a convex combination of the risk of $X^{(n)}$ and the risk of δ_{JS}. Estimators of this form are useful when the value of σ^2 that appears in $w(X^{(n)})$ is not known but instead it is known that σ^2 is bounded below by $\alpha_0 > 0$, so that $\alpha_0/\sigma^2 \le 1$. By taking $\alpha = \alpha_0/\sigma^2$, the shrinkage weight $\alpha w(X^{(n)})$ is equal to $\alpha_0/d(X^{(n)}, \psi^{(n)})^2$ when the event $\{X^{(n)} : \sigma^2/d(X^{(n)}, \psi^{(n)})^2 \le 1\}$ occurs. Consequently, the estimator $[X^{(n)}, \psi^{(n)}]_{\tilde{w}}$ where $\tilde{w} = 1 \wedge \alpha_0/d(X^{(n)}, \psi^{(n)})^2$ will have the same large sample risk properties as δ_{JS}.

5. Analysis of the Bayes risk of δ_{JS}. Efron and Morris [25] show that the James–Stein estimator may be interpreted as an empirical Bayes procedure as follows: If $X^{(n)} \sim N_n(\theta^{(n)}, \sigma^2 I)$ and the prior distribution for $\theta^{(n)}$ is $\theta^{(n)} \sim N_n(\mu^{(n)}, \tau^2 I)$, then the posterior mean estimator of $\theta^{(n)}$ is the linear shrinkage estimator $(1-t)X^{(n)} + t\mu^{(n)}$, with $t = \sigma^2/(\sigma^2 + \tau^2)$. If an appropriate choice of τ^2 is not available, Efron and Morris suggest empirically estimating its value from the data. Specifically, they show that $(n-2)/\sum_{i=1}^n (X_i^{(n)} - \mu_i^{(n)})^2$ is an unbiased estimator of $1/(\sigma^2 + \tau^2)$ with respect to the marginal distribution of X.

Plugging this into the expression for t yields the James–Stein estimator δ_{JS}. Whereas Stein's results on risk concerned frequentist risk, that is, risk as a function of $\theta^{(n)}$, Efron and Morris obtained results on the Bayes risk, the average frequentist risk with respect to the prior distribution $\theta^{(n)} \sim N_n(\mu^{(n)}, \tau^2 I)$. They showed that not only is δ_{JS} better than $X^{(n)}$ with respect to Bayes risk, δ_{JS} is almost as good as the posterior mean estimator, which is Bayes-risk optimal. For any value of τ^2, the Bayes risk of δ_{JS} approaches that of the optimal posterior mean estimator as $n \to \infty$.

In this section, we consider similar results for the geodesic James–Stein estimator. We first examine the Bayes risk of the geodesic James–Stein estimator in the case that the shrinkage point is fixed at $\psi^{(n)}$. In this case, the Bayes risk is bounded above in terms of the distance between the shrinkage point $\psi^{(n)}$ and the prior Fréchet mean of $\theta^{(n)}$. If the dimension n is sufficiently large, δ_{JS} will have a smaller Bayes risk than $X^{(n)}$. However, there is no guarantee that the risk of δ_{JS} will asymptotically approach the minimum Bayes risk as $n \to \infty$. The absence of such a result is not surprising, since in general the Bayes estimator may not be a geodesic shrinkage estimator of the form $[X^{(n)}, \psi^{(n)}]_t$. For example, even for Euclidean sample spaces, Bayes estimators will not generally be linear shrinkage estimators unless the model is an exponential family and the prior distribution is conjugate [19].

Next, we compare the Bayes risk of $X^{(n)}$ to that of a potentially more useful shrinkage estimator, one for which the shrinkage point is empirically estimated from the data $X^{(n)}$. This is done in a setting that generalizes the simple hierarchical normal model $X^{(n)} \sim N_n(\theta^{(n)}, \sigma^2 I)$ and $\theta^{(n)} \sim N_n(\tilde{\mu} 1, \tau^2 I)$, where $\tilde{\mu} \in \mathbb{R}$ and 1 is an n-vector of all ones. Empirical Bayes estimation of both $\tilde{\mu}$ and τ^2 is possible since they are common to all elements of $\theta^{(n)}$ and, therefore, common to all elements of $X^{(n)}$. We consider an analogous scenario in which the prior Fréchet mean of each element of $\theta^{(n)}$ is equal to a common value $\tilde{\mu}$. Under this assumption, $\tilde{\mu}$ can approximately be estimated by the sample Fréchet mean $\bar{X}^{(n)}$ of $X_1^{(n)}, \ldots, X_n^{(n)}$. The resulting estimator δ_{JS} has a smaller Bayes risk than $X^{(n)}$, where unlike in the frequentist case, this result is global and does not only apply to a subfamily of $\mathcal{P}_m^{(n)}$. Recall that the primary difficulty in obtaining a global domination result of δ_{JS} over $X^{(n)}$ in the frequentist case was that the shrinkage point may be far away from $\theta^{(n)}$. By adaptively choosing the shrinkage point in the Bayesian setting, there is no longer this concern as $\bar{X}^{(n)}$ will be reasonably close to $\theta^{(n)}$ with high probability.

5.1. *Bayes risk of δ_{JS}.* Throughout this section, we work with a prior distribution $Q^{(n)} = Q_1^{(n)} \times \cdots \times Q_n^{(n)}$ for the estimand $\theta^{(n)} = (\theta_1^{(n)}, \ldots, \theta_n^{(n)})$, so that the components $\theta_i^{(n)}$ of $\theta^{(n)}$ are mutually independent under this prior distribution. Let $\mu^{(n)} \in \mathcal{X}^{(n)}$ be the Fréchet mean of $Q^{(n)}$ and take τ^2 to be the Fréchet variance of $Q^{(n)}$. Conditional on $\theta^{(n)}$ the distribution $P^{(n)}_{i,\theta_i^{(n)}}$ of $X_i^{(n)}$ is assumed to have Fréchet mean $\theta_i^{(n)}$ and Fréchet variance σ_i^2. Furthermore, we assume conditional independence of the $X_i^{(n)}$ given $\theta^{(n)}$ so that this conditional distribution is denoted by $P^{(n)}_{\theta^{(n)}} = P^{(n)}_{1,\theta_1^{(n)}} \times \cdots \times P^{(n)}_{n,\theta_n^{(n)}}$. Lastly, we assume some additional moment conditions so that $Q^{(n)} \in \mathcal{P}_l^{(n)}$, with $\mathcal{P}_l^{(n)}$ defined as in (16), for some sequence $l = \{l_c\}_{c=1}^{\infty}$ and $P_\theta^{(n)} \in \mathcal{P}_m^{(n)}$ for every $\theta \in \mathcal{X}^{(n)}$ for some sequence $m = \{m_c\}_{c=1}^{\infty}$. In summary, the joint distribution of X and θ has the form

$$
\begin{aligned}
\theta^{(n)} &\sim Q^{(n)} = Q_1^{(n)} \times \cdots \times Q_n^{(n)} \in \mathcal{P}_l^{(n)}, \\
E_2 Q^{(n)} &= \mu^{(n)}, \qquad V_2 Q^{(n)} = \tau^2, \\
X^{(n)} | \theta^{(n)} &\sim P^{(n)}_{\theta^{(n)}} = P^{(n)}_{1,\theta_1^{(n)}} \times \cdots \times P^{(n)}_{n,\theta_n^{(n)}} \in \mathcal{P}_m^{(n)}, \\
E_2 P^{(n)}_{\theta^{(n)}} &= \theta^{(n)}, \qquad V_2 P_\theta^{(n)} = \sigma^2.
\end{aligned} \tag{18}
$$

The results of this section remain nonparametric as they apply to any choice $Q^{(n)}$ and $P^{(n)}_{\theta^{(n)}}$ that satisfy (18). Notice that the model formulation in (18) still does not explicitly posit any relationship between the distributions of the various $(X_i^{(n)}, \theta_i^{(n)})$'s. As an example, the standard Gaussian hierarchical model is encompassed by (18) by taking $P^{(n)}_{\theta^{(n)}} = N_n(\theta^{(n)}, \sigma^2 I)$ and $Q^{(n)} = N_n(\mu^{(n)}, \tau^2 I)$.

As in Section 4, the estimation problem of interest is to estimate $\theta^{(n)}$ under squared distance loss where the only known quantities in (18) are $X^{(n)}$ and σ^2. Theorem 4.4 extends to this setting where a prior distribution is placed on $\theta^{(n)}$ by evaluating the performance of $\delta_{JS}(X^{(n)})$ in terms of its Bayes risk.

THEOREM 5.1. *Under the distributional assumptions in* (18), *suppose that there is a sequence* $a_n \to \infty$ *such that* $d(\mu^{(n)}, \psi^{(n)})^2 \leq n/a_n$. *There exists an* $n^*(m, l, \{a_n\})$ *such that if* $n \geq n^*$ *then the Bayes risk satisfies* $E(R(P^{(n)}_{\theta^{(n)}}, \delta_{JS})) < E(R(P^{(n)}_{\theta^{(n)}}, X^{(n)}))$.

The bound on the distance $d(\theta^{(n)}, \psi^{(n)})^2$ that appears in Theorem 4.4 is replaced by a bound on $d(\mu^{(n)}, \psi^{(n)})^2$ in Theorem 5.1. A special submodel of (18) where the condition $d(\mu^{(n)}, \psi^{(n)})^2/n = o(1)$ is easily satisfied is where $\mathcal{X}_i = \mathcal{X}$ for all i and $Q^{(n)}$ has the form $Q^{(n)} = \tilde{Q} \times \cdots \times \tilde{Q}$ for all n. Throughout this section, tildes will be used to denote points, metrics and distributions on $\mathcal{X}$ when $\mathcal{X}^{(n)} = \mathcal{X}^n$ is a Cartesian product of $\mathcal{X}$. If $\psi^{(n)} = (\tilde{\psi}, \ldots, \tilde{\psi})$ is chosen to have identical componentwise entries for all n, then $d(\mu^{(n)}, \psi^{(n)})^2 = \tilde{d}(\tilde{\mu}, \tilde{\psi})^2$ is constant over n and so it is $o(n)$. Using such a sequence of $\psi^{(n)}$'s, Theorem 5.1 guarantees the existence of an n^* for which δ_{JS} has a smaller Bayes risk than $X^{(n)}$ for $n \geq n^*$. The dimension that is needed for this smaller Bayes risk is still shrinkage point dependent since it is contingent on the value of $\tilde{d}(\tilde{\mu}, \tilde{\psi})^2$. In this case, we can write $n^*(m, l, \{a_n\})$ as $n^*(m, l, \tilde{d}(\tilde{\mu}, \tilde{\psi}))$.

Theorem 5.1 applies to the location family example introduced in the previous section where $X^{(n)} = \theta^{(n)} + \epsilon^{(n)}$. The only modification needed is that $\theta^{(n)}$ is now assumed to have the distribution $\theta_i^{(n)} \sim \tilde{Q} \in \mathcal{P}_l^{(1)}$ independently for $i = 1, \ldots, n$. Even in this specific example, the class of distributions on $\theta^{(n)}$ and $\epsilon^{(n)}$ to which these results hold is very broad. Suppose that the shrinkage point is taken to have equal componentwise entries, $\tilde{\psi}$. The same dimension $n^*(m, l, \tilde{d}(\tilde{\mu}, \tilde{\psi}))$ works for any mean zero error distribution of $\epsilon^{(n)}$ that is in $\mathcal{P}_m^{(n)}$ with $V_2\epsilon^{(n)} = \sigma^2$. Likewise, $n^*(m, l, \tilde{d}(\tilde{\mu}, \tilde{\psi}))$ applies to any distribution $\tilde{Q} \in \mathcal{P}_l^{(1)}$ as long as $\tilde{d}(E_2\tilde{Q}, \tilde{\psi}) \leq \tilde{d}(\tilde{\mu}, \tilde{\psi})$.

Theorem 4.6 can similarly be extended to a Bayesian setting.

THEOREM 5.2. *Let* $X^{(n)} \sim P^{(n)}_{\theta^{(n)}}$, $n \in \mathbb{N}$ *and* $E_2X^{(n)} = \theta^{(n)} \sim Q^{(n)}$, $n \in \mathbb{N}$ *satisfy the distributional assumptions in* (18). *If* $d(\mu^{(n)}, \psi^{(n)})^2 \to \infty$ *for a sequence of shrinkage points* $\{\psi^{(n)}\}_{n=1}^{\infty}$, *then* $\limsup_n E(R(P^{(n)}_{\theta^{(n)}}, \delta_{JS})) = \lim_n E(R(P^{(n)}_{\theta^{(n)}}, X^{(n)}))$. *By Theorem* 5.1, *for any sequence of* $\psi^{(n)}$*'s,* $\limsup_n E(R(P^{(n)}_{\theta^{(n)}}, \delta_{JS})) \leq \lim_n E(R(P^{(n)}_{\theta^{(n)}}, X^{(n)}))$, *with strict inequality if* $d(\mu^{(n)}, \psi^{(n)})^2/n = o(1)$. *Additionally, we have that for all* $\epsilon > 0$, $\lim_n P(L(\theta^{(n)}, \delta_{JS}) > \sigma^2 + \epsilon) = 0$.

It should be noted that the distributional assumptions in (18) do not constitute a fully Bayesian model since $P^{(n)}_{\theta^{(n)}}$ and the prior distribution $Q^{(n)}$, although constrained, are both left unspecified. By leaving $P^{(n)}_{\theta^{(n)}}$ and $Q^{(n)}$ unspecified the results above can be regarded as part of a robust Bayesian analysis that compares the Bayes risk of δ_{JS} to $X^{(n)}$ over a wide class of joint distributions for $(X^{(n)}, \theta^{(n)})$ [9]. A fully Bayesian model can be obtained from (18) if hyperpriors are placed on both $P^{(n)}_{\theta^{(n)}}$ and $Q^{(n)}$.

5.2. *Bayes risk for an adaptively chosen shrinkage point.* In scenarios where the distributions of $(X_i^{(n)}, \theta_i^{(n)}), i = 1, \ldots, n$ are exchangeable, it is reasonable to require that an estimator of $\theta^{(n)}$ be equivariant under permutations of indices. This symmetry consideration suggests that the shrinkage point $\psi^{(n)}$ used in δ_{JS} should have identical componentwise entries.

It is intuitively clear that a good choice of $\psi^{(n)}$ should be close to $\theta^{(n)}$ on average. In the proof of Theorem 5.2, it is shown that $\lim_n E[(a) + (c)] = 0$, for the terms (a), (c) in (14). We make the further assumption that for all $n \in \mathbb{N}$,

$$Q^{(n)} = \tilde{Q} \times \cdots \times \tilde{Q} \quad \text{and} \quad P^{(n)}_{\theta^{(n)}} = \tilde{P}_{\theta_1^{(n)}} \times \cdots \times \tilde{P}_{\theta_n^{(n)}}.$$

Therefore, the joint distribution of $(X_i^{(n)}, \theta_i^{(n)})$ is the same for each group. By the definition of $Q^{(n)}$, $\mu^{(n)} = (\tilde{\mu}, \ldots, \tilde{\mu})$, and if $\psi^{(n)} = (\tilde{\psi}, \ldots, \tilde{\psi})$ has identical componentwise entries, this implies

$$\begin{aligned}\limsup_{n\to\infty} E\big(R\big(P^{(n)}_{\theta^{(n)}}, \delta_{JS}\big)\big) &\le \limsup_{n\to\infty} E\left[I_A \frac{d(\theta^{(n)}, \psi^{(n)})^2}{d(X^{(n)}, \psi^{(n)})^2}\right]\sigma^2 = \frac{E(d(\theta^{(n)}, \psi^{(n)})^2)}{E(d(X^{(n)}, \psi^{(n)})^2)}\sigma^2 \\ &\le \frac{E(d(\theta^{(n)}, \psi^{(n)})^2)}{\sigma^2 + E(d(\theta^{(n)}, \psi^{(n)})^2)}\sigma^2.\end{aligned} \tag{19}$$

The second equality in (19) holds since the integrand is uniformly integrable because it is in $L^{1+\epsilon}(\mathbb{R})$ for some $\epsilon > 0$ since $I_A/d(X^{(n)}, \psi^{(n)})^2 \le 1/\sigma^2$. The strong law of large numbers shows that $d(\theta^{(n)}, \psi^{(n)})^2 \xrightarrow{a.s} E(d(\theta^{(n)}, \psi^{(n)})^2)$ and $d(X^{(n)}, \psi^{(n)})^2 \xrightarrow{a.s} E(d(X^{(n)}, \psi^{(n)})^2)$ from which the second equality follows. The last inequality is a result of the Hadamard bias variance inequality (7) applied to $E(d(X^{(n)}, \psi^{(n)})^2|\theta^{(n)})$. The upper bound of (19) is minimized over $\tilde{\psi}$ when $\tilde{\psi} = \operatorname{argmin}_{\tilde{\psi}\in\mathcal{X}} E(d(\theta^{(n)}, \psi^{(n)})^2) = \operatorname{argmin}_{\tilde{\psi}\in\mathcal{X}} E(\tilde{d}(\theta_1^{(n)}, \tilde{\psi})^2)$. By the definition of $E_2\theta_1^{(n)}$, $\tilde{\psi} = E_2\theta_1^{(n)} = \tilde{\mu}$ is the minimizer of the asymptotic risk upper bound in (19). At this optimal value of $\psi^{(n)}$, the asymptotic Bayes risk of δ_{JS} is at most $\tau^2/(\sigma^2 + \tau^2)$ percent of the risk of $X^{(n)}$. If either of the inequalities in (19) are strict, δ_{JS} may offer an even greater improvement over $X^{(n)}$.

The preceding discussion confirms the intuition that $\tilde{\psi}$ should be chosen so that it is close to $\tilde{\mu}$. From the observations $X^{(n)} = (X_1^{(n)}, \ldots, X_n^{(n)})$, an estimate of $\tilde{\mu}$ can be obtained by calculating the sample Fréchet mean (9) of $X^{(n)}$. In Euclidean space, the sample Fréchet mean is simply the sample mean. Under regularity conditions, the sample Fréchet mean of an independent and identically distributed sample $\{X_i^{(n)}\}_{i=1}^n$, converges in $L^2(\mathcal{X})$ to $E_2X_1^{(n)}$ as $n \to \infty$. Consequently, we propose using the data dependent shrinkage point, $\tilde{\psi} = \overline{X}^{(n)}$, where $\overline{X}^{(n)}$ is the sample Fréchet mean of $X_1^{(n)}, \ldots, X_n^{(n)}$. It may not, however, be the case that $E_2X_1^{(n)}$ is the asymptotically optimal point $\tilde{\mu}$. The point $\tilde{\mu}$ is defined by $\tilde{\mu} = E_2\theta_1^{(n)} = E_2(E_2(X_1^{(n)}|\theta_1^{(n)}))$, which is not guaranteed to equal $E_2X_1^{(n)}$ as the tower rule does not always hold in a general Hadamard space (see the Supplementary Material [53]).

It was shown in Theorem 5.1 that the dimension needed for δ_{JS} to outperform X, n^*, is a function of m, l and $\tilde{d}(\tilde{\mu}, \tilde{\psi})$. If $\overline{X}^{(n)}$ is sufficiently close to $E_2X_1^{(n)}$, then the n^* needed when using this adaptive shrinkage point will approximately be a function of m, l and $\tilde{d}(\tilde{\mu}, E_2X_1^{(n)})$. The bias-variance inequality shows that $\tilde{d}(\tilde{\mu}, E_2X_1^{(n)})^2 \le E(\tilde{d}(X_1^{(n)}, \tilde{\mu})^2)$, while the triangle inequality $\tilde{d}(X_1^{(n)}, \tilde{\mu}) \le \tilde{d}(X_1^{(n)}, \theta_1^{(n)}) + \tilde{d}(\theta_1^{(n)}, \tilde{\mu})$ can be used to show that $\tilde{d}(\tilde{\mu}, E_2X_1^{(n)})$ can be bounded above entirely in terms of m and l. The next theorem makes this reasoning precise and proves the existence of an $n^*(m, l)$ for which the James–Stein estimator with an adaptive shrinkage point has a smaller Bayes risk than X.

THEOREM 5.3. *Assume that* $X^{(n)} \sim P^{(n)}_{\theta^{(n)}} = \tilde{P}_{\theta_1^{(n)}} \times \cdots \times \tilde{P}_{\theta_n^{(n)}}$ *and* $\theta^{(n)} \sim Q^{(n)} = \tilde{Q} \times \cdots \times \tilde{Q}$ *for all* $n \in \mathbb{N}$. *If* $E(\tilde{d}(\overline{X}^{(n)}, E_2 X_1^{(n)})^2) = O(n^{-1})$ *with the multiplicative constant in* $O(n^{-1})$ *only depending on* m *and* l, *then there exists an* $n^*(m,l)$ *such that for* $n \geq n^*$ *then* $E(R(P^{(n)}_{\theta^{(n)}}, \delta_{JS})) < E(R(P^{(n)}_{\theta^{(n)}}, X^{(n)}))$, *where* δ_{JS} *is the adaptive shrinkage estimator given by* (13) *with* $\psi_i^{(n)} = \overline{X}^{(n)}$. *Furthermore, the same* n^* *is valid for any distributions* $\tilde{P}^{(n)}_{\theta_i} \in \mathcal{P}_m^{(1)}$ *and* $\tilde{Q}^{(n)} \in \mathcal{P}_l^{(1)}$.

This result demonstrates that by choosing the shrinkage point adaptively there is no longer any concern that $d(\mu^{(n)}, \psi^{(n)})^2$ grows at too fast a rate. The shrinkage point $\overline{X}^{(n)}$ is on average close enough to $\tilde{\mu}$ so that it is beneficial to shrink $X^{(n)}$ toward $\overline{X}^{(n)}$. Fixing the conditional distribution $P^{(n)}_{\theta^{(n)}}$, Theorem 5.3 shows that δ_{JS} has a strictly smaller $\mathcal{P}_l^{(n)}$-Bayes risk $\sup_{Q^{(n)} \in \mathcal{P}_l^{(n)}} E(R(P^{(n)}_{\theta^{(n)}}, \delta_{JS}))$ than $X^{(n)}$ for $n \geq n^*$ [9].

The condition $E(\tilde{d}(\overline{X}^{(n)}, E_2 X_1^{(n)})^2) = O(n^{-1})$ in Theorem 5.3 is not overly restrictive. For example, if $\mathcal{X}$ is a Hilbert space then $E(\tilde{d}(\overline{X}^{(n)}, E_2 X_1^{(n)})^2) = (\sigma^2 + \tau^2)/n$. More generally, it is shown in [61] that if $\mathcal{X}$ satisfies the entropy condition $\sqrt{\log(N(B_\alpha(\mu), r))} \leq c(\alpha/r)^s$ for any $\alpha, r > 0$ and fixed numbers $c, s \in \mathbb{R}^+$ with $s < 1$ then the desired condition holds with a multiplicative constant that only depends on m and l. The number $N(B_\alpha(\mu), r)$ is defined as the covering number of the ball of radius α centered at μ by balls of radius r. Many spaces of interest, such as the metric tree space with vertex degrees that are bounded above and edge lengths that are bounded below, will satisfy this covering number condition. In fact, it is not fully necessary that $E(\tilde{d}(\overline{X}^{(n)}, E_2 X_1^{(n)})^2)$ be $O(n^{-1})$ for the conclusion of Theorem 5.3 hold; all that is needed is $E(\tilde{d}(\overline{X}^{(n)}, E_2 X_1^{(n)})^2) = o(1)$. However, in such a case the $n^*(m,l)$ needed will also depend on the rate of convergence of $E(\tilde{d}(\overline{X}^{(n)}, E_2 X_1^{(n)})^2)$ to zero.

5.3. *Asymptotic optimality of* δ_{JS}. As mentioned, it is too much to expect that δ_{JS} asymptotically attain the optimal Bayes risk for a given sampling model, as a Bayes estimator may not take the form of a shrinkage estimator. The asymptotic Bayes risk of δ_{JS} can instead be compared against the risk of the best possible shrinkage estimator. We define the minimum shrinkage Bayes risk of the model in 5.2 as

$$\inf_{\tilde{\psi} \in \mathcal{X}, t \in [0,1]} E\big(d\big([X^{(n)}, \psi^{(n)}]_t, \theta^{(n)}\big)^2\big).$$

The same derivation used in (11) shows that for a given $\psi = (\tilde{\psi}, \ldots, \tilde{\psi})$ the shrinkage weight that minimizes the CAT(0) upper bound is

$$\tilde{t} = \frac{\sigma^2 + \rho(X^{(n)}, \psi^{(n)})^2 - \rho(\theta^{(n)}, \psi^{(n)})^2}{2\rho(X^{(n)}, \psi^{(n)})^2}. \tag{20}$$

As the James–Stein shrinkage weight $w(X)$ converges to $\sigma^2/\rho(X^{(n)}, \psi^{(n)})^2$ almost surely, δ_{JS} only minimizes the CAT(0) bound asymptotically if $\rho(X^{(n)}, \psi^{(n)})^2 - \rho(\theta^{(n)}, \psi^{(n)})^2 = \sigma^2$. If $\mathcal{X}$ has negative curvature, it is typical that $\rho(X^{(n)}, \psi^{(n)})^2 - \rho(\theta^{(n)}, \psi^{(n)})^2 > \sigma^2$ so that δ_{JS} asymptotically performs less shrinkage than is needed to minimize the CAT(0) bound.

Determining the minimizer of the CAT(0) bound with respect to ψ is more complex. If the above value of $\tilde{t}$ is substituted into the CAT(0) bound, then the resulting expression is

$$\begin{aligned}
&\tilde{t}\sigma^2 + (1-\tilde{t})\rho\big(\theta^{(n)}, \psi^{(n)}\big)^2 - \tilde{t}(1-\tilde{t})\rho\big(X^{(n)}, \psi^{(n)}\big)^2 \\
&\quad = \rho\big(\theta^{(n)}, \psi^{(n)}\big)^2 - \frac{(\rho(\theta^{(n)}, \psi^{(n)})^2 + \rho(X^{(n)}, \psi^{(n)})^2 - \sigma^2)^2}{4\rho(X^{(n)}, \psi^{(n)})^2}.
\end{aligned}$$

The above expression can also be simplified in the special case when $\rho(X^{(n)}, \psi^{(n)})^2 = \rho(\theta^{(n)}, \psi^{(n)})^2 + \sigma^2$, where it equals $\sigma^2\rho(\theta^{(n)}, \psi^{(n)})^2/(\sigma^2 + \rho(\theta^{(n)}, \psi^{(n)})^2)$. In this case, it is seen that the optimal choice of $\psi^{(n)}$ is $E_2\theta^{(n)}$ as this minimizes $\rho(\theta^{(n)}, \psi^{(n)})^2$. The condition $\rho(X^{(n)}, \psi^{(n)})^2 = \rho(\theta^{(n)}, \psi^{(n)})^2 + \sigma^2$ is satisfied in any Hilbert space, as this is just the bias-variance decomposition. Furthermore, the CAT(0) bound holds with equality in a Hilbert space so the shrinkage estimator minimizing the Bayes risk is the familiar estimator, $[X^{(n)}, E_2\theta^{(n)}]_{\sigma^2/(\sigma^2+\tau^2)}$. The tower rule also holds in a Hilbert space so $\overline{X}^{(n)} \to E_2\theta_1^{(n)}$ in $L^2(\mathcal{X})$. The bound in (19) thus shows that δ_{JS} attains the minimum Bayes shrinkage risk asymptotically in a Hilbert space. For example, in the location family example in $\mathbb{R}^n$, the Bayes risk of the adaptive James–Stein estimator approaches the minimum Bayes risk out of all linear estimators of $\theta^{(n)}$ as $n \to \infty$.

Without any additional assumptions on the metric in a Hadamard space with negative Alexandrov curvature, not much more can be said about the asymptotic optimality of δ_{JS}. The CAT(0) upper bound may not fully reflect the behavior of the risk function in such a space.

6. Simulations and examples. In this section, the empirical performance of the geodesic James–Stein estimator is examined in the space of phylogenetic trees and the space of symmetric positive definite matrices. Simulation results suggest that the n^* appearing Theorems 4.4 and 5.1 is not prohibitively large in practice.

6.1. *Estimation of gene trees in the BHV tree space.* A phylogenetic *species tree* describes the evolutionary history of a collection of different species. As the true evolutionary history is not known, it has to be inferred [26]. This is often done by comparing various DNA and amino acid sequences across the species. When comparing the nucleotide sequence of a single gene across species, an estimate of the *gene tree* can be reconstructed, where a gene tree chronicles the evolutionary history of the particular gene. Using these gene trees, one heuristic method for estimating the species tree is to average the gene trees by taking their Fréchet mean with respect to a chosen metric [54]. This estimator assumes that the gene trees on average resemble the species tree. However, gene trees may differ from each other and from the species tree due to factors such as horizontal gene transfer and incomplete lineage sorting [50, 69]. The estimation and modeling of gene trees is currently an active area of phylogenetics research [1, 60]. In the remainder of this section, we detail how the geodesic James–Stein estimator can be used to construct gene tree estimates, and provide simulation results for these estimates. Before doing this, we first discuss more precisely what a phylogenetic tree is and how to define a metric on the space of trees.

A (weighted and rooted) k-phylogenetic tree is a weighted graph with nonnegative edge weights that is a tree with k leaves (degree 1 vertices). The interior vertices of this tree have degree at least 3, except for the root vertex ρ, which possibly has degree two [62]. Each leaf in the tree is associated with a unique label from the set $\{1, \ldots, k\}$, while one of the interior vertices of the tree is labeled as the root. A k-phylogenetic tree can be interpreted as a species tree where the labels $\{1, \ldots, k\}$ represent extant species and interior vertices represent ancestral species, with the root vertex being a common ancestor to all species. Each interior vertex signifies a speciation event where one species diverges into two or more different species. The edge weights between a species and its immediate ancestor in a species tree quantifies the "amount of evolution" that has occurred between these species. If the rate of mutation is constant over time, the edge weights can be interpreted as the amount of time elapsed between speciation events.

The Billera–Holmes–Vogtmann (BHV) treespace $\mathcal{T}_k$ is a Hadamard space where each point in $\mathcal{T}_k$ is a k-phylogenetic tree [11]. For pedagogical purposes, we restrict our discussion

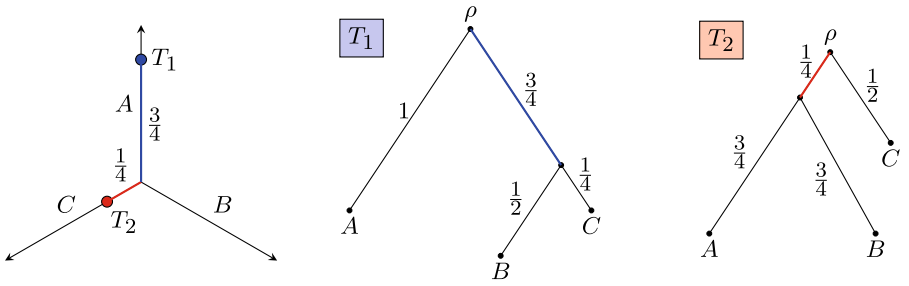

FIG. 2. *Left to right*: *The tripod space on the taxa* $\{A, B, C\}$. *The tree* T_1 *with* $(s, l_I, l_A, l_B, l_C) = (A, \frac{3}{4}, 1, \frac{1}{2}, \frac{1}{4})$, *with the interior edge highlighted in blue. The tree* T_2 *with* $(s, l_I, l_A, l_B, l_C) = (C, \frac{1}{4}, \frac{3}{4}, \frac{3}{4}, \frac{1}{2})$, *with the interior edge highlighted in red.*

to $\mathcal{T}_3$, denoting the three taxa by A, B and C. However, note that everything that follows applies to treespaces with $k > 3$. Any tree in $\mathcal{T}_3$ has four edges: three that have an endpoint which is a leaf and one interior edge that has the root as an endpoint. Given a collection of edge weights, the tree is determined by which of the three species diverged from the other two first, which is referred to as the *tree topology*. Therefore, any tree in $\mathcal{T}_3$ is determined by $(s, l_I, l_A, l_B, l_C) \in \{A, B, C\} \times \mathbb{R}^4_+$ where s is the species that first diverges from the root, l_I is the weight or length of the interior edge and $l_\alpha, \alpha \in \{A, B, C\}$ is the length of the edge connected to species α. Representing a tree by these data, the treespace metric is given by

$$d(T, T^*) = \begin{cases} \sqrt{(l_I - l_I^*)^2 + \sum_\alpha (l_\alpha - l_\alpha^*)^2} & \text{if } s = s^*, \\ \sqrt{(l_I + l_I^*)^2 + \sum_\alpha (l_\alpha - l_\alpha^*)^2} & \text{if } s \neq s^*. \end{cases} \tag{21}$$

Note that T and T^* represent the same tree if $l_I = l_I^* = 0$ and $l_\alpha = l_\alpha^*$, $\alpha \in \{A, B, C\}$. Geometrically, the interior edge of a tree can be represented as the tripod space that takes three rays and glues these rays together at their endpoints (see Figure 2). A point that is a distance of c away from the endpoint of ray α corresponds to trees with $s = \alpha$ and $l_I = c$. The space $\mathcal{T}_3$ is the Cartesian product of the tripod space with the nonnegative orthant $\mathbb{R}^3_+$, where the lengths of the noninterior edges of a tree are represented in this orthant. Higher-dimensional tree spaces are constructed similarly, although there are many more than three types of tree topologies when $k > 3$ [62].

The Jukes–Cantor model [40] is a basic model that describes the mutation of nucleotide sequences with respect to a given k-phylogenetic tree τ. Let $z \in \{A, T, C, G\}$ be a base pair at the root vertex ρ of τ and take v to be a vertex that is adjacent to the root and a distance of ℓ away. Under the Jukes–Cantor model the probability that the base pair at v is still z is $(1 + 3e^{-\ell})/4$ while the probability that z mutates into one of the three other base pairs is $(1 - e^{-\ell})/4$ for each of the other base pairs. Once a base pair is generated at v, the process is repeated for vertices adjacent to v that have not already been assigned a base pair until a base pair is has been placed at every vertex on τ. To generate an entire nucleotide sequence, this process is run independently for each base pair in the starting nucleotide sequence at ρ. In practice, one usually observes only the nucleotide sequences resulting from the Jukes–Cantor process at each leaf of the tree. From these sequences, the tree τ, which can be viewed as a parameter in the Jukes–Cantor model, can be estimated via maximum likelihood estimation or the method of moments [26, 38]. This process of estimating a gene tree motivates the distribution over trees that we consider in the simulations to follow.

Under the BHV distance, the gene tree estimation problem can be formulated as a problem of jointly estimating Fréchet means. Let $T := (T_1, \ldots, T_n) \in \mathcal{T}_k \times \cdots \times \mathcal{T}_k$ be an observation

of gene trees across k different species for n different genes. The goal of the gene tree estimation problem is to estimate the treespace Fréchet mean of $(E_2(T_1), \ldots, E_2(T_n))$ of T under the loss function $L(E_2(T), \delta) = \sum_{i=1}^{n} d(\delta_i, E_2(T_i))^2$. As $\mathcal{T}_k$ is a Hadamard space, the geodesic James–Stein estimator of $E_2(T)$ can be used as an alternative to the estimator T.

To compare the performance of $\delta_{JS}(T)$ to T, we compare the risk functions of these estimators in the treespace $\mathcal{T}_3$. Given n, fix a collection of trees $(\tau_1, \ldots, \tau_n) \in \mathcal{T}_3^n$. For each $i = 1, \ldots, n$, we generate a tree T_i corresponding to τ_i by:

1. Running the Jukes–Cantor process on τ_i with respect to a nucleotide sequence of length 50.
2. Taking the observation T_i to be a method of moments estimator of τ_i constructed from the above nucleotide sequences on the leaves of τ_i.

From the observed gene trees $(T_1, \ldots, T_n)$, we obtain Monte Carlo estimates of the risk of the estimator T and the geodesic James–Stein estimator. To efficiently explore the parameter space of trees, we draw independent observations of $\tau_i = (s_i, l_{I,i}, l_{A,i}, l_{B,i}, l_{C,i})$ with

$$
\begin{aligned}
&s_i \sim \text{Multinomial}(p_A, p_B, p_C), \qquad L_i \sim \text{Exp}(\lambda), \qquad U_i \sim \text{Unif}(0,1),\\
&l_{I,i} = U_i L_i, \qquad l_{s_i,i} = L_i, \qquad l_{\alpha,i} = (1-U_i)L_i, \quad \alpha \neq s_i
\end{aligned} \tag{22}
$$

The random variable s_i reflects the distribution of the tree topology of τ_i, L_i represents the distance from every leaf to the root node and U_i represents the ratio of length of the interior edge to the distance from the root to a leaf node. Our Monte Carlo study proceeds as follows:

1. Fix parameter values for (p_A, p_B, p_C) and λ and generate 50 batches of trees $(\tau_1, \ldots, \tau_n)$ from the aforementioned distribution with these parameter values.
2. For each of the 50 batches, we compute Monte Carlo estimates of the risk of the sample Fréchet mean and the geodesic James–Stein estimator.

Two different shrinkage points are used in the geodesic James–Stein estimator: the first is the sample Fréchet mean of $T_1, \ldots, T_n$, while the second corresponds to the tree $(A, 4, 8, 4, 4)$ with $s_i = A$, $L_i = 8$ and $U_i = 0.5$. The second shrinkage point is intentionally chosen so that it is not close to the typical location of the Fréchet means under (22). The results of this study are summarized in Table 1, which displays the average (Bayes) risk ratio across all batches of the James–Stein estimator to the estimator T. The proportion of the 50 batches where the James–Stein estimator has a smaller risk than T is also displayed in parentheses.

Table 1 demonstrates that in general, the Bayes risk of the geodesic James–Stein estimator is less than the Bayes risk of T. At $n = 50$, the conclusion of Theorem 5.1 holds empirically.

TABLE 1

Risk ratios $E(R(P_\theta, \delta_{JS}))/E(R(P_\theta, T))$ where P_θ is given by (22) with $\theta = (p_A, p_B, p_C, \lambda)$. In parentheses, the proportion of batches with $R(P_\theta, \delta_{JS}) < R(P_\theta, T)$

Shrinkage point		$\psi_i = \overline{T}$			$\psi_i = (A, 4, 8, 4, 4)$		
n		3	15	50	3	15	50
(p_A, p_B, p_C)	λ						
$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$\lambda = 0.5$	0.70 (0.74)	0.48 (1.00)	0.45 (1.00)	0.87 (0.82)	0.88 (1.00)	0.90 (1.00)
	$\lambda = 1$	0.79 (0.50)	0.61 (0.98)	0.54 (1.00)	0.92 (0.70)	0.94 (0.92)	0.95 (0.98)
	$\lambda = 2$	1.00 (0.22)	0.91 (0.52)	0.71 (0.98)	0.98 (0.52)	1.00 (0.44)	0.99 (0.64)
$(\frac{1}{10}, \frac{1}{10}, \frac{8}{10})$	$\lambda = 0.5$	0.67 (0.76)	0.50 (1.00)	0.44 (1.00)	0.87 (0.88)	0.89 (0.98)	0.89 (1.00)
	$\lambda = 1$	0.84 (0.46)	0.61 (1.00)	0.55 (1.00)	0.94 (0.74)	0.95 (0.80)	0.96 (0.96)
	$\lambda = 2$	1.05 (0.14)	0.89 (0.48)	0.69 (0.98)	0.99 (0.48)	0.99 (0.56)	0.99 (0.58)

This difference in risks is especially significant when the sample Fréchet mean shrinkage point is used in δ_{JS}, where the Bayes risk can be less than half of that of T. Even with poorly chosen, second shrinkage point, δ_{JS} has smaller Bayes risks than T. As the number of gene trees increases, the proportion of batches where δ_{JS} outperforms the sample Fréchet mean also increases. When using the sample Fréchet mean as the shrinkage point, at $n = 50$ the estimator δ_{JS} outperforms T for nearly every batch.

It is also seen from the table that the value of λ significantly influences the Bayes risk ratios. Simulations show that the ratio of the between-group Fréchet variance to the within-group Fréchet variance increases as λ increases. Therefore, the geodesic James–Stein estimator has better relative performance for small λ values, where shrinkage is especially helpful in reducing the variability of the tree estimates. Surprisingly, the multinomial probabilities have little impact on the Bayes risk ratios, in part because the these probabilities do not appreciably alter the within-to-between group Fréchet variance ratio.

One might wonder if the Euclidean orthant $\mathbb{R}^3_+$ portion of $\mathcal{T}_3$ is the primary factor that explains why the James–Stein estimator generally outperforms the sample Fréchet mean. A separate simulation study involving a random walk on a metric tree is provided in the the Supplementary Material [53]. The results of this study demonstrate that δ_{JS} performs especially well on this negatively curved metric tree space, which is formed by gluing multiple tripod spaces together. This agrees with the intuition that the inward bending comparison triangles in a negatively curved space enhance the performance of shrinkage relative to a space with Euclidean comparison triangles.

A property of sample Fréchet means for an observation in the treespace $\mathcal{T}_k$ and more generally on stratified spaces that has garnered recent interest is that the Fréchet mean can be "sticky" [36]. Stickiness in the tripod space roughly amounts to the tendency of the sample Fréchet means to lie exactly on the central vertex as the sample size increases. When the population Fréchet mean is also this same vertex, stickiness implies that for large sample sizes there is a high probability that the sample Fréchet mean is equal to the population Fréchet mean. Shrinkage is therefore only beneficial if the sample Fréchet mean does not stick to the population Fréchet mean. Despite this property of the sample Fréchet mean in the sticky regime, it can still be improved upon by the geodesic James–Stein estimator. That is, if there are n populations and a random sample of fixed size m is taken from each population, then under the conditions of Theorem 4.4, δ_{JS} outperforms the groupwise sample Fréchet mean estimator if n is large enough. Simulations demonstrating this are provided in the Supplementary Material [53]. When the sample Fréchet mean is sticky, its Fréchet variance is small so only a mild amount of shrinkage is used in δ_{JS}. Shrinkage estimation is most potent when sample sizes are small. For large sample sizes, δ_{JS} will behave almost identically to the sample Fréchet mean.

6.2. *Spatial smoothing of DTI imaging data.* Diffusion tensor imaging (DTI) is a medical imaging technique where the diffusion of water molecules in a tissue is measured under the influence of an external magnetic field. Various magnetic field gradients are applied, making it possible to estimate the prominent directions of diffusion at each tissue voxel. DTI is particularly useful for imaging the brain, as the primary directions of diffusion correlate with the directions of white matter fiber tracts. The DTI data we examine below consist of a collection of 3×3 symmetric positive definite (SPD) matrices, one at each voxel in three-dimensional space. Each matrix is an estimate of the covariance matrix of the diffusion process at the corresponding voxel. The eigenstructure of these SPD matrices is especially of interest. For example, in each voxel the principle eigenvector represents the primary axial direction of diffusion and the mean of the eigenvalues represents the overall diffusivity of the medium. The eigenstructure however can be sensitive to noise and so spatial smoothing is often applied to the SPD matrices [70]. One approach to smoothing is to use kernel-weighted sample

Fréchet means under either the log-Euclidean, affine-invariant or Euclidean distances within a window around a voxel [18]. The log-Euclidean and affine-invariant distances, respectively, are

$$d_{LE}(S_1, S_2) = \|\log(S_1) - \log(S_2)\|_F,$$
$$d_{AI}(S_1, S_2) = \|\log(S_1^{-1/2} S_2 S_1^{-1/2})\|_F,$$

where $\log(\cdot)$ is the matrix logarithm and $\|\cdot\|_F$ is the Frobenius norm [4, 57]. The collection of SPD matrices is a Hadamard space under both of these distances, where the log-Euclidean distance has vanishing Alexandrov curvature.

The basic kernel-weighted Fréchet mean estimator has the drawback that the same pattern of weights or kernel bandwidth is used at each voxel. As an alternative, the geodesic James–Stein estimator, which is known to have connections to smoothing in the Euclidean setting [41, 47], can be used to adaptively smooth DTI data. Shrinkage estimators of covariance matrices have been considered extensively in the literature where shrinking the eigenvalues of the sample covariance matrix can result in estimators that dominate the sample covariance matrix [35, 39, 44]. These results typically pertain to squared error loss or Stein's loss with a notable exception being the recent work [73] where the log-Euclidean distance is used as a loss function. Letting S_{ijk} be the observed SPD matrix at voxel (i, j, k), our objective is to jointly estimate the Fréchet means $E_2 S_{ijk}$ as (i, j, k) ranges over all voxels, under the loss $\sum_{ijk} d(\delta_{ijk}, E_2 S_{ijk})^2$. Define $\bar{S}_{ijk}$ to be the sample Fréchet mean of the 27 matrices in the $3 \times 3 \times 3$ window $B_{ijk} = \{(a, b, c) : \|(i, j, k) - (a, b, c)\|_{\ell_1} \le 1\}$, centered at voxel (i, j, k). Using $\bar{S}_{ijk}$ as a shrinkage point, we define the following variant of the geodesic James–Stein estimator:

$$\tilde{\delta}_{ijk} = [S_{ijk}, \bar{S}_{ijk}]_{w_{ijk}}, \quad w_{ijk} = 1 \wedge \frac{\hat{\sigma}^2_{B_{ijk}}}{\frac{1}{27}\sum_{(a,b,c)\in B_{ijk}} d(S_{abc}, \bar{S}_{ijk})^2}. \tag{23}$$

The distance d determines the type of geodesic used in $\tilde{\delta}$ and can be any one of d_{LE}, d_{AI} or the Euclidean distance. The quantity $\hat{\sigma}^2_{B_{ijk}}$ is an estimate of the total Fréchet variance $\sigma^2_{B_{ijk}} := \sum_{(a,b,c)\in B_{ijk}} V_2(S_{abc})/27$ of the tensors in B_{ijk}. Large values of $\hat{\sigma}^2_{B_{ijk}}$ result in the window smoother $\tilde{\delta}_{ijk} \approx \bar{S}_{ijk}$ while small values yield $\tilde{\delta}_{ijk} \approx S_{ijk}$. The only difference between the estimator $\tilde{\delta}$ and δ_{JS} is that a different shrinkage point is used at each voxel.

Before applying $\tilde{\delta}$ to real data we evaluate the performance of the geodesic James–Stein estimator on simulated tensor field data. This is done by perturbing two different template tensor fields T_{ijk} with noise. The two templates chosen are shown in Figure 3 and have the same checkerboard and circular patterns, respectively, for each (i, j) plane slice. The patterns have the respective dimensions $12 \times 12 \times 5$ and $7 \times 7 \times 5$. Simulated data S_{ijk} that are centered around the respective templates are generated according to one of the following log-Gaussian or Wishart distributions:

$$\log(S_{ijk}) \sim N\big(\log(T_{ijk}), 0.25I\big) \quad \text{or} \quad 8S_{ijk} \sim \text{Wishart}(T_{ijk}, 8).$$

Fifty such data sets are simulated for each noise model, and the geodesic James–Stein estimators (23) under the affine-invariant and log-Euclidean distances are computed for each data set. A total of 720 and 245 tensors are estimated in the checkerboard and circular templates, respectively, for each data set. Monte Carlo estimates are used to compute the Fréchet variances in (23), which are assumed to be known. The average value of $\sum_{l=1}^{3}(\log(\lambda_l) - \log(\hat{\lambda}_l))^2$ where λ_l is the lth eigenvalue of T_{ijk} and $\hat{\lambda}_l$ is the lth eigenvalue of various estimators of T_{ijk} is presented in Table 2. The average axial angle between the principle eigenvectors in the template and the various estimates of the template is also provided.

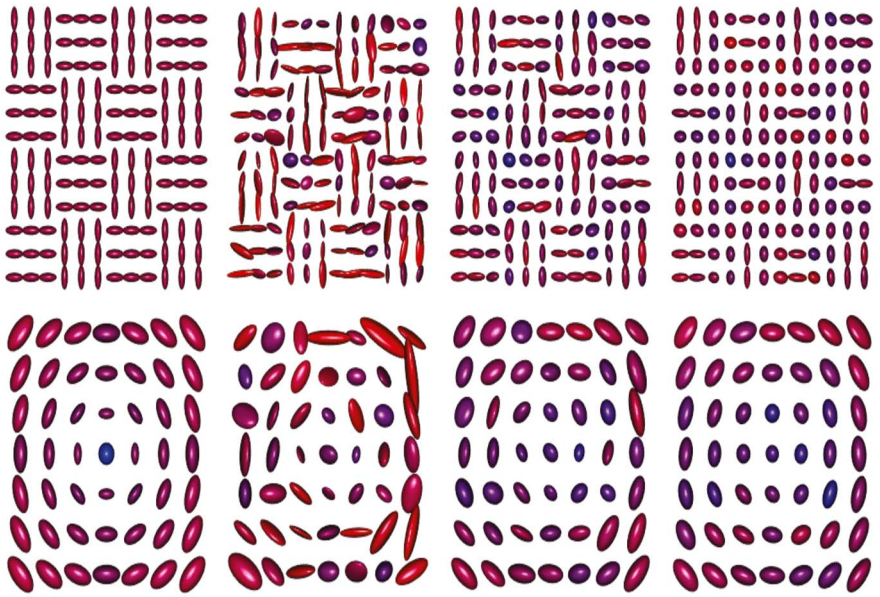

FIG. 3. *Left to right*: *template*, *template with added noise*, *affine-invariant geodesic James–Stein smoothing*, *log-Euclidean window smoother. Tensors are coloured by their fractional anisotropy.*

The results in Table 2 demonstrate that the different geodesic James–Stein estimator variants perform well, with the estimator that uses the log-Euclidean metric generally outperforming the estimator that uses the affine-invariant metric. They are both significantly better than the window smoother in the checkerboard template since the window smoother applies too much smoothing around the boundary of each square, making the estimated tensors in these voxels overly isotropic. In the circular template, the James–Stein and window estimators are comparable as both estimators smooth the data by a significant amount. The benefit of using the geodesic James–Stein estimator variant introduced here is that the shrinkage weight

TABLE 2
Risk of estimating eigenvalues and principle eigenvectors

Template	Noise	Estimator	Eigenvalue risk	Principle eigenvector risk
Checker	Log-Gauss	Noised Data	0.68	0.43
		LE-JS	0.26	0.33
		AI-JS	0.29	0.37
		Log-Window	0.71	0.45
Checker	Wishart	Noised Data	1.25	0.43
		LE-JS	0.55	0.33
		AI-JS	0.60	0.38
		Log-Window	0.92	0.46
Circle	Log-Gauss	Noised Data	0.79	0.47
		LE-JS	0.16	0.24
		AI-JS	0.21	0.30
		Log-Window	0.22	0.20
Circle	Wishart	Noised Data	1.43	0.47
		LE-JS	0.45	0.21
		AI-JS	0.52	0.31
		Log-Window	0.47	0.21

used in $\tilde{\delta}$ adapts to the amount of local variability of the Fréchet mean $E_2 S_{ijk}$ as a function of voxel location. Along boundary regions where the tensor structure is highly variable, w_{ijk} will be small and less shrinkage will be applied as compared to regions that are more uniform. This explains why $\tilde{\delta}$ shrinks by a large amount in the smoothly varying circular template and shrinks less in the checkerboard template.

We now apply the estimator (23) to the "Sherbrooke 3-shell" data set that is publicly available from the DIPY Python package [31]. The data are a $71 \times 88 \times 62$ spatial grid of 3×3 SPD matrices representing the estimated diffusion tensor at various locations in the brain of a single patient. The log-Euclidean metric is used in (23), as sample Fréchet means can be computed quickly in this metric. The remarks at the end of Section 4 suggest that $\delta_{JS}(S)$ will outperform S if a lower bound for $\sigma^2_{B_{ijk}}$ is used in the estimator (23). To find such a lower bound, we first make the simplifying assumption that $\sigma^2_{B_{ijk}} = \sigma^2$ is constant across (i, j, k). Due to the vanishing Alexandrov curvature of d_{LE}, if $E_2 S_{ijk} \approx E_2 S_{abc}$ then $E(d_{LE}(S_{ijk}, S_{abc})^2) \approx 2\sigma^2$. If $NN(ijk)$ is the index of the neighboring tensor of voxel (i, j, k) that is closest to S_{ijk} in d_{LE}, then we assume that $E_2 S_{ijk} \approx E_2 S_{NN(ijk)}$ and

$$\hat{\sigma}^2 = \frac{1}{2|R|} \sum_{(a,b,c)\in R} d(S_{abc}, S_{NN(abc)})^2$$

is taken as a conservative lower bound of σ^2 in $\tilde{\delta}$. The region R is chosen to be a representative portion of the brain that excludes the outer border of the image where the tensors are constant.

Figure 4 displays a two-dimensional slice of the field of principle eigenvectors for the original data, the window smoother that computes the log-Euclidean Fréchet mean within each $3 \times 3 \times 3$ window and $\tilde{\delta}$ under the log-Euclidean distance. Zero tensors where no diffusion occurs are not included when calculating the window mean in the window smoother. Moreover, we do not apply any shrinkage to voxels that contain zero tensors in the window smoother. These additional steps are taken to ensure that the window smoother is a reasonable estimator to which the James–Stein estimator can be compared against. The estimates

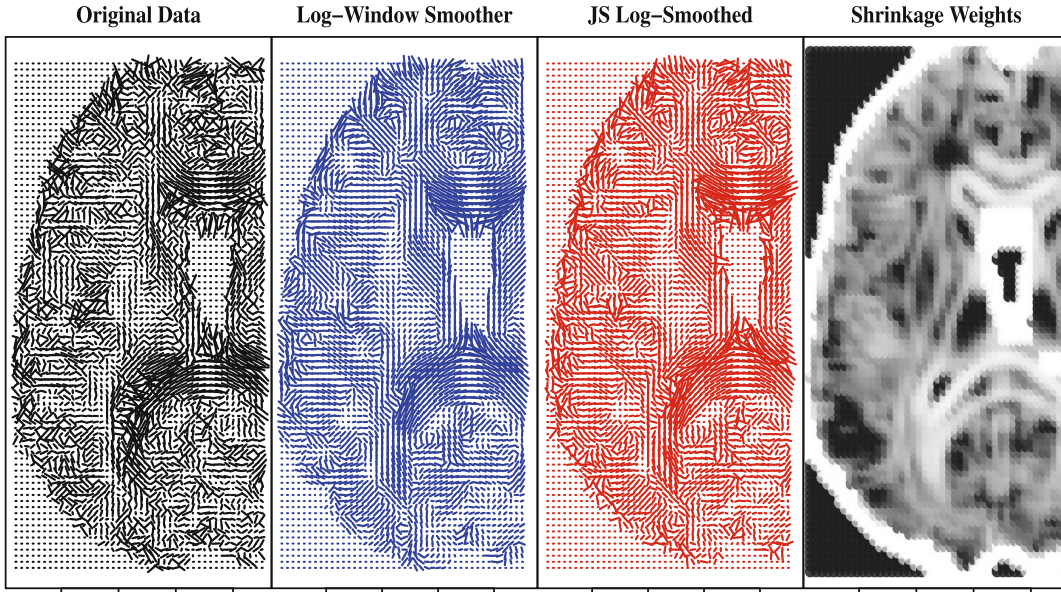

FIG. 4. *Principle eigenvectors in a transverse section of the brain and the shrinkage weights used in the log-Euclidean geodesic James–Stein estimator. Black represents weights that are* 1 (*total shrinkage*) *and white represents weights that are* 0 (*no shrinkage*).

of both $\tilde{\delta}$ and the window smoother depend on tensors lying in adjacent two-dimensional slices that are not pictured here. The rightmost plot in Figure 4 illustrates how the $\tilde{\delta}$ shrinkage weights w_{ijk} vary over the given slice. It is seen that in regions where the tensor field abruptly changes, such as the outer border of the brain and the center of the brain, almost no shrinkage is applied in $\tilde{\delta}$. Without the above modifications to how the window smoother treats zero tensors, the window smoother will smooth data in these regions excessively, blurring the boundaries apparent in the original data between regions with disparate structure. In summary, the log-Euclidean geodesic James–Stein estimator provides compromise between the two extremes of the not smoothing at all and utilizing a window smoother. It adaptively chooses the amount of smoothing to be applied an each voxel, which helps to mitigate the risk of oversmoothing in regions where the underlying mean tensor field is highly variable.

7. Discussion. In this article, we have primarily considered the risk properties of the geodesic James–Stein estimator for multiple Fréchet means. The primary result of this work, Theorem 4.4, shows that under mild conditions the geodesic James–Stein estimator outperforms X in a simultaneous Fréchet mean estimation problem if there are enough groups present and the shrinkage point is reasonably chosen. It is the nonpositive Alexandrov curvature of the metric space that forms the foundation of this result, as it implies that the squared distance function is metrically convex.

One may wonder if the results of this article can be extended to arbitrary geodesic metric spaces. In general the answer is no. To see this, consider the sphere $\mathbb{S}^2 \subset \mathbb{R}^3$ with its intrinsic, angular metric. The squared distance metric on the sphere is not metrically convex due to its positive sectional, and thus Alexandrov, curvature. For example, any two points x, y that lie on the equator of the sphere have $d([x, y]_t, N) = d(x, N) = d(y, N)$ for all $t \in [0, 1]$ where N is the north pole. As a result, no point of the geodesic $[x, y]$ is closer to N than x itself. A more extreme example on $\mathbb{S}^1$ is presented in Supplementary Material [53] where for a certain ψ and distribution of X, $[X, \psi]_t$ is has a larger risk than X for all $t > 0$. As $\mathbb{S}^1$ is compact, Corollary 4.5 fails to hold in a general metric space. Shrinkage may still be beneficial under specific circumstances. In the case of a Riemannian manifold, if a distribution is concentrated in a small enough region of the manifold, the effect of curvature on the metric will not be pronounced and results from the Euclidean case will approximately apply. If reliable prior information, suggesting that $E_2 X$ is close to ψ, is available then the shrinkage estimator $[X, \psi]_t$ will likely have reasonable performance even if the metric space has positive Alexandrov curvature.

Another extension of the geodesic James–Stein estimator presented here would be to cases where σ^2 is unknown and a plug-in estimator is used for σ^2 in the expression for the geodesic James–Stein estimator. The theoretical properties of such an estimator are more complex because multiple observations per group are required to obtain an estimate of σ^2. A property like the Hadamard bias-variance inequality will no longer be applicable since the sample Fréchet means of i.i.d. observations may not be unbiased for the underlying Fréchet mean. Results from [33, 34] further show that there is no Stein phenomenon for a family of distributions with finite support. More specifically, admissible estimators for individual decision problems remain admissible when combined into an estimator for the joint decision problem whose loss function is the sum of the losses for the individual problems. For example, if $X_i \sim \text{Bin}(n_i, \theta_i)$ then $(X_1, \ldots, X_n)$ is admissible for estimating $(\theta_1, \ldots, \theta_n)$ under squared error loss because X_i is admissible for estimating θ_i. This shows that Corollary 4.5 will not hold in general if σ^2 is unknown, since the estimator X is admissible in this binomial example. We again remark that σ^2 does not have to be known exactly in order to use δ_{JS}. Rather, all that is needed is a nonzero lower bound on σ^2 from which this lower bound can be used in place of σ^2 in (13). All the theoretical results in in Sections 4 and 5 will apply to the James–Stein estimator that uses such a lower bound, as shown by (17).

The hierarchical model introduced in Section 4 of this article represents one of the most basic Fréchet mean and variance structures possible on metric space valued data. Recent work on Fréchet regression [59] and geodesic regression [27] provide examples of reasonable Fréchet mean functions of a Euclidean covariate for metric space valued data. In these works, the mean functions depend on more general covariates in $\mathbb{R}^k$, rather than just indicator functions of group membership. Another area of recent interest is modeling the joint distributions of random objects on metric spaces. The Bayesian hierarchical model of Section 5 provides a basic example of this, for if multiple observations were obtained within each group, then observations within the same group are more "correlated" with each other than observations in different groups. Various notions of covariance on metric spaces have been proposed in [22, 49, 68]. There is substantial scope for the development of parametric and nonparametric models that incorporate these notions of covariance and permit tractable inference. The geodesic James–Stein estimator solves the simple weighted Fréchet mean problem, $\delta_{JS,i} = \operatorname{argmin}_{z\in\mathcal{X}}(1 - w(X))d(X_i, z)^2 + w(X)d(\psi, z)^2$. It is anticipated that a typical inferential procedure for estimating the Fréchet means of correlated metric space data will result in solving similar weighted sample Fréchet mean problems.

Acknowledgments. We thank the Associate Editor and referees for their helpful and constructive comments.

SUPPLEMENTARY MATERIAL

Supplement A: Proofs (DOI: 10.1214/22-AOS2245SUPPA; .pdf). Proofs of the results in this article can be found in Supplement A [52].

Supplement B: Counterexamples, numerical results and algorithms (DOI: 10.1214/22-AOS2245SUPPB; .pdf). Supplement B [53] contains counterexamples related to the tower rule and unbiasedness of sample Fréchet means. Simulation results on a metric tree and in a regime where the Fréchet mean is sticky are provided.

REFERENCES

[1] Åkerborg, Ö., Sennblad, B., Arvestad, L. and Lagergren, J. (2009). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl. Acad. Sci. USA* **106** 5714–5719.

[2] Aleksandrov, A. D. (1951). A theorem on triangles in a metric space and some of its applications. *Tr. Mat. Inst. Steklova* **38** 5–23. MR0049584

[3] Alexander, S., Kapovitch, V. and Petrunin, A. (2019). *An Invitation to Alexandrov Geometry. CAT*(0) *Spaces. SpringerBriefs in Math.* **22**. Springer, Cham. MR3930625 https://doi.org/10.1007/978-3-030-05312-3

[4] Arsigny, V., Fillard, P., Pennec, X. and Ayache, N. (2006). 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

[5] Bačák, M. (2014). *Convex Analysis and Optimization in Hadamard Spaces. De Gruyter Series in Nonlinear Analysis and Applications* **22**. de Gruyter, Berlin. MR3241330 https://doi.org/10.1515/9783110361629

[6] Baranchik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Stat.* **41** 642–645. MR0253461 https://doi.org/10.1214/aoms/1177697104

[7] Beran, R. (2010). The unbearable transparency of Stein estimation. In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis*: *A Festschrift in Honor of Professor Jana Jurečková. Inst. Math. Stat.* (*IMS*) *Collect.* **7** 25–34. IMS, Beachwood, OH. MR2808363

[8] Berger, J. (1975). Minimax estimation of location vectors for a wide class of densities. *Ann. Statist.* **3** 1318–1328. MR0386080

[9] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR0804611 https://doi.org/10.1007/978-1-4757-4286-2

[10] 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

[11] 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
[12] BRANDWEIN, A. C. and STRAWDERMAN, W. E. (1991). Generalizations of James–Stein estimators under spherical symmetry. *Ann. Statist.* **19** 1639–1650. MR1126343 https://doi.org/10.1214/aos/1176348267
[13] BRANDWEIN, A. C. and STRAWDERMAN, W. E. (2012). Stein estimation for spherically symmetric distributions: Recent developments. *Statist. Sci.* **27** 11–23. MR2953492 https://doi.org/10.1214/10-STS323
[14] 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
[15] BROWN, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Stat.* **37** 1087–1136. MR0216647 https://doi.org/10.1214/aoms/1177699259
[16] BROWN, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Stat.* **42** 855–903. MR0286209 https://doi.org/10.1214/aoms/1177693318
[17] 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
[18] CARMICHAEL, O., CHEN, J., PAUL, D. and PENG, J. (2013). Diffusion tensor smoothing through weighted Karcher means. *Electron. J. Stat.* **7** 1913–1956. MR3084676 https://doi.org/10.1214/13-EJS825
[19] DIACONIS, P. and YLVISAKER, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281. MR0520238
[20] DO CARMO, M. P. (1992). *Riemannian Geometry*. *Mathematics*: *Theory & Applications*. Birkhäuser, Inc., Boston, MA. MR1138207 https://doi.org/10.1007/978-1-4757-2201-7
[21] DUBEY, P. and MÜLLER, H.-G. (2019). Fréchet analysis of variance for random objects. *Biometrika* **106** 803–821. MR4031200 https://doi.org/10.1093/biomet/asz052
[22] DUBEY, P. and MÜLLER, H.-G. (2020). Functional models for time-varying random objects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 275–327. MR4084166
[23] DUDLEY, R. M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge Univ. Press, Cambridge. MR1932358 https://doi.org/10.1017/CBO9780511755347
[24] EFRON, B. and MORRIS, C. (1972). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika* **59** 335–347. MR0334386 https://doi.org/10.1093/biomet/59.2.335
[25] EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597
[26] FELSENSTEIN, J. and FELENSTEIN, J. (2004). *Inferring Phylogenies* **2**. Sinauer Associates, Sunderland, MA.
[27] 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
[28] FOURDRINIER, D., STRAWDERMAN, W. E. and WELLS, M. T. (2003). Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix. *J. Multivariate Anal.* **85** 24–39. MR1978175 https://doi.org/10.1016/S0047-259X(02)00023-4
[29] FOURDRINIER, D., STRAWDERMAN, W. E. and WELLS, M. T. (2018). *Shrinkage Estimation*. *Springer Series in Statistics*. Springer, Cham. MR3887633 https://doi.org/10.1007/978-3-030-02185-6
[30] FRÉCHET, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. Henri Poincaré* **10** 215–310. MR0027464
[31] GARYFALLIDIS, E., BRETT, M., AMIRBEKIAN, B., ROKEM, A., VAN DER WALT, S., DESCOTEAUX, M., NIMMO-SMITH, I. and DIPY CONTRIBUTORS (2014). DIPY, a library for the analysis of diffusion MRI data. *Front. Neuroinform.* **8** 8. https://doi.org/10.3389/fninf.2014.00008
[32] GINESTET, C. E. (2012). Strong consistency of Fréchet sample mean sets for graph-valued random variables. Preprint. Available at arXiv:1204.3183.
[33] GUTMANN, S. (1982). Stein's paradox is impossible in problems with finite sample space. *Ann. Statist.* **10** 1017–1020. MR0663454
[34] GUTMANN, S. (1984). Decisions immune to Stein's effect. *Sankhyā Ser. A* **46** 186–194. MR0778869
[35] HAFF, L. R. (1991). The variational form of certain Bayes estimators. *Ann. Statist.* **19** 1163–1190. MR1126320 https://doi.org/10.1214/aos/1176348244
[36] 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
[37] HUDSON, H. M. (1978). A natural identity for exponential families with applications in multiparameter estimation. *Ann. Statist.* **6** 473–484. MR0467991
[38] HUELSENBECK, J. P. (1995). Performance of phylogenetic methods in simulation. *Syst. Biol.* **44** 17–48.
[39] JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

[40] JUKES, T. H. and CANTOR, C. R. (1969). Evolution of protein molecules. *Mammalian Prot. Metab.* **3** 21–132.

[41] KNEIP, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22** 835–866. MR1292543 https://doi.org/10.1214/aos/1176325498

[42] KUBOKAWA, T. (1998). The Stein phenomenon in simultaneous estimation: A review. In *Applied Statistical Science*, *III* 143–173. Nova Sci. Publ., Commack, NY. MR1673649

[43] KUBOKAWA, T. and SRIVASTAVA, M. S. (1999). Robust improvement in estimation of a covariance matrix in an elliptically contoured distribution. *Ann. Statist.* **27** 600–609. MR1714715 https://doi.org/10.1214/aos/1018031209

[44] LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. MR2985942 https://doi.org/10.1214/12-AOS989

[45] LEE, J. M. (2018). *Introduction to Riemannian Manifolds*. *Graduate Texts in Mathematics* **176**. Springer, Cham. MR3887684

[46] LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR1639875

[47] LI, K.-C. and HWANG, J. T. (1984). The data-smoothing aspect of Stein estimates. *Ann. Statist.* **12** 887–897. MR0751280 https://doi.org/10.1214/aos/1176346709

[48] LIU, X., LIU, L. and HU, J. (2017). James-Stein estimation problem for a multivariate normal random matrix and an improved estimator. *Linear Algebra Appl.* **532** 231–256. MR3688639 https://doi.org/10.1016/j.laa.2017.06.032

[49] LYONS, R. (2013). Distance covariance in metric spaces. *Ann. Probab.* **41** 3284–3305. MR3127883 https://doi.org/10.1214/12-AOP803

[50] MADDISON, W. P. (1997). Gene trees in species trees. *Syst. Biol.* **46** 523–536.

[51] MARCHAND, E. and STRAWDERMAN, W. E. (2004). Estimation in restricted parameter spaces: A review. In *A Festschrift for Herman Rubin*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **45** 21–44. IMS, Beachwood, OH. MR2126884 https://doi.org/10.1214/lnms/1196285377

[52] MCCORMACK, A. and HOFF, P. (2022). Supplement A to "The Stein effect for Fréchet means": Proofs. https://doi.org/10.1214/22-AOS2245SUPPA

[53] MCCORMACK, A. and HOFF, P. (2022). Supplement B to "The Stein effect for Fréchet means": Counterexamples, numerical results and algorithms. https://doi.org/10.1214/22-AOS2245SUPPB

[54] MILLER, E., OWEN, M. and PROVAN, J. S. (2015). Polyhedral computational geometry for averaging metric phylogenetic trees. *Adv. in Appl. Math.* **68** 51–91. MR3345895 https://doi.org/10.1016/j.aam.2015.04.002

[55] PATRANGENARU, V. and ELLINGSON, L. (2016). *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*. CRC Press, Boca Raton, FL. MR3444169

[56] PENNEC, X. (2006). Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *J. Math. Imaging Vision* **25** 127–154. MR2254442 https://doi.org/10.1007/s10851-006-6228-4

[57] PENNEC, X., FILLARD, P., AYACHE, N. and EPIDAURE, P. (2006). A Riemannian framework for tensor computing. *Int. J. Comput. Vis.* **66** 41–66. https://doi.org/10.1007/s11263-005-3222-z

[58] PETERSEN, A. and MÜLLER, H.-G. (2019). Wasserstein covariance for multiple random densities. *Biometrika* **106** 339–351. MR3949307 https://doi.org/10.1093/biomet/asz005

[59] 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

[60] RASMUSSEN, M. D. and KELLIS, M. (2011). A Bayesian approach for fast and accurate gene tree reconstruction. *Mol. Biol. Evol.* **28** 273–290. https://doi.org/10.1093/molbev/msq189

[61] SCHÖTZ, C. (2019). Convergence rates for the generalized Fréchet mean via the quadruple inequality. *Electron. J. Stat.* **13** 4280–4345. MR4023955 https://doi.org/10.1214/19-EJS1618

[62] SEMPLE, C. and STEEL, M. (2003). *Phylogenetics*. *Oxford Lecture Series in Mathematics and Its Applications* **24**. Oxford Univ. Press, Oxford. MR2060009

[63] SHAO, P. Y.-S. and STRAWDERMAN, W. E. (1994). Improving on the James–Stein positive-part estimator. *Ann. Statist.* **22** 1517–1538. MR1311987 https://doi.org/10.1214/aos/1176325640

[64] STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 197–206. Univ. California Press, Berkeley–Los Angeles, CA. MR0084922

[65] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

[66] STURM, K.-T. (2002). Nonlinear martingale theory for processes with values in metric spaces of nonpositive curvature. *Ann. Probab.* **30** 1195–1222. MR1920105 https://doi.org/10.1214/aop/1029867125

[67] 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

[68] SZÉKELY, G. J. and RIZZO, M. L. (2013). Energy statistics: A class of statistics based on distances. *J*. *Statist*. *Plann*. *Inference* **143** 1249–1272. MR3055745 https://doi.org/10.1016/j.jspi.2013.03.018

[69] SZÖLLŐSI, G. J., TANNIER, E., DAUBIN, V. and BOUSSAU, B. (2015). The inference of gene trees with species trees. *Syst*. *Biol*. **64** e42–e62. https://doi.org/10.1093/sysbio/syu048

[70] TABELOW, K., POLZEHL, J., SPOKOINY, V. and VOSS, H. U. (2008). Diffusion tensor imaging: Structural adaptive smoothing. *NeuroImage* **39** 1763–1773. https://doi.org/10.1016/j.neuroimage.2007.10.024

[71] TSUKUMA, H. and KUBOKAWA, T. (2007). Methods for improvement in estimation of a normal mean matrix. *J*. *Multivariate Anal*. **98** 1592–1610. MR2370109 https://doi.org/10.1016/j.jmva.2007.04.009

[72] XIE, X., KOU, S. C. and BROWN, L. D. (2012). SURE estimates for a heteroscedastic hierarchical model. *J*. *Amer*. *Statist*. *Assoc*. **107** 1465–1479. MR3036408 https://doi.org/10.1080/01621459.2012.728154

[73] YANG, C. and VEMURI, B. C. (2019). Shrinkage estimation on the manifold of symmetric positive-definite matrices with applications to neuroimaging. In *International Conference on Information Processing in Medical Imaging* 566–578. Springer, Berlin.

[74] YANG, C. and VEMURI, B. C. (2020). Shrinkage estimation of the Frechet mean in Lie groups. Preprint. Available at arXiv:2009.13020.

[75] ZIEZOLD, H. (1977). On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. In *Transactions of the Seventh Prague Conference on Information Theory*, *Statistical Decision Functions*, *Random Processes and of the Eighth European Meeting of Statisticians* (*Tech*. *Univ*. *Prague*, *Prague*, 1974), *Vol*. *A* 591–602. Reidel, Dordrecht. MR0501230