On the wald Space for Phylogenetic Trees

Stephan F. Huckemann$^{[0000-0001-5990-1741]}$,
Mahshid Mirhashemi$^{[0009-0002-0362-4460]}$ and Tom M. W. Nye$^{[0000-0002-8737-777X]}$

Abstract Most existing metrics between phylogenetic trees directly measure differences in topology and edge weights, and are unrelated to the models of evolution used to infer trees. We describe metrics which instead are based on distances between the probability models of discrete or continuous characters induced by trees. We describe how construction of information-based geodesics leads to the recent [3] wald space of phylogenetic trees. As a point set, it sits between the BHV space [4] and the edge-product space [5]. It has a natural embedding into the space of symmetric positive definite matrices, equipped with the information geometry. Thus, singularities such as overlapping leaves are infinitely far away, proper forests, however, comprising the "BHV-boundary at infinity", are part of the wald space, adding boundary correspondences to groves (corresponding to orthants in the BHV space). In fact, the wald space contracts to the completely disconnected forest. Further, it is a geodesic space, exhibiting the structure of a Whitney stratified space of type (A) where strata carry compatible Riemannian metrics. We explore some more geometric properties, but the full picture remains open. We conclude by identifying open problems we deem interesting [1].

Stephan F. Huckemann
Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences, Georg-August-Universität Göttingen, Germany, e-mail: huckeman@math.uni-goettingen.de

Mahshid Mirhashemi
Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences, Georg-August-Universität Göttingen, Germany, e-mail: mahshid.mirhashemi@uni-goettingen.de

Tom M. W. Nye
School of Mathematics, Statistics and Physics, Newcastle University, UK e-mail: tom.nye@ncl.ac.uk

1 Introduction

With the postulations of *evolutionary hypotheses* that were introduced to the scientific community, e.g. by [6, 7, 8, 9], which had enormous impact in the second half of the 19th century, the question arose how to measure evolution not only qualitatively but also quantitatively. For this, a fundamental building block is the concept of a *distance between taxa* (species). Given such mutual distances, one can incorporate unobserved ancestors, to arrive at a weighted phylogenetic tree, explaining descendance and the number of accumulated mutations. Efforts to measure distances between taxa has, among others, led to the development of morphometric methods e.g. by [10, 11]. With the discovery of the structure of DNA, among others by [12, 13], distances between taxa could now also be based on differing genetic material. Notably, for morphology, the issue of defining a distance seemed rather canonical: the Euclidean distance of configurations of landmarks placed on homologous loci of every taxum, modulo the group of similarity transformations, e.g. [14, 15, 16]. Choosing a distance between two genetic sequences, which are words in the alphabet of the four nucleic bases of DNA (A = adenine, C = cytosine, F = guanine and T = thymine), seemed less straightforward. A simple approach is the distance proposed by [17]: Add one for every site having different letters. Sites, however, are rather correlated, among others due to the three dimensional geometry for thus encoded RNA and proteins, which relate directly to biological function and thus to evolutionary fitness. For this reason a plethora of biologically more realistic genetic distances have been proposed, e.g. [18]. There are more difficulties. Although there are *starting* and *ending codons* (certain letter triplets) encoding the beginning and end of a gene, practical *alignment* of genes common to all taxa of concern, with, usually, thousands of bases is often faulty, e.g. [19]. Further, there may be many possible trees compatible with empirically found distances, or none at all (satisfying the *four-point-condition* from (2)). In fact, due to the high number of possible *tree topologies* (see Section 2), tree building methods, usually traversing different *neighboring* tree topologies, return one, or a few, but not necessarily all possible trees – and the returned trees may depend on the method used.

To make it even worse, building trees on different common genes usually yields different phylogenetic trees. Eventually this leads to issue of *averaging* in some sense over different phylogenetic trees to obtain an "expected" phylogeny. [20] has extended the concept of an expected value to a random variable X in a metric space (Q, d) simply by finding a minimizer of the squared expected distance, a *barycenter*, also called a *Fréchet mean* in his honor:

$$\mu \in \operatorname{argmin}_{q \in Q} \mathbb{E}[d(X, q)^2] \,.$$

Indeed, if (Q, d) is a Euclidean space and X square integrable, then the minimizer is uniquely given by $\mathbb{E}[X]$. In general it may be nonunique (e.g. for the uniform distribution on a sphere) or even void (e.g. for a standard normal distribution on a Euclidean space, punctured at its origin).

Fortunately, in Hadamard spaces (see [21]) means are unique, and there is a stochastic algorithm proposed by [21], converging in probability (even a.s. if the support of X is bounded) to the Fréchet mean. This algorithm requires frequent computation of *geodesics* (length minimizing curves).

There is a price to pay, however: For a sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$, their *Fréchet sample mean*

$$\hat{\mu}_n \in \operatorname{argmin}_{q \in Q} \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[d(X_j, q)^2]$$

may *stick* to the *Fréchet population mean* μ, as discovered by [22] and illustrated in Example 1.

Definition 1 (Sample stickiness) A sample mean μ_n sticks to a population mean μ if there is a random number N such that $\mu_n = \mu$ a.s. for all $n \geq N$.

This potential lack of an asymptotic fluctuation is a dead end for asymptotic statistics, a two-sample test, say. Various *flavors of stickiness* beyond sample stickiness have been explored in [23] including *testing for the degree of stickiness* as a workaround in case of sample stickiness.

In this contribution we address the fundamental issue of constructing a suitable metric space for phylogenetic trees. Most prominent to date is the BHV space (the acronym stands for the inventors: Billera, Holmes and Vogtmann [4]), it has been well-studied with an abundance of numerical methods available, among others to compute Fréchet means and principal modes of variation, see e.g. [24, 25]. While mathematically elegant, it falls short on reflecting "true" biology "near infinity", i.e. for taxa phylogenetically nearly independent (see Section 2). Allowing for phylogenetically independent taxa, which results in proper forests (disconnected trees), the EP (*edge product*) space has been developed by [26, 5], see Section 3. This is a superset of the BHV space and in order to arrive at a closed topological simplicial complex, different taxa separated by zero distance are included. However, the EP space is not a metric space. The *wald space* from [3, 1] extends the BHV space by independent taxa, i.e. by phylogenetic forests, borrowing the topology of EP space, but not allowing for overlapping taxa. It takes its geometry from an approximation of the information geometry of a two-state biological Markov model. Eventually this corresponds to a subset of the space of symmetric positive definite matrices equipped with a Cartan-Hadamard structure yielding a nonpositive curvature geometry e.g. [27, Chapter XII], aka Killing or affine invariant geometry.

Due to including proper forests, topologies obtained by *pruning and regrafting* (removing a subtree and attaching it somewhere else, see [28]) are *neighboring*. The wald space's geometry therefore conveniently facilitates exploring a wide set of topologies in order to find a tree most likely to yield a given set of genetic sequences. On the other hand, it appears that due to curvature effects the Fréchet mean is repulsed by proper forests. This is advantageous since evolutionary biologists believe all species evolved from a single common ancestor, and so the mean should be a tree.

The geometric structure of the wald space, biologically appealing, comes with new mathematical and numerical challenges. We propose some of these interesting open problems at the end of this paper.

As usual for a set A, its *cardinality* is $|A|$. For finite $|A|$, a *partition of* A is a set $\{A_1, \dots, A_k\}$, $1 \leq k \leq |A|$ such that $A_1 \cup \dots \cup A_n = A$ and $A_i \cap A_j = \emptyset$ for all $1 \leq i < j \leq k$ in case of $k > 1$.

2 The BHV Space

Let $N \in \mathbb{N}$ denote the number of taxa plus one root, such that a *rooted phylogenetic tree* can be modeled as a set of compatible splits of the *leaf set* $\mathcal{L} = \{1, \dots, N\}$.

Usually, in the literature dealing with BHV spaces, N denotes the number of taxa, not including the root, so that N from that literature corresponds to our $N - 1$ here.

Here, a *split* $s = \{A, B\}$ is a partition of $\mathcal{L}$ (i.e. $A \cup B = N, A \cap B = \emptyset$). Since every edge in a tree "splits" the leaf set, every edge corresponds to a split. Vice versa, if no vertices in the tree have degree 2, every split also corresponds to an edge. Two splits $s_i = \{A_i, B_i\}$, $i = 1, 2$, are *compatible* if one of the following sets is empty

$$A_1 \cap A_2, \quad A_1 \cap B_2, \quad B_1 \cap A_2, \quad B_1 \cap A_2 \,.$$

It its easy to verify that a set of splits corresponds to the edges of a tree if and only if the splits are pairwise compatible.

A split $\{A, B\}$ is a *pendant split* if $|A| = 1$ or $|B| = 1$, else it is an *interior split*. The BHV tree spaces is the set of all rooted phylogenetic trees where each tree contains at least all pendant splits, and it contains zero or a positive number of interior splits.

In effect, a BHV tree models equivalence classes of graph-theoretical trees modulo relabeling of interior vertices (see Definition 2), where all interior vertices have a degree of at least three and the leaves are exactly the vertices with degree one. Then interior splits correspond uniquely to interior edges and pendant splits to pendant edges.

Notably, there are N pendant edges and $M = 2^{N-1} - N - 1$ possible interior edges. Let us number them s_j, $1 \leq j \leq M$. Of these at most $N - 3$ can be present in a single BHV tree.

Every edge (split) s carries a length $\ell(s) > 0$ modeling evolutionary distance, so that trees including their splits' lengths are elements in

$$\mathbb{R}_+^N \times \mathrm{BHV}_N \hookrightarrow \mathbb{R}^N \times \mathbb{R}^M$$

with

$$\mathrm{BHV}_N := \bigsqcup_{\substack{J = \{i_1, \dots, i_k\} \subset \{1, \dots, M\} \\ \text{giving compatible splits,} \\ 1 \leq k \leq N-3}} \mathbb{R}_+^J \hookrightarrow \mathbb{R}^M \,, \tag{1}$$

modeling the interior edges, if present, or by a single point, if not present. Here "$\sqcup$" stands for the disjoint union and $\mathbb{R}_+^\emptyset$ is identified with $\{0\} \subset \mathbb{R}^M$. Then $BHV_3 = \{0\}$ and, usually, BHV spaces with at least 3 leaves and one root ($N \geq 4$) are considered.

The natural embedding in the Euclidean space equips BHV_N with the canonical trace metric, turning this disjoint union BHV_N into a CAT(0) space (see [4]). As the combinatorial structure is modeled by BHV_N, usually, the cartesian product with the space $\mathbb{R}_+^N$ modeling the pendant edges is ignored. Each $\mathbb{R}_+^J$ above is called an *orthant* and, by induction, the number of top-dimensional (of dimension $N-3$) orthants in BHV_N is

$$1 \cdot 3 \cdot \dots \cdot (2N-5) = \sqrt{2}\left(\frac{2N}{e}\right)^{N-2} (1 + o(1)) \text{ as } N \to \infty$$

Thus, while the geometry is rather simple, the challenge lies in the combinatorics. In spite of the number of orthants growing exponentially, [33] have proposed an algorithm that determines a geodesic between arbitrary trees in BHV_N within $O(N^4)$,

Another issue manifesting in BHV space is that of *stickiness*, which is already present in BHV_4.

Example 1 As there are three possible interior edges, BHV_4 is the tripod obtained from three positive coordinate axes joined at the origin. Then a straightforward computation yields that the Fréchet sample mean μ_n, sampling n points uniformly from the three points $(1,0,0), (0,1,0), (0,0,1)$, lies at the origin μ (the Fréchet population mean) unless one of the points appears more often than $2n/3$ in the sample. By the strong law of large numbers, the frequency of every point in a sample converges to $1/3$ almost surely, so that there is a.s. a random integer N such that for all $n > N$ the sample means μ_n coincides with (sticks to) the population tree.

In general, for suitable distributions, sample means may stick to any lower dimensional orthant (e.g. [34]), with the effect of stickiness strongest at the completely unresolved tree, the star tree having no interior edges, modeled by $\{0\}$. Due to varying topologies, many realistic data sets have their Fréchet means at the star tree, featuring stickiness (e.g. [29]).

More problematic for BHV spaces, rather than the issue of stickiness, seems biological modeling. Two trees with all edges very long, regardless of their topologies encode phylogenies that are close to one with independent taxa. If their topologies are not neighboring, as edge lengths increase, their BHV distance tends to infinity, while they are getting biologically more similar to the topology of a totally disconnected forest. Such forests and hence the "BHV boundary at infinity" will be modeled by the edge product space below.

3 The Edge Product Space

Evolution of genetic sequences can be viewed as a stochastic process switching between states as laid out in [35, 36, 18]: In full biological complexity, states correspond to words in the four letter alphabet of nucleotide bases. In a simple mathematical model, there are only two states, and the stochastic process is indexed along a true but unknown phylogenetic tree. Further, most simply, this process is assumed to be a stationary, time-reversible Markov process. Of the underlying probability distribution, only the marginal, namely correlation of leaves $x, y \in \mathcal{L}$ is observed, which is linked to their evolutionary distances

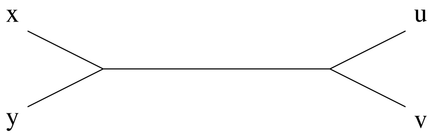
$$d(x,y) = \sum_{e \in \text{ path from } x \text{ to } y} \ell(e)$$

by

$$\rho(x,y) := e^{-d(x,y)} .$$

At this point recall that a symmetric matrix $D = (d(x,y))_{x,y\in\mathcal{L}}$ with nonnegative entries, vanishing along the diagonal, encodes the leaf-distances within a tree if and only if the *four point condition* (2) below is satisfied.

$$\begin{cases} \text{For all } x,y,u,v \in \mathcal{L}, \\ \text{below, two are equal and larger than the third} \\ d(x,y) + d(u,v), \quad d(x,u) + d(y,v), \quad d(x,v) + d(y,u) . \end{cases} \tag{2}$$



Notably, this condition implies the triangle inequality. In correlation notation it reads as

$$\begin{cases} \text{for all } x,y,u,v \in \mathcal{L}, \\ \text{below, two are equal and smaller than the third} \\ \rho(x,y)\rho(u,v), \quad \rho(x,u)\rho(y,v), \quad \rho(x,v)\rho(y,u) . \end{cases} \tag{3}$$

Thus, [5] showed that the *edge product space*

$$\begin{aligned} \mathrm{EP}_N := \{\rho = (\rho(x,y))_{x,y\in\mathcal{L}} \in [0,1]^{N\times N} :\ & \rho \text{ is symmetric}, \\ & \text{it satisfies (3) and } \rho(x,x) = 1 \text{ for all } x \in \mathcal{L}\} , \end{aligned}$$

uniquely models phylogenetic forests with edge lengths, modulo relabeling of internal vertices. Thus $\mathbb{R}_+^N \times \mathrm{BHV}_N$ can be naturally viewed as a subset of EP_N. It extends BHV_N by the following two aspects. For two leaves $x, y \in \mathcal{L}$, $x \neq y$,

(A1) $\rho(x,y) = 1$ is possible in EP_N meaning that $d(x,y) = 0$, i.e. x and y "overlap"
(A2) $\rho(x,y) = 0$ is possible in EP_N meaning that $d(x,y) = \infty$, i.e. x and y are in different subtrees

In particular the completely disconnected forest with $\rho(x,y) = \delta_{x,y}$, $x, y \in \mathcal{L}$ is an element of EP_N. Topologically EP_N can thus be viewed as a compactification of BHV_N. This compactification is not compatible with the BHV_N geometry, however.

Theorem 1 ([5]) EP_N*, equipped with the natural trace topology inherited from Euclidean* $\mathbb{R}^{N\times N}$*, is a finite CW complex, that is contractible.*

In the following, we bring the edge product space even closer to biology by excluding taxa that are different but yet identical (aspect (A1)) and equipping the corresponding subset with a biologically motivated geometry that will be fundamentally different from the BHV_N geometry.

4 The wald Space

Definition 2 We call a triple $F = (V, E, \ell)$ a *phylogenetic forest* over the leaf set $\mathcal{L}$ if

(F1) V is a finite set called *vertices* with $\mathcal{L} \subset V$,
(F2) $E \subset \{\{(u, v)\}, u, v \in V\}$, called the *edge set*, turns (V, E) into a graph-theoretical forest (disjoint union of trees),
(F3) $\ell \in (0, \infty)^E$, which encodes edge lengths,
(F4) $\deg(u) \in \{1, 2\}$ for $u \in V$ if and only if $u \in \mathcal{L}$, where the *degree* $\deg(u)$ of a vertex $u \in V$ is the cardinality of $\big\{\{u, v\} \in E : v \in V\big\}$, of incident edges.

Further, we say that two phylogenetic forests $F = (V, E, \ell)$ and $F' = (V', E', \ell')$ over a common leaf set $\mathcal{L}$ are *equivalent modulo relabeling of internal vertices* and write $F \sim F'$ if there is a bijection $f : V \longrightarrow V'$ satisfying

(E1) $f(x) = x$ for all $x \in \mathcal{L}$,
(E2) $(u, v) \in E \Leftrightarrow (f(u), f(v)) \in E'$,
(E3) $\ell(u, v) = \ell'(f(u), f(v))$ for all $\{u, v\} \in E$.

Throughout the rest of this paper,

$$\mathrm{SPD}(N) = \{A \in \mathbb{R}^{N\times N} : A = A^T > 0\}$$

denotes the set of symmetric positive definite matrices.

Theorem 2 ([3]) *Every phylogenetic forest* $F = (V, E, \ell)$ *over a leaf set* $\mathcal{L} = \{1, \ldots, N\}$ *has a representative* $\rho \in \mathrm{SPD}(N)$ *satisfying:*

(i) $0 \leq \rho(x, y) \leq 1 = \rho(x, x)$ *for all* $x, y \in L$*, and*
(ii)the 4-pt condition (3).

Vice versa, every $\rho \in \mathrm{SPD}(N)$ *satisfying (i) and (ii) is a representative of some graph-theoretical forest* F*, and if* F' *is another representative, then* $F \sim F'$*.*

Definition 3 ([3]) Every such equivalence class of phylogenetic forests over the leaf set $\mathcal{L} = \{1, \dots, N\}$ modulo relabeling of internal vertices is called a *wald* and the space of all such equivalence classes is the *wald space*, denoted by $\mathcal{W}_N$.

Recall that for a BHV tree, every edge e carried a length $\ell(e) \in (0, \infty)$. In order to metrically compactify at "infinity" we introduce edge weights in λ-*notation*:

$$\ell(e) \mapsto \lambda(e) := 1 - \exp(-\ell(e)) \in (0,1) \text{ such that } 0 \mapsto 0, \infty \mapsto 1\,.$$

Modeling forests instead of trees requires additional notation. Letting

$$\begin{aligned} \mathcal{P}_N &:= \{\{\mathcal{L}_1, \dots, \mathcal{L}_l\} \text{ is a partition of } \mathcal{L} \,:\, 1 \le l \le N\} \text{ and} \\ \mathcal{E}^{\mathcal{L}_j} &:= \{\{(A_i, B_i\} \,:\, 1 \le i \le k\} \text{ are compatible splits of } \mathcal{L}_j\} \end{aligned}$$

we identify

$$\mathcal{W}_N := \{W_\infty\} \sqcup \bigsqcup_{\substack{E = E_1 \cup \ldots \cup E_l \\ E_j \in \mathcal{E}^{\mathcal{L}_j},\, 1 \le j \le l \\ \{\mathcal{L}_1, \dots, \mathcal{L}_l\} \in \mathcal{P}_N \\ 1 \le l \le N-1}} (0,1)^E\,, \tag{4}$$

where W_∞ denotes the completely disconnected forest with partition $\{\{j\} : j = 1, \dots, N\}$ and empty edge (split) sets $\mathcal{E}^{\{j\}}$, $1 \le j \le N$. Every open cube $(0,1)^E$ is called a *grove* and in case of $l = 1$, a grove corresponds to the product of a BHV orthant times the pendant edges' orthant: $\mathbb{R}^N_+ \times \mathbb{R}^J_+$ with suitable J, see (1).

Definition 4 With (4) we denote the elements of $\mathcal{W}_N$ by (E, λ) with suitable $E = E_1 \cup \ldots \cup E_j$ and $\lambda \in (0,1)^E$. Further, let $\phi : \mathcal{W}_N \to \mathrm{SPD}(N)$ be the injective mapping guaranteed by Theorem 2, where we consider $\mathrm{SPD}(N) \hookrightarrow \mathbb{R}^{N \times N}$.

Theorem 3 ([1]) *With the above notation, the following hold:*

(i) $\phi(\mathcal{W}_N)$ *is star shaped in* $\mathbb{R}^{N\times N}$ *with respect to the unit matrix, which is* $\phi(W_\infty)$.

(ii) The pullback topology of $\mathcal{W}_N$ *(pulling back the Euclidean topology of* $\mathbb{R}^{N\times N}$ *under* ϕ*) agrees with the topology induced from the pullback metric under* ϕ *generated by the infimum of lengths of curves*

$$\phi(\mathcal{W}) \supset \gamma \textit{ continuous in } \mathbb{R}^{N\times N} \textit{connecting } p, q \in \phi(\mathcal{W})\,.$$

From now on, we assume that $\mathcal{W}_N$ is equipped with the pullback topology under ϕ. In the next step we derive a stratified structure of $\mathcal{W}$ comprising analytic manifold strata. Since the mapping ϕ restricted to a single grove has, by construction, the simple form

$$\begin{aligned} \phi_E : (0,1)^E &\to \mathrm{SPD}(N) \\ (\lambda_e)_{e\in E} &\mapsto \left(\prod_{e \in \text{ path from } x \text{ to } y} (1 - \lambda_e) \right)_{x,y=1}^N \end{aligned}$$

(empty products are set to 1), it is real analytic (it can be even analytically extended to all of $\mathbb{R}^E$). Hence the canonical analytic manifold structure of $\mathrm{SPD}(N) \hookrightarrow \mathbb{R}^{N\times N}$ can be pulled back under ϕ to every orthant making it an analytic manifold. As detailed in the following theorem, thus the entire wald space is a Whitney stratified space of type (A).

Theorem 4 ([1]) *The image of $\mathcal{W}_N$ under ϕ in $\mathbb{R}^{N\times N}$ carries a natural structure of a Whitney stratified space of type (A): it comprises disjoint analytic manifold strata M_j of every dimension $0 \le j \le 2N-3$ such that for all $0 \le i < j \le 2N-3$,*

(i) if $M_i \cap \overline{M_j} \neq \emptyset$ then $M_i \subset \overline{M_j}$,
(ii) if $M_j \ni q_1, q_2, \ldots \longrightarrow p \in M_i$ and $T_{q_1}, T_{q_2}, \ldots \longrightarrow T$ in the Grassmannian $G(\mathbb{R}^{N\times N}, j)$, then $T_p M_i \subset T$.

In particular, $M_0 = \{\phi(W_\infty)\}$.

In the final step we equip the wald space with a biologically motivated metric. To this end recall the Markov process from Section 3 associated with a phylogenetic tree T, and denote it by X_t^T. It assumes values in a state space Ω, which in our context of the two-state model is just $\{0,1\}$. This process is indexed in loci $t \in T$ where T is viewed as the point set of the phylogenetic tree and it is determined by the Bernoulli probabilities and variances

$$\left\{\begin{array}{l} \mathbb{P}\{X_{t_1}^T = X_{t_2}^T\} = \frac{1}{2}\Big(1 + \exp(-d(t_1,t_2))\Big) \\ \mathrm{Var}[X_{t_1}|X_{t_2}] = \frac{1}{4}\Big(1 - \exp(-2d(t_1,t_2))\Big) \end{array}\right\}, \quad t_1, t_2 \in T\,, \tag{5}$$

where d denotes the distance on T. Then we have the marginal $\mathbb{P}\{X_t = 1\} = \frac{1}{2}$ for all $t \in T$. In particular there is a one-to-one relationship between such processes X^T and all probability distributions p_T on Ω^N, i.e. on the leaf set [30, 31].

A natural distance for such probability distributions is an f-divergence, for instance, a Kullback-Leibler divergence, a Jenson-Shannon divergence or a Hellinger distance. It turns out that the corresponding information metrics (parameters are edge lengths) are equivalent for all f-divergences if f is convex and $f(1) = 0$ [3]. Computations within this geometry, e.g. of distances (which build on computations of geodesics), involve summation over Ω^N, which become computationally costly for high N. For this reason we approximate instead these probability distributions by a continuous Gaussian process $Z_t, t \in T$ that has equal first and second moments,

$$\big(Z_{t_2}|Z_{t_1} = z\big) \;\sim\; \mathcal{N}(ze^{-d(t_1,t_2)}, 1 - e^{-2d(t_1,t_2)})\,,$$

yielding $\mathrm{Var}[Z_x] = 1$, and $\mathrm{cov}[Z_x.Z_y] = \exp(-d(x,y))$ for all $x, y \in \mathcal{L}$. Similarly employing their information geometry for the space of covariances, i.e. the positive definite matrices, determining full rank Gaussians with zero mean, yields the well known universal nonpositive curvature geometry, aka Killing or affine invariant Riemannian, or Fisher information metric for $\mathrm{SPD}(N)$ (e.g. [27, 37]). This metric allows for straightforward computation of geodesics, given by

$$\gamma(t) = \sqrt{p}\,\mathrm{Exp}(t\mathrm{Log}(\sqrt{p}^{-1}q\sqrt{p}^{-1}))\sqrt{p} \tag{6}$$

from $p = \gamma(0)$ to $q = \gamma(1)$, involving the matrix exponential and logarithm, as well as the unique symmetric positive definite matrix roots. This geometry can be restricted to $\phi(\mathcal{W})$ and thus pulled back to $\mathcal{W}$.

For the following recall (e.g. from [38]) that with the notation of Theorem 4, a Whitney stratified space of type (A) is a *Riemann stratified space*, if every stratum M_j carries a Riemannian geometry with Riemannian tensor $g_p^{(j)} : T_pM_j \times T_pM_j \to \mathbb{R}$, $p \in M_j$, $j = 1, \dots, 2N-3$, such that, whenever $M_j \ni q_1, q_2, \dots \to p \in M_i$, $1 \leq i < j \leq 2N-3$, and $T_{q_1}, T_{q_2}, \dots \to T$ in the Grassmannian $G(\mathbb{R}^{N\times N}, j)$, then $g_{q_k}^{(j)}$ converges, as $k \to \infty$, to some $g_p^* : T \times T \to \mathbb{R}$, which, restricted to $T_pM_i \times T_pM_i$ agrees with $g_p^{(i)}$.

Theorem 5 ([1]) *The wald space $\mathcal{W}_N$ equipped with the pullback of the information geometry on* SPD(N) *is a Riemann stratified space, that is geodesic (i.e. any two curves can be joined by a geodesic, i.e. a length minimizing curve).*

Remark 1 The EP$_N$ trees/forests with overlapping leaves excluded in $\mathcal{W}_N$ would map under the natural extension of ϕ to the degenerate symmetric positive semi-definite matrices on the "boundary" of SPD(N), which, in the information geometry, is infinitely far away.

Remark 2 Geodesics in wald space are pre-images under ϕ of shortest curves in the geometry of SPD(N) that satisfy the four point condition (3) throughout. Numerical experiments suggest that SPD(N) geodesics given by (6) between two different matrices satisfying the four point conditions, only satisfy them at the boundary points. In fact, effectively numerically computing wald space geodesics is an active field of current research, cf. [2], [1].

5 Open questions

The investigation of the wald space has only recently begun, many questions have not been answered yet. Here we state some of them, some with conjectured answers.

1. Are grove closures geodesically convex? We conjecture: Yes.
   Notably, orthants in BHV space are already geodesically convex, i.e. BHV geodesics between two trees with common tree topology are given by linear interpolation of edge lengths and thus stay within the respective topology. Numerical experiments for wald geodesics between two trees within a common grove, however, may traverse the boundary of degenerate topologies, see Figure 1.
2. Is there an analog of the algorithm of [33], which computes the groves to traverse for geodesics between pairs of trees (with different) topologies?

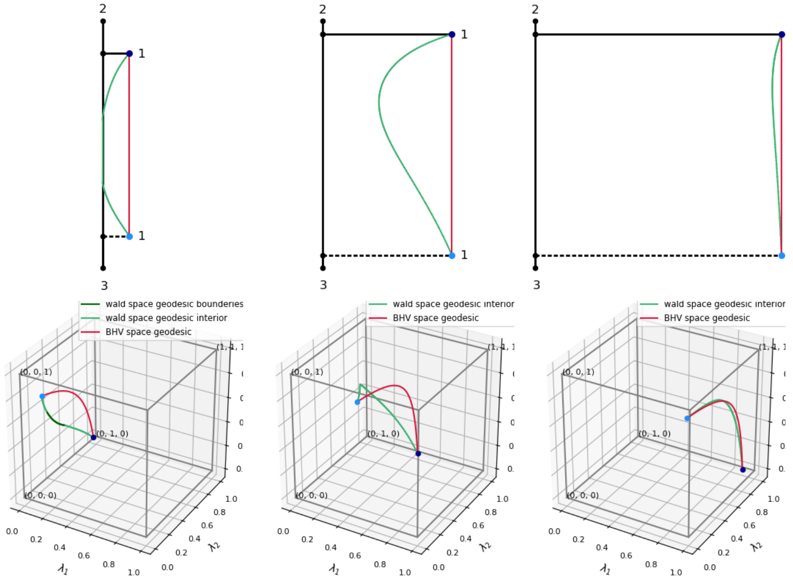


Fig. 1 *Geodesics in $\mathcal{W}_3$ (green) and $\mathbb{R}^3_+ \times \mathrm{BHV}_3$ (red). The initial tree (black) differs from the final tree (leaves 2 and 3 unchanged in black, leaf 1 (blue)). Along the geodesics in $\mathcal{W}_3$ (bottom) the products $\lambda_1 \cdot \lambda_2$ of the pendant weights is only almost constant; depicting it as constant (top) results in seemingly nonsymmetric geodesics (green, top). The curves traversed by the third pendant weight λ_3 is depicted in the top row. Notably, the first wald space geodesic (left column) sojourns (dark green) on the boundary stratum determined by $\lambda_2 = 0$. For initial trees further away from the boundary at zero, their joining wald space geodesics remains in the top-dimensional grove (middle and right column). Even for initial trees near the boundary at ∞, numerical experiments show that wald space geodesics are pushed away from the totally disconnected forest, suggesting that it is repulsive as opposed to the attractive boundary at zero.*

The BHV geodesic between two trees with split sets E_1 and E_2 only traverses orthants involving splits from $E_1 \cup E_2$, as [33] showed. In the light of Question (1), a similar result may be expected.

3. "Regularity" of forests: Can proper forests be on geodesics between trees? Conjecture: No.

   First numerical computations showed that the vantage point W_∞ is repulsive, see Figures 1 and 2.

4. Antipodes: Are there cut loci (pairs of wälder admitting different shortest geodesics)? Conjecture: Yes but exotic.

   Recall that $\mathcal{W}_N$ is viewed as a subspace of $\mathrm{SPD}(N)$ with the information metric, yielding a space of global nonpositive curvature (e.g. [27]), featuring no cut loci [21]. Numerical experiments, however, show that $\mathcal{W}_N$ also features positive curvatures (geodesic triangles featuring sums of Alexandrov angles greater than π), see Figure 2. Such spaces tend to exhibit antipodes, as, for instance, spheres.

Natural candidates for such antipodal points would be trees or forests exhibiting some symmetries.

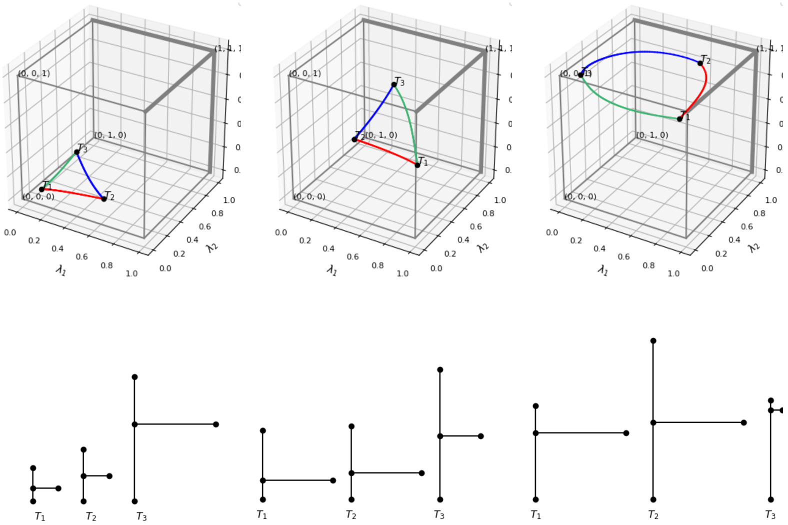


Fig. 2 *Top: geodesic triangles in $\mathcal{W}_3$ getting "fatter" as they move closer to the vantage point W_∞, the totally disconnected forest (corresponding to the fat gray edges and their intersection point $(1,1,1)$), the sum of their Alexandrov angles are $159.26^0, 193.32^0, 427.90^0$ from left to right. Bottom: the corresponding trees in with leaf labeling as in Figure 1.*

5. Which conditions provide uniqueness of geodesics and Fréchet means?
   If the answer to Question (4) is positive, most likely concentration conditions for the underlying probability distributions. For instance restricting the support to open "geodesic half balls" as in [39], ensures uniqueness on manifolds. Also [40] guarantee uniqueness for sample means of distributions featuring densities with respect to Riemannian measures on manifolds.
6. Limit theorems for the Fréchet mean:

(6.a) We expect stickiness (collapsing parametric asymptotics) and
(6.a) smeariness (exploding parametric asymptotics),
(6.a) but anticipate the validity of the bootstrap in such scenarios.

Samples $X_1, \dots, X_n$ of a random variable X in a Euclidean space with finite second moment satisfy the classical central limit theorem

$$\sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n} X_i - \mathbb{E}[X]\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \operatorname{cov}[X])\,,$$

as $n \longrightarrow \infty$. For circles and spheres, it has been shown that Fréchet sample means may exhibit slower convergence rates than the parametric rate of $n^{-1/2}$ [41, 42, 43] and this phenomenon has been called *smeariness*. As mentioned in the introduction, probability distributions in BHV space may feature stickiness, resulting in faster asymptotic rates. We expect that both types of asymptotics occur on the wald space. As smeariness is related to nonuniqueness of Fréchet means [32] this links to Question (5).

In case of smeariness or stickiness, quantile based statistics can no longer be used. For the circle it has been shown by [44] that the bootstrap still provides consistency.

7. Is $\mathcal{W}_N$ also a Whitney stratified space of type (B)? Conjecture: Yes at BHV boundaries, not at forest boundaries.

   In addition to a suitable convergence of tangent spaces, Whitney stratified spaces of type (B) also feature a suitable convergence of tangent vectors: In the context of the notation from Theorem 4 (recall that all strata are embedded in a Euclidean space) one requires additionally that whenever $M_j \ni q_1, q_2, \ldots \to p \in M_i$ and $M_i \ni p_1, p_2, \ldots \to p$, $1 \leq i < j \leq 2N - 3$, and the secant lines c_k between p_k and q_k $(k = 1, 2, \ldots)$ converge to a line c, then c is contained in T.

Acknowledgements

The first author gratefully acknowledges support by the Niedersachen Vorab of the Volkswagen Foundation, DFG HU 1575/7 and DFG SFB 1456. The second author acknowledges funding by DFG GK 2088.

All authors thank the anonymous referee for helpful comments, including the following consideration: The fact that wald space includes disconnected forests might be counterintuitive to an evolutionary biologist. Since disconnected forest are conjectured to be repulsive, they may not ordinarily appear as barycenters. But they may be contained in a confidence set comprising multiple topologies. This would carry the statistical interpretation that the data set at hand suffices to determine some species relations but not others, not that these relations do not exist

References

1. J. Lueg, M. K. Garba, T. M. W. Nye, S. F. Huckemann, ”Foundations of wald space for statistics of phylogenetic trees,” *Journal of the London Mathematical Society*, vol, 109, no. 5, e12893, 2024.
2. J. Lueg, M. K. Garba, T. M. W. Nye, S. F. Huckemann, ”Wald space for phylogenetic trees,” *Proceedings of Geometric Science of Information: 5th International Conference, GSI 2021, Paris, France, July 21–23*, pp. 710–717, 2021, Springer.
3. M. K. Garba, T. M. W. Nye, J. Lueg, S. F. Huckemann, ”Information geometry for phylogenetic trees,” *Journal of Mathematical Biology*, vol. 82, no. 3, pp. 1–39, 2021, Springer.

4. L. J. Billera, S. P. Holmes, K. Vogtmann, "Geometry of the Space of Phylogenetic Trees," *Advances in Applied Mathematics*, vol. 27, no. 4, pp. 733–767, 2001, Elsevier.
5. V. Moulton, M. Steel, "Peeling phylogenetic 'oranges'," *Advances in Applied Mathematics*, vol. 33, no. 4, pp. 710–727, 2004, Elsevier.
6. J.-B. Lamarck, *Philosophie Zoologique*, vol. 1, 1873, Paris: F. Savy.
7. A. R. Wallace, "XVIII.—On the law which has regulated the introduction of new species," *Annals and Magazine of Natural History*, vol. 16, no. 93, pp. 184–196, 1855, Taylor & Francis.
8. C. Darwin, *On the Origins of Species by Means of Natural Selection*, London: Murray, 1859.
9. G. J. Mendel, "Versuche über Pflanzenhybriden," *Verhandlungen des naturforschenden Vereines in Brünn für das Jahr 1865, Abhandlungen*, vol. 4, pp. 3–47, 1866.
10. E. Haeckel, *Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformierte Descendenz-Theorie. Band 1: Allgemeine Anatomie. Band 2: Allgemeine Entwicklungsgeschichte*, de Gruyter, 1866.
11. D. W. Thompson, *On Growth and Form*, Cambridge University Press, 1917.
12. R. E. Franklin and R. G. Gosling, "The structure of sodium thymonucleate fibres. I. The influence of water content," *Acta Crystallographica*, vol. 6, no. 8-9, pp. 673–677, 1953, International Union of Crystallography.
13. J. D. Watson and F. H. C. Crick, "Molecular structure of nucleic acids," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.
14. J. C. Gower, "Generalized Procrustes analysis," *Psychometrika*, vol. 40, pp. 33–51, 1975.
15. D. G. Kendall, "The diffusion of shape," *Advances in Applied Probability*, vol. 9, pp. 428–430, 1977.
16. F. L. Bookstein, *The Measurement of Biological Shape and Shape Change*, 2nd edition, Lecture Notes in Biomathematics, vol. 24, Springer-Verlag, New York, 1978.
17. R. W. Hamming, "Error detecting and error correcting codes," *The Bell System Technical Journal*, vol. 29, no. 2, pp. 147–160, 1950, Nokia Bell Labs.
18. Z. Yang, *Computational Molecular Evolution*, OUP Oxford, 2006.
19. K. M. Wong, M. A. Suchard, J. P. Huelsenbeck, "Alignment uncertainty and genomic analysis," *Science*, vol. 319, no. 5862, pp. 473–476, 2008, American Association for the Advancement of Science.
20. M. Fréchet, "Les éléments aléatoires de nature quelconque dans un espace distancié," *Annales de l'Institut Henri Poincaré*, vol. 10, no. 4, pp. 215–310, 1948.
21. K. T. Sturm, "Probability measures on metric spaces of nonpositive curvature," *Contemporary Mathematics*, vol. 338, pp. 357–390, 2003, Providence, RI: American Mathematical Society.
22. T. Hotz, S. Huckemann, H. Le, J. S. Marron, J. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, S. Skwerer, "Sticky Central Limit Theorems on Open Books," *Annals of Applied Probability*, vol. 23, no. 6, pp. 2238–2258, 2013.
23. L. Lammers, D. T. Van, S. F. Huckemann, "Sticky Flavors," *arXiv preprint arXiv:2311.08846*, 2023.
24. E. Miller, M. Owen, J. S. Provan, "Polyhedral computational geometry for averaging metric phylogenetic trees," *Advances in Applied Mathematics*, vol. 68, pp. 51–91, 2015, Elsevier.
25. T. M. W. Nye, X. Tang, G. Weyenberg, R. Yoshida, "Principal component analysis and the locus of the Fréchet mean in the space of phylogenetic trees," *Biometrika*, vol. 104, no. 4, pp. 901–922, 2017, Oxford University Press.
26. J. Kim, "Slicing hyperdimensional oranges: the geometry of phylogenetic estimation," *Molecular Phylogenetics and Evolution*, vol. 17, no. 1, pp. 58–75, 2000, Elsevier.
27. S. Lang, *Fundamentals of Differential Geometry*, Springer, 1999.
28. B. L. Allen, M. Steel, "Subtree transfer operations and their induced metrics on evolutionary trees," *Annals of Combinatorics*, vol. 5, pp. 1–15, 2001, Springer.
29. S. Skwerer, E. Bullitt, S. Huckemann, E. Miller, I. Oguz, M. Owen, V. Patrangenaru, S. Provan, J. S. Marron, "Tree-oriented analysis of brain artery structure," *In revision*, 2013.
30. J. S. Rogers, "On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences," *Systematic Biology*, vol. 46, no. 2, pp. 354–357, 1997, JSTOR.

31. E. S. Allman, C. Ané, J. A. Rhodes, ”Identifiability of a Markovian model of molecular evolution with gamma-distributed rates,” *Advances in Applied Probability*, vol. 40, no. 1, pp. 229–249, 2008, Cambridge University Press.
32. B. Eltzner, ”Geometrical smeariness–a new phenomenon of Fréchet means,” *Bernoulli*, vol. 28, no. 1, pp. 239–254, 2022, Bernoulli Society for Mathematical Statistics and Probability.
33. M. Owen, J. S. Provan, ”A fast algorithm for computing geodesic distances in tree space,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 8, no. 1, pp. 2–13, 2011, IEEE Computer Society Press.
34. D. Barden, H. Le, M. Owen, ”Central limit theorems for Fréchet means in the space of phylogenetic trees,” *Electronic Journal of Probability*, vol. 18, no. 25, pp. 1–25, 2013.
35. C. Semple and M. Steel, *Phylogenetics*, Oxford University Press on Demand, 2003.
36. J. Felsenstein, *Inferring Phylogenies*, Oxford University Press, 2004.
37. C. Lenglet, M. Rousson, R. Deriche, O. Faugeras, ”Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor MRI processing,” *Journal of Mathematical Imaging and Vision*, vol. 25, no. 3, pp. 423–444, 2006.
38. S. F. Huckemann and B. Eltzner, ”Statistical Methods Generalizing Principal Component Analysis to Non-Euclidean Spaces,” in *Handbook of Variational Methods for Nonlinear Geometric Data*, eds. P. Grohs, M. Holler, A. Weinmann, Springer, 2020, pp. 317–388.
39. B. Afsari, ”Riemannian L^p center of mass: existence, uniqueness, and convexity,” *Proceedings of the American Mathematical Society*, vol. 139, pp. 655–773, 2011.
40. M. Arnaudon and L. Miclo, ”Means in complete manifolds: uniqueness and approximation,” *ESAIM: Probability and Statistics*, vol. 18, pp. 185–206, 2014, EDP Sciences.
41. T. Hotz and S. Huckemann, ”Intrinsic Means on the Circle: Uniqueness, Locus and Asymptotics,” *Annals of the Institute of Statistical Mathematics*, vol. 67, no. 1, pp. 177–193, 2015.
42. B. Eltzner and S. F. Huckemann, ”A smeary central limit theorem for manifolds with application to high-dimensional spheres,” *Annals of Statistics*, vol. 47, no. 6, pp. 3360–3381, 2019. DOI: 10.1214/18-AOS1781.
43. S. Hundrieser, B. Eltzner, S. F. Huckemann, ”A Lower Bound for Estimating Fréchet Means,” 2024. arXiv:2402.12290.
44. S. Hundrieser, B. Eltzner, S. F. Huckemann, ”Finite Sample Smeariness of Fréchet Means and Application to Climate,” *Electronic Journal of Statistics*, vol. 18, no. 2, pp. 3274–3309, 2024. DOI: 10.1214/24-EJS2276.