




## Information geometry for phylogenetic trees

M. K. Garba<sup>1,2</sup> · T. M. W. Nye<sup>1</sup> · J. Lueg<sup>3</sup> · S. F. Huckemann<sup>3</sup> 

Received: 28 March 2020 / Revised: 12 October 2020 / Accepted: 21 October 2020 /  
Published online: 15 February 2021  
© The Author(s) 2021

### Abstract

We propose a new space of phylogenetic trees which we call *wald space*. The motivation is to develop a space suitable for statistical analysis of phylogenies, but with a geometry based on more biologically principled assumptions than existing spaces: in wald space, trees are close if they induce similar distributions on genetic sequence data. As a point set, wald space contains the previously developed Billera–Holmes–Vogtmann (BHV) tree space; it also contains disconnected forests, like the edge-product (EP) space but without certain singularities of the EP space. We investigate two related geometries on wald space. The first is the geometry of the Fisher information metric of character distributions induced by the two-state symmetric Markov substitution process on each tree. Infinitesimally, the metric is proportional to the Kullback–Leibler divergence, or equivalently, as we show, to any  $f$ -divergence. The second geometry is obtained analogously but using a related continuous-valued Gaussian process on each tree, and it can be viewed as the trace metric of the affine-invariant metric for covariance matrices. We derive a gradient descent algorithm to project from the ambient space of covariance matrices to wald space. For both geometries we derive computational methods to compute geodesics in polynomial time and show numerically that the two information geometries (discrete and continuous) are very similar. In particular, geodesics are approximated extrinsically. Comparison with the BHV geometry shows that our canonical and biologically motivated space is substantially different.

**Keywords** Phylogenetic tree · Information geometry · Tree space

**Mathematics Subject Classification** 92D15 · 53A35 · 94A17

---

✉ S. F. Huckemann  
huckeman@math.uni-goettingen.de

Extended author information available on the last page of the article

## 1 Introduction

Evolutionary relationships between species are represented by phylogenetic trees, in which the leaves represent present-day species, internal vertices represent speciation events, and edge lengths represent the degree of evolutionary divergence between species (Semple and Steel 2003). Evolutionary relationships are often subject to a high degree of uncertainty, and so it is natural to consider the space of all possible relationships and probability distributions on this space. Billera et al. (2001) were the first to construct a space of all phylogenetic trees on a fixed set of leaves. This space, known as Billera–Holmes–Vogtmann or BHV tree space, has a very rich geometry: in particular there is a unique geodesic, or shortest length path, between any two points in the space. BHV tree space is a so-called  $CAT(0)$  space (Billera et al. 2001), meaning it has globally non-positive curvature, and many of its attractive geometric properties follow from this condition. A polynomial time algorithm for computing geodesics and their lengths was subsequently developed (Owen and Provan 2011). A number of statistical methods for analysing samples of phylogenetic trees have been established, which rely fundamentally on the geometry of BHV tree space by transferring conventional multivariate statistical methods into the new geometrical context. Algorithms have been developed for computing sample means (Bačák 2014; Miller et al. 2015), for constructing confidence regions for the population mean (Willis 2019), and for performing principal component analysis (Nye 2011; Feragen et al. 2013; Nye 2014; Nye et al. 2017). An alternative geometry for phylogenetic trees, known as the tropical tree space (Speyer and Sturmfels 2004; Lin et al. 2018), arises from regarding phylogenetic trees as distance matrices between the species at the leaves. Statistical methods such as calculation of sample means (Lin and Yoshida 2018) and principal component analysis (Yoshida et al. 2019), have also been developed in tropical tree space.

In the BHV and tropical tree spaces, trees are regarded primarily as geometric or algebraic objects, without specific consideration to how phylogenetic trees are estimated or interpreted. Phylogenetic trees are typically inferred from genetic sequence data via Markov models of sequence evolution over the edges of the tree (Yang 2006), and we are only concerned with such trees. Each phylogenetic tree can therefore be regarded as a probability model for genetic sequence data, and a space of all tree-like probability models can be constructed. This idea was first considered by Kim (2000), and then developed more formally in subsequent papers (Moulton and Steel 2004; Gill et al. 2008). The space is known as the *phylogenetic orange space* or *edge-product space*. While the space has been studied from the viewpoint of algebraic geometry (Zwiernik and Smith 2012; Engström et al. 2013), metric geometry on the space has received little attention. Recently, methods for approximately computing ‘probabilistic’ metrics on the edge-product space have been developed (Garba et al. 2018). These metrics are defined by mapping each tree to its associated distribution on sequence data, and using a metric between these probability distributions. Specifically, each tree represents a distribution on characters, where a character is a map from the  $N$  leaves of the tree to some alphabet of letters  $\Omega$ . The Hellinger and Jensen–Shannon metrics are defined between distributions on  $\Omega^N$  and are pulled back to give metrics between trees. Exact calculation of these metrics involves summation over all possible

characters, and so when  $N$  is large, Garba et al. (2018) use a simulation procedure to estimate the distance between any pair of trees. The probabilistic metrics have substantially different properties than the BHV and tropical metrics. For example, if all the edge lengths in a pair of given trees are scaled up linearly, then the BHV and tropical distance between the trees both scale in the same way, while in contrast the probabilistic metrics eventually tend to zero. This is because the letters at the leaves of the trees become independent from one another as the edge lengths increase, due to genetic saturation. The distributions on characters represented by the two trees therefore converge to one another as the edge lengths are scaled up, and the distance tends to zero (see Fig. 2 in Garba et al. (2018)).

The metrics studied by Garba et al. (2018) arise from embedding tree space into the larger ‘ambient’ space of all distributions on characters. They are obtained from the lengths of ‘chordal paths’ in the ambient space which do not generally lie within the embedded tree space, and are hence called *extrinsic* metrics. In contrast, the BHV metric is an *intrinsic* metric, since it is obtained from the lengths of paths lying within tree space. For trees sharing a common branching pattern the BHV metric agrees with the corresponding *extrinsic* metric obtained via an embedding into Euclidean space. The statistical methods developed on BHV tree space rely heavily on the intrinsic nature of the metric, and this motivated us to seek intrinsic analogs of the probabilistic metrics.

The aim of this article is to realize intrinsic metrics and their associated geodesics in a new space of forests, the *wald space*,<sup>1</sup> that is related to the edge-product tree space (for the subtle, yet essential differences see the discussion in Sect. 6), when the underlying assumptions are similar to those for the probabilistic metrics. We assume that the infinitesimal distance between two trees is measured using the Fisher information matrix. We show that this is equivalent to assuming the infinitesimal squared distance is the Kullback–Leibler divergence, or equivalently, any  $f$ -divergence. Our approach uses ideas from information geometry, which is the study of Riemannian differential geometry on spaces of probability distributions. The purpose of developing this geometry on this space of forests is with the ultimate aim of obtaining statistical methods analogous to those on other tree spaces. The probabilistic metrics and the information geometry have an important advantage over the BHV and tropical geometries: they have by definition a direct biological interpretation in terms of the evolution of genetic sequences. In the information geometry, two trees are close when they determine similar distributions of characters, and as a result they would be potentially indistinguishable if inferred from experimental samples of sequence data. Conversely, trees are distant in the information geometry when they induce substantially different distributions. In contrast, the BHV and tropical metrics are defined more abstractly without reference to evolutionary models or processes. Examples of the biological interpretation of the probabilistic metrics were given by Garba et al. (2018).

Our approach has two main parts. First, we consider geodesics in the information geometry when the model associated with each phylogenetic tree is the two-state symmetric Markov process. This is the simplest discrete Markov model of sequence

---

<sup>1</sup> This space was first discussed by the authors at the Oberwolfach 1804 meeting “Statistics for Data with Geometric Structure” in the *Schwarzwald* (Black Forest) in 2018.

evolution, for which there are two letters in the alphabet,  $\Omega = \{0, 1\}$ . This model is introduced in Sect. 2 along with a formal definition of the wald space and a brief review of BHV space. The thesis of Garba (2019) contains some comparisons of results obtained using the two-state model versus models with the DNA alphabet. Geodesics in wald space are constructed locally by numerically integrating a certain differential equation determined by the assumptions on the Riemannian metric. We explore geodesics on the space of unrooted trees with 5 leaves, for which visualization is relatively straightforward, and compare the results with those for BHV tree space. This forms Sect. 3 of the paper. Secondly, in order to improve computational tractability, we consider an alternative continuous-valued model of evolution on each tree. This consists of a Gaussian process which approximates the two-state Markov process by matching its moments. The continuous random variables at the leaves of the tree have a multivariate normal distribution with zero mean, for which the covariance matrix is related to the matrix of path lengths between the leaves. Numerically solving the differential equations for geodesics is much faster under this set of assumptions, and the geodesics closely resemble those for the two-state model. However, solutions are still restricted to trees sharing a common branching pattern, or topology. The definition of the Gaussian process on trees and numerical solution of geodesics in the corresponding information geometry are described in Sect. 4. The information geometry of multivariate normal distributions with zero mean corresponds to a certain geometry on the space of symmetric positive definite matrices, known as the Fisher–Rao or affine-invariant geometry, and the map from the wald space to covariance matrices is an isometric embedding in this space. The geometry on the space of symmetric positive definite matrices is analytically tractable, and geodesics can be computed in polynomial time. The embedding therefore gives intrinsic and extrinsic metrics on the wald space. We describe a projection algorithm from the space of symmetric positive definite matrices into the embedded wald space. We then use this algorithm to project geodesics in the ambient space down into wald space in various ways to obtain approximate geodesics between trees with different topologies. The embedding in the space of symmetric positive definite matrices and the associated geometry is described in Sect. 5. We conclude in Sect. 6 with a detailed discussion of the promises and challenges of our new wald space.

## 2 Background and the new wald space

### 2.1 Phylogenetic trees

For  $N = 2, 3, \dots$  we define  $U_N$  to be the set of unrooted phylogenetic trees on  $N$  taxa. More specifically, a tree  $T$  is an element of  $U_N$  if it satisfies the following conditions. First,  $T$  contains exactly  $N$  vertices with degree 1, which are called *leaves*, and these are bijectively labelled  $1, \dots, N$ . Secondly,  $T$  must contain no vertices with degree 2. Thirdly, each edge  $e$  in  $T$  is assigned a *length*  $\ell^e \geq 0$  with  $\ell^e \neq 0$  if  $e$  contains a leaf. An edge in a tree is called a *pendant edge* if it contains a leaf; otherwise it is called an *internal edge*. Similarly, the vertices which are not leaves are called *internal vertices*.

The edge lengths  $\ell^e$  on any given tree  $T \in U_N$  define a path length distance between any pair of leaves. The path length on  $T$  between  $u, v \in \{1, \dots, N\}$  will be denoted  $\ell_{uv}$ .

Each tree  $T \in U_N$  contains at most  $2N - 3$  edges, in which case the tree is called *fully resolved* or *bifurcating*, and all internal vertices have degree 3. Trees with fewer edges are called *unresolved*, and for  $N > 3$ , these contain at least one vertex with degree 4 or more. Trees which contain only the  $N$  pendant edges joined at a single degree- $N$  internal vertex are called *star trees*.

A tree  $T$  is *rooted* when some internal point  $\rho \in T$  is labelled as being the root. This is conveniently achieved by adding an additional taxon labelled 0 to the tree via a pendant edge of length zero. It follows that the set of rooted phylogenetic trees satisfies the same conditions as  $U_N$ , except the leaves are bijectively labelled  $0, 1, \dots, N$ , and the pendant edge containing taxon 0 has zero length. We will work with unrooted trees, but our results are easily transferred to the space of rooted trees via this relationship.

Every fully resolved tree will correspond to a fully resolved *BHV-tree* (reviewed in Sect. 2.2) and to a fully resolved *wald*, as introduced below in Sect. 2.4. In both BHV tree space and in wald space, unresolved trees will be identified with other trees with certain internal edges having zero length, so that conceptually a missing edge is the same as a zero length edge.

## 2.2 Billera–Holmes–Vogtmann tree space

Billera et al. (2001) defined a space of phylogenetic trees, subsequently known as BHV tree space, and described its geometry. BHV tree space can be described via an embedding in  $\mathbb{R}^d$  for dimension  $d$  which increases exponentially with the number of leaves. However, we have chosen to describe BHV tree space in a way different from the original authors, and we define it as a quotient space. As a result, the *wald space* introduced in the next section is a superset of BHV tree space when the spaces are regarded simply as sets, clarifying the relationship between the two spaces. Importantly, we allow internal edges on trees to have length zero, and under the quotient these are equivalent to trees with those edges missing. A second difference is that while Billera et al. (2001) worked with rooted trees, we work with unrooted trees. As described in Sect. 2.3, the distribution on binary characters determined by a tree does not depend on the root position under the two-state symmetric model, and so unrooted trees are more natural to work with.

BHV tree space is defined using the notion of splits, where a *split* is a bipartition of the leaf labels  $1, \dots, N$  into two disjoint sets. Cutting an edge of a tree induces such a bipartition of the leaves, and so each edge on a tree corresponds to a split, and the terms *split* and *edge* can be used interchangeably. The set of splits represented by a tree is called its *topology*.

Arbitrary sets of splits do not typically determine valid tree topologies: the splits of a tree must satisfy a compatibility condition. For example, the splits  $\{1, 2\}$ ,  $\{3, 4, \dots, N\}$  and  $\{1, 3\}$ ,  $\{2, 4, \dots, N\}$  are incompatible, since leaf 1 cannot be grouped next to both 2 and 3 on the same tree. For any topology  $\tau$  with  $k$  internal edges,  $0 \leq k \leq N - 3$ , the set of trees in  $U_N$  with that topology is bijectively parametrized by  $\mathbb{R}_{>0}^N \times \mathcal{O}_\tau$  where

the first term in the product parametrizes the pendant edge lengths that, by definition, are strictly positive, and  $\mathcal{O}_\tau = \mathbb{R}_{\geq 0}^k$  parametrizes the internal edge lengths.

The set  $\mathcal{O}_\tau$  is called the *orthant* associated with topology  $\tau$ , and we identify the set of all trees with topology  $\tau$  with  $\mathbb{R}_{> 0}^N \times \mathcal{O}_\tau$ . Under this identification, the set of all trees  $U_N$ , as defined in Sect. 2.1, is the disjoint union

$$U_N = \mathbb{R}_{> 0}^N \times \bigsqcup_{\tau} \mathcal{O}_\tau$$

where the disjoint union is taken over all possible topologies  $\tau$ .

The unrooted BHV tree space  $\mathcal{U}_N$  is obtained by taking the quotient of  $U_N$  with respect to an equivalence relation:

$$\mathcal{U}_N = U_N / \sim .$$

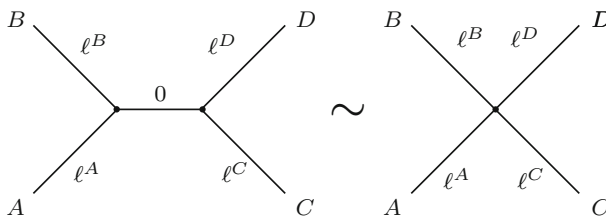
Two trees in  $U_N$  are equivalent under  $\sim$  if and only if they are identical modulo the presence of internal splits with zero length, as shown in Fig. 1.

The quotient space factorizes as

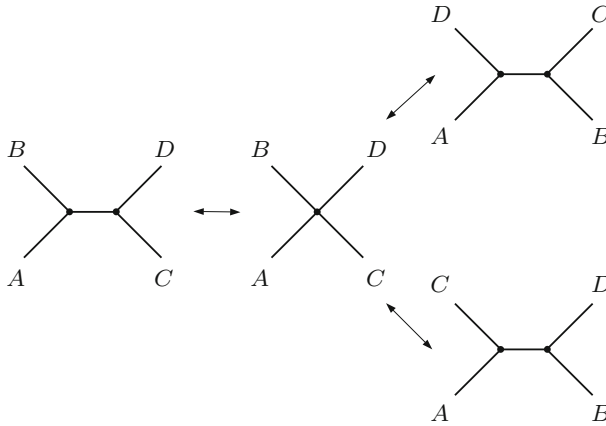
$$\mathcal{U}_N = \mathbb{R}_{> 0}^N \times \text{BHV}_N$$

where the first term parametrizes the lengths of the pendant edges and the space  $\text{BHV}_N$  parametrizes the topology and internal edge lengths of the BHV-trees. When  $\tau$  is fully resolved,  $\mathcal{O}_\tau$  is called a *maximal orthant*. Unresolved trees correspond to points on the boundaries of maximal orthants; they can be obtained from fully resolved trees by shrinking internal edge lengths down to zero.

Since there are  $(2N - 5)!!$  fully resolved unrooted topologies,  $\text{BHV}_N$  can be thought of as being constructed by gluing this number of maximal orthants together along their boundaries, where two points are identified if they correspond to the same tree. For example, when  $N = 4$ , there are three fully resolved topologies, each of which contains a single internal edge. The space  $\text{BHV}_4$  therefore consists of three copies of  $\mathbb{R}_{\geq 0}$  glued together at the origin. The origin corresponds to the star trees, while the location along



**Fig. 1** Two trees in  $U_N$  are equivalent under the relation  $\sim$  when they are identical after internal edges with length zero are removed, and the vertices at the end of every such edge are merged.  $A, B, C, D$  represent different subtrees joined by edges of length  $\ell^A, \ell^B, \ell^C, \ell^D$  to an internal edge with length  $\ell = 0$  on the left. The Markov process  $X(t)$  cannot change state on any edge with length zero, so the distribution on  $X_1, \dots, X_N$  is unchanged by removing such edges in this way



**Fig. 2** When an internal edge from a fully resolved topology is contracted down to length zero (left to centre), there are two fully resolved topologies which can be obtained by expanding out an alternative edge (right).  $A, B, C, D$  represent subtrees. The operation of contracting an internal edge and expanding out an alternative edge is called *nearest neighbour interchange*. It follows that at each codimension-1 boundary, three maximal orthants are glued together

each of the three copies of  $\mathbb{R}_{\geq 0}$  gives the length of the internal edge in each of the three possible fully resolved topologies. For  $N = 5$  there are 15 possible unrooted tree topologies, each of which contains two internal edges. It follows that  $BHV_5$  consists of 15 copies of  $\mathbb{R}_{\geq 0}^2$  glued along their boundaries. At each codimension-1 boundary, three maximal orthants are joined together. This is because when a single internal edge is contracted to length zero, a degree 4 vertex is obtained, and there are 3 possible ways to add in an edge, including the original edge, in order to obtain a fully resolved topology, as illustrated by Fig. 2.

The metric on  $BHV_N$  is constructed as follows. The basic idea is that for trees with the same fully-resolved topology but different vectors of internal edge lengths, say  $\ell_1$  and  $\ell_2$ , the distance is the Euclidean distance  $\|\ell_1 - \ell_2\|$ , and the corresponding geodesic is the straight line segment in the orthant containing the trees. Billera et al. (2001) showed that there exists a unique shortest path between any two points in  $BHV_N$ , for which path length is measured using the Euclidean distance in each orthant, and the length of these defines a metric on  $BHV_N$  which we denote  $d_{BHV}$ . A metric on  $\mathcal{U}_N$ , denoted  $d_{\mathcal{U}_N}$ , is obtained as the product metric when the metric on pendant edges is taken to be the Euclidean distance. An algorithm has been developed which constructs geodesics and calculates their lengths in  $O(N^4)$  time (Owen and Provan 2011).

### 2.3 The two-state symmetric Markov model

Genetic sequence evolution is typically modelled using discrete-valued continuous-time Markov processes defined over the edges of a tree  $T$  (Yang 2006; Bryant et al. 2005). DNA sequence evolution is modelled by associating to each point  $t \in T$ , a random variable  $X(t)$  which takes values in an alphabet  $\{A, C, G, T\}$ . In this paper, however, we will consider the two-state symmetric Markov process with alphabet

$\Omega = \{0, 1\}$ . This simplification is made in order to make the mathematics more tractable and for computational speed. Nonetheless, some of the calculations using the two-state symmetric can readily be performed using DNA models. More details are given in the thesis of Garba (2019) in which simulations show similarity of geometries obtained from the two- and the four-state process. The transition probability of the symmetric two-state model is defined in terms of the path length  $\ell_{t_1 t_2}$  between any two points  $t_1, t_2 \in T$ :

$$\begin{aligned} \Pr(X(t_2) = X(t_1)) &= \frac{1}{2} \left( 1 + e^{-\ell_{t_1 t_2}} \right), \quad \text{and} \\ \Pr(X(t_2) \neq X(t_1)) &= \frac{1}{2} \left( 1 - e^{-\ell_{t_1 t_2}} \right). \end{aligned} \tag{2.1}$$

The stationary distribution of this Markov process is  $Bern(1/2)$ , and the process is assumed to be in its stationary state over the tree. As a result, for all  $t \in T$ ,  $X(t)$  has a marginal Bernoulli distribution,  $X(t) \sim Bern(1/2)$ . While the random variables  $X_1, \dots, X_N$  at the leaves of the tree have the same marginal distributions, they are not independent since the tree imposes a dependence structure. The following lemma determines certain moments of the process  $X(t)$  giving insight on the dependence structure of  $X_1, \dots, X_N$ . The proof is straightforward using the transition probabilities in Eq. (2.1).

- Lemma 2.1** 1. If  $X_1, \dots, X_N$  are the random variables at the leaves of a tree  $T \in U_N$  determined by the discrete Markov process defined above, then  $\text{Cov}(X_u, X_v) = \frac{1}{4} \exp(-\ell_{uv})$  where  $\ell_{uv}$  is the path length between leaves  $u$  and  $v$ .
2. If  $t_1, t_2 \in T$  are path length  $\ell_{t_1 t_2}$  apart, then the conditional distribution of  $X(t_2)$  given  $X(t_1) = \omega \in \{0, 1\}$  has variance  $\frac{1}{4}(1 - \exp(-2\ell_{t_1 t_2}))$ .

It is straightforward to simulate realizations of  $X(t)$  in the following way. First simulate a Poisson process with rate 1 independently on each edge of the tree. The positions of the simulated events correspond to points at which  $X(t)$  changes parity. Secondly, pick any point  $t_0 \in T$  which is not a change point and sample  $X(t_0)$  from  $Bern(1/2)$ . The change points generated from the Poisson process then determine the value of  $X(t)$  for all other  $t \in T$ . The distribution obtained is independent of the choice of  $t_0$ , because the Markov process is reversible. In particular, the Markov process is independent of the choice of  $t_0$ , which could be considered as a root.

Under the model, each edge length can be interpreted as the expected number of change points that occur over the edge. Internal edges are allowed to have length zero, which means that no change in  $X(t)$  occurs over the edge. On the other hand, when edges are long, the number of changes is likely to be large, and the letters at either end of the edge are weakly correlated. Biologists refer to this effect as *saturation*. A fixed change of edge length  $\delta\ell$  therefore has more effect on the distribution of characters when applied to a short edge as opposed to a long edge in some given tree. For example, an increase of  $\delta\ell = 0.1$  to an edge with length  $\ell = 0.1$  approximately doubles the probability that the letters at either end of the edge are different, but the same change to an edge of length  $\ell = 10$  has almost no effect on this probability, which due to saturation is very close to  $1/2$ . This idea becomes important when we consider



defining distances between trees via the information they represent, in particular using the probability mass function of the nontrivial distribution of  $(X_1, \dots, X_N)$ .

**Remark 2.1** 1. The probability mass function of  $(X_1, \dots, X_N)$  determined by  $T$  is denoted  $p_T(s)$  where  $s \in \{0, 1\}^N$  is called a *binary character*. Given any binary character  $s$ , the values of  $p_T(s)$  can be evaluated via a recursive algorithm (Bryant et al. 2005), described in Appendix A. Appendix A also contains a modified form of the algorithm which is used to compute exactly the derivatives of  $p_T(s)$  with respect to the edge lengths.

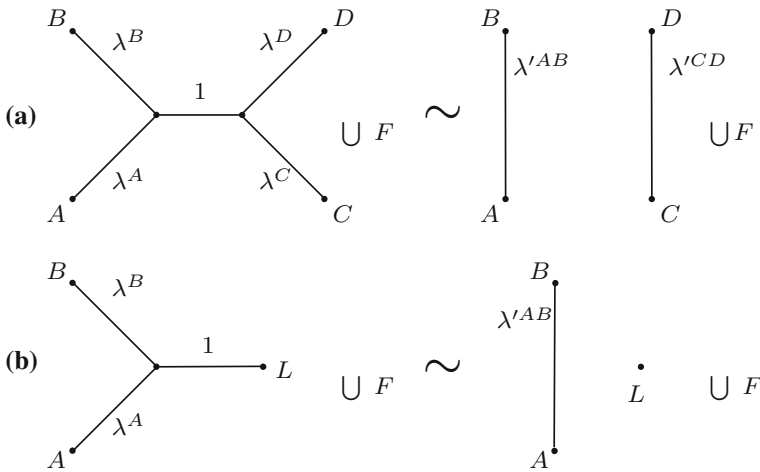
2. Also, as shown in Appendix A,  $p_T(s)$  is a multivariate polynomial in  $1 + e^{\ell^k}$  and  $1 - e^{\ell^k}$  where  $\ell^k$  ranges over the edge lengths in  $T$ .
3. Crucially, the map  $T \mapsto p_T$  from fully resolved trees in  $\mathcal{U}_N$  to the space of probability mass functions on  $\{0, 1\}^N$  is injective up to the equivalence relation introduced in Sect. 2.2 (Rogers 1997; Allman et al. 2008). This has two implications: first that the probability mass function  $p_T$  uniquely characterizes each element of  $\mathcal{U}_N$ , and secondly that metrics on distributions of characters pull back to define a metric between fully resolved trees, as described in Sect. 3.2.

## 2.4 A new forest space: the wald space

The following wald space gives an alternative viewpoint of phylogenetic trees by regarding them as Markov models for sequence evolution (Kim 2000; Moulton and Steel 2004; Gill et al. 2008). We will give a description of the wald space that is related to the *edge-product* space of previous authors, by defining it as a quotient space which adds trees with infinitely long edges to BHV tree space. As described in the introduction, each probabilistic metric considered by Garba et al. (2018) converges to zero in the limit as all the edge lengths in a pair of trees simultaneously tend to infinity. This behaviour indicates that shortest paths might cross a tree with infinitely long edges, which is why we add this *boundary at infinity*. Furthermore, we allow pendant edges of length zero under certain conditions. Thus, in the wald space, edge lengths  $\ell^e$  take values in  $\mathbb{R}_{\geq 0} \cup \{\infty\}$ . It is convenient to reparametrize to the  $\lambda$ -*parametrization* by defining *weights*  $\lambda^e = 1 - \exp(-\ell^e)$ , so  $\lambda^e \in [0, 1]$ . Under this transformation,  $\text{BHV}_N$  becomes a set of unit cubes, rather than orthants, glued along faces where  $\lambda^e = 0$  for one or more edges. The wald space is defined by imposing additional gluing rules on faces where  $\lambda^e = 1$ .

In order to be able to identify trees with infinitely long edges along faces where  $\lambda^e = 1$ , we construct the wald space from forests, that is, disjoint unions of unrooted trees. We start with a preliminary space leading to the definition of the wald space further below. Let  $W_N$  be the collection of forests satisfying the following necessary and sufficient conditions for each  $F \in W_N$ .

1. The forest  $F$  contains exactly  $N$  labelled vertices, these are called leaves and labelled  $1, \dots, N$ .
2. There are no unlabelled vertices in  $F$  of degree 0, 1 and 2.
3. For every pair of leaves  $u, v$  in the same tree in  $F$ , at least one edge  $e$  on the unique path from  $u$  to  $v$  satisfies  $\lambda^e > 0$ .



**Fig. 3** Illustration of *boundary at infinity* rule used to define  $\mathcal{W}_N$ . In both (a) and (b), the forests on the left are equivalent to the forests on the right.  $A, B, C, D$  are subtrees. **a** Internal edge with weight 1: the edge is deleted, disconnecting the tree. The resulting edges between subtrees  $A$  and  $B$  are replaced by a single edge with weight  $\lambda^{AB} = \lambda^A + \lambda^B - \lambda^A \lambda^B$  and similarly for  $C, D$ . **b** Pendant edge, where  $L$  is a leaf. The pendant edge with weight 1 is removed, and the resulting edges between  $A$  and  $B$  are replaced by a single edge with weight  $\lambda^{AB}$ . The term  $F$  in both panels refers to other disconnected components in the forests

Clearly,  $U_N \subset W_N$ . The condition on the edge weights ensures that no pair of leaves is coincident and consequently that metrics are always well-defined, as described in Sect. 3.2. We impose an equivalence relation  $\sim$  on  $W_N$ , defined by the following two rules.

**BHV boundary rule:** Given  $F_1, F_2 \in W_N$ , suppose all internal edges with  $\lambda^e = 0$  are removed from the forests, and the vertices at either end of each such edge are merged. If the resulting forests are identical, then  $F_1 \sim F_2$ . The rule is the same as that in Fig. 1.

**Boundary at infinity rule:** Suppose  $F \in W_N$  contains an edge with  $\lambda^e = 1$ , and that  $F$  is modified as follows. The edge with  $\lambda^e = 1$  is removed, disconnecting the tree it belongs to. If this results in any unlabelled vertex having degree 2, then those vertices are removed. If  $v$  is such a vertex, and the two adjacent edges  $e, \tilde{e}$  have weights  $\lambda^e, \lambda^{\tilde{e}} \in [0, 1]$ , then  $e, \tilde{e}$  are replaced by a single edge with weight  $\lambda^e + \lambda^{\tilde{e}} - \lambda^e \lambda^{\tilde{e}}$ , as is further explained below. Now suppose  $F_1, F_2 \in W_N$ , and this process of modifying unit-weight edges is applied to both forests. Then  $F_1 \sim F_2$  if the resultant forests are identical, as illustrated in Fig. 3.

The *wald space*  $\mathcal{W}_N$  is defined to be the quotient  $W_N / \sim$  and it immediately follows that as sets  $\mathcal{U}_N \subset \mathcal{W}_N$ , but the geometry imposed on  $\mathcal{W}_N$  will be completely different from the geometry of the BHV space.

The boundary rule at infinity requires some explanation. The rule declares that edges of weight  $\lambda^e = 1$  (or equivalently length  $\ell^e = \infty$ ) can be deleted from a forest  $F$ , but unlike the BHV rule for which the vertices at the ends of the edge are merged, edge removal disconnects a tree in  $F$ . When resulting degree-2 vertices are

removed, the edge length is preserved so that the new edge has length  $\ell^e + \ell^{\tilde{e}}$ . The corresponding weight  $\lambda$  is given by  $\lambda = 1 - \exp(-(\ell^e + \ell^{\tilde{e}})) = \lambda^e + \lambda^{\tilde{e}} - \lambda^e \lambda^{\tilde{e}}$ . Unlike the BHV boundary rule, in which finitely many trees are identified in each equivalence class, infinitely many combinations of edge weights  $\lambda^e, \lambda^{\tilde{e}}$  give rise to the same value  $\lambda^e + \lambda^{\tilde{e}} - \lambda^e \lambda^{\tilde{e}}$ . It follows that an uncountable collection of forests can be identified into a single equivalence class in  $\mathcal{W}_N$ .

In the edge-product space (Moulton and Steel 2004; Gill et al. 2008), an alternative parametrization is used, defining  $\mu^e = 1 - \lambda^e = \exp(-\ell^e)$  to be the weight of edge  $e$ . This parametrization has the advantage that sums of edge lengths  $\ell^1 + \dots + \ell^m$  become products of edge weights  $\mu^1 \times \dots \times \mu^m$  (hence the name ‘edge-product’). The boundary at infinity rule is simpler under this parametrization: the weights in Fig. 3 panel (b) become  $\mu^e, \mu^{\tilde{e}}$  and 0 on the left and  $\mu^e \mu^{\tilde{e}}$  on the right. However, under the  $\mu$ -parametrization, the BHV boundary with  $\ell^e = 0$  lies on faces of cubes with  $\mu^e = 1$ , whereas the boundary at infinity has  $\mu^e = 0$ . We prefer to work with the  $\lambda$ -parametrization since it gives a more intuitive interpretation of the weights, i.e.  $\ell^e = 0$  corresponds to  $\lambda^e = 0$  and  $\ell^e = \infty$  corresponds to  $\lambda^e = 1$ . Forests which contain more than one connected component lie in the faces of cubes with at least one  $\lambda^e = 1$ . Since the pendant edges can be expanded out to infinite length, they are also subject to the boundary at infinity rule, and so the representation of pendant edge lengths in  $\mathcal{W}_N$  is not via a product geometry, as it is for BHV tree space. While the star trees correspond to all internal edges having zero length,  $\mathcal{W}_N$  also contains a point which consists of  $N$  isolated vertices.

The BHV boundaries enable tree topologies to be changed via nearest neighbour interchange (NNI) operations (as illustrated by Fig. 2). The boundary at infinity corresponds to a different topological operation, called tree bisection and reconnection (TBR) (Allen and Steel 2001). Under this operation, an edge  $e$  in a tree can be expanded up to the boundary  $\lambda^e = 1$ . Removing the edge bisects the tree, and the two components can be reconnected by an edge  $\tilde{e}$  with  $\lambda^{\tilde{e}} = 1$  placed arbitrarily between the two trees. Reducing the weight  $\lambda^{\tilde{e}}$  down from 1 then gives a tree with a topology different from the original tree. It follows that there exist continuous paths in the wald space between trees with different topologies, which pass through the boundary at infinity and, as a result, change tree topology via TBR operations. This is in contrast to BHV tree space in which paths between trees of different topologies involve only NNI operations, as edges are contracted down to length zero and alternative edges are expanded out.

While the set  $\mathcal{W}_N$  was defined above via an equivalence relation on forests, we also need to understand how it parametrizes Markov models and then characterise its elements again as probability mass functions on  $\{0, 1\}^N$ . The two-state symmetric Markov process extends from being defined on trees to forests by taking the process on each connected component in a forest to be independent of the other components. This defines a distribution  $p_F$  on  $\{0, 1\}^N$  for each  $F \in \mathcal{W}_N$ . In fact the distribution uniquely determines the equivalence class of  $F$ , and vice versa, as the following lemma shows.

**Lemma 2.2** *Given  $F_1, F_2 \in \mathcal{W}_N$ , then  $F_1 \sim F_2$  if and only if  $p_{F_1}(s) = p_{F_2}(s)$  for all  $s$ .*

A proof is given in the Appendix.

Note that the forest consisting of  $N$  isolated vertices corresponds to the random variables  $X_1, \dots, X_N$  being independent, and this can be obtained from any tree by expanding all edges one after the other, as, by definition, there is at least one edge between any two leaves.

### 3 Information geometry for the two-state symmetric model

Information geometry provides methods for constructing metrics and geodesics on parametrized sets of probability distributions. In this section we embed wald space in the space of distributions of two-state characters, and investigate the corresponding information geometry analytically and computationally.

#### 3.1 Geometry of embeddings

Suppose that  $\theta : X \rightarrow Y$  where  $(Y, d)$  is a metric space and  $\theta$  is injective. We will say that  $X$  is embedded in  $Y$ , and refer to  $Y$  as the *ambient space*. The embedding can be used to construct certain metrics on  $X$ . First, since  $\theta$  is injective,  $d$  pulls back to define a metric on  $X$  which we denote  $d_X$ :

$$d_X(x_1, x_2) = d(\theta(x_1), \theta(x_2))$$

for all  $x_1, x_2 \in X$ . The pullback metric is often called the induced *extrinsic metric* and, it is simply the restriction of  $d$  to  $X \subseteq Y$ , and so when the context is clear, it is also denoted  $d$ . The probabilistic metrics described in Sect. 3.2 are constructed in this way. A second metric, called the induced *intrinsic metric* and denoted  $d^*(x_1, x_2)$ , is defined as the infimum of the length of all possible paths in  $X$  between  $x_1, x_2 \in X \subseteq Y$  when path length is measured using the metric  $d$ . If no path with finite length exists between  $x_1$  and  $x_2$  then  $d^*(x_1, x_2) = \infty$ , in which case  $d^*$  is not a metric. Details of this construction of the induced intrinsic metric are given by Bridson and Haefliger (2011). The metrics  $d$  and  $d^*$ , if well-defined, give  $X$  the structure of a *length space*, which is a space in which the metric between points  $x_1, x_2$  is the infimum of the lengths of paths between those points. Length spaces are similar to geodesic metric spaces, except that the infimum is not necessarily achieved by a path lying within the space; in a geodesic metric space a minimum length path exists between every pair of points, and so every geodesic metric space is a length space. An example of a length space which is not a geodesic metric space is  $\mathbb{R}^2$  with the origin removed and the Euclidean metric. Points antipodal to the origin cannot be joined by a geodesic, but the distance between them is the infimum of the lengths of paths joining the points.

In order to illustrate the relationship between  $d$  and  $d^*$ , consider the example of the embedding of the unit sphere  $X = S^2$  in  $Y = \mathbb{R}^3$  equipped with the Euclidean metric  $d$ . For any two points  $x_1, x_2$ ,  $d(x_1, x_2)$  is the length of the straight line segment in the ambient space  $\mathbb{R}^3$  joining the points. This metric is usually called the *chordal metric* on  $S^2$ . However, when we consider paths between  $x_1, x_2$  which are restricted to lie in  $S^2$ , the shortest paths (with respect to  $d$ ) are great circles, and the induced metric  $d^*$

is the arc length metric. In fact,  $S^2$  is a geodesic metric space, since the infimum of path length is always achieved by a great circle.

In the following Sect. 3.2, the wald space  $\mathcal{W}_N$  will be embedded in the space of distributions of characters. Later in Sect. 4 it will be embedded in the space of  $N \times N$  symmetric positive definite matrices. Each embedding will be used to construct metrics on  $\mathcal{W}_N$ .

### 3.2 Probabilistic metrics

Here we briefly describe the probabilistic metrics developed by Garba et al. (2018) since these will be used for comparison with other metrics. The Kullback–Leibler divergence is a commonly used measure of the difference between two distributions. Given two probability mass functions  $p, q$  on characters  $\{0, 1\}^N$ , the Kullback–Leibler divergence from  $q$  to  $p$  is defined as

$$D_{KL}(p; q) = \sum_{s \in \{0,1\}^N} p(s) \log \left( \frac{p(s)}{q(s)} \right)$$

provided  $p(s) = 0$  only when  $q(s) = 0$ . The Kullback–Leibler divergence is not a metric since it is not symmetric. However, metrics can be defined as follows: the Jensen–Shannon metric  $d_{JS}$  is defined by

$$d_{JS}(p, q)^2 = \frac{1}{2} D_{KL} \left( p; \frac{p+q}{2} \right) + \frac{1}{2} D_{KL} \left( q; \frac{p+q}{2} \right)$$

and the Hellinger metric  $d_H$  is defined by

$$d_H(p, q)^2 = \sum_{s \in \{0,1\}^N} \left( \sqrt{p(s)} - \sqrt{q(s)} \right)^2.$$

Recently, probabilistic metrics have been developed which are based on distributions of gene trees instead of distributions of characters (Adams and Castoe 2020).

The Kullback–Leibler divergence, squared Jensen–Shannon metric and squared Hellinger metric are all examples of a more general class of distances between probability distributions known as  $f$ -divergences. Given any convex function  $f(t)$  such that  $f(1) = 0$ , the  $f$ -divergence of  $p$  from  $q$  is defined as

$$D_f(p; q) = \sum_{s \in \{0,1\}^N} q(s) f \left( \frac{p(s)}{q(s)} \right). \tag{3.1}$$

The Kullback–Leibler divergence  $D_{KL}(p; q)$  is obtained by taking  $f(t) = t \log t$ , while the reversed divergence  $D_{KL}(q; p)$  is obtained with  $f(t) = -\log(t)$ . The squared Jensen–Shannon metric and squared Hellinger metric can also be obtained by using more complicated functions  $f$ , cf. Sason and Verdu (2016).

Now, let  $F \in W_N$  be a forest representative of a wald  $[F] \in \mathcal{W}_N$ . As described in Sect. 2.4, the distributions at leaves of different trees of  $F$  are independent. For two leaves in the same tree in  $F$ , some degree of evolution occurs between them since by definition of  $W_N$  no two leaves are coincident. Therefore, all characters are possible, giving

$$p_F(s) \neq 0 \text{ for all } s \in \{0, 1\}^N. \tag{3.2}$$

It follows that the Kullback–Leibler divergence is always well-defined between distributions of characters corresponding to forest representatives from the wald space. Since by Lemma 2.2 the map  $[F] \mapsto p_F$  is injective for  $[F] \in \mathcal{W}_N$  the Jensen–Shannon and Hellinger metrics pull back to define extrinsic metrics on the wald space  $\mathcal{W}_N$  (analogously to (Garba et al. 2018)). As already mentioned in the introduction, statistical methods rely heavily on the intrinsic nature of metrics, and thus we aim for more geometrical structure in the next section by imposing the Fisher information metric (a Riemannian metric) onto the wald space.

### 3.3 A two-state process geometry for the wald space

BHV tree space  $\mathcal{U}_N$  and wald space  $\mathcal{W}_N$  both do not have the structure of a manifold globally, but the interior of each maximal orthant is a manifold parametrized by  $\ell$  or  $\lambda$ . Therefore we consider first the information geometry on the subspaces of wald space corresponding to a fixed fully resolved tree topology—here every wald has only one single tree representative, since every wald corresponding to a forest with more than one component, as well as a wald containing a pendant edge with length zero, lies on the boundary of unit cubes corresponding to fully resolved tree topologies. Secondly, we establish global results about the constructed geometry of  $\mathcal{W}_N$ .

Thus suppose  $\tau$  is a fully resolved tree topology, and that trees with this topology are parametrized by  $\ell = (\ell^1, \dots, \ell^{2N-3}) \in \mathbb{R}_{>0}^N \times \mathcal{O}_\tau$ .

Let  $p_\ell(s)$  be the probability mass function  $p_T(s)$  associated with tree  $T$  determined by  $\tau, \ell$ . Recalling that  $p_\ell(s) > 0$  for all  $s$ , due to (3.2), the Fisher information matrix at  $\ell$  is

$$g_{ij}(\ell) = \sum_{s \in \{0,1\}^N} p_\ell(s) \left( \partial_i \log p_\ell(s) \right) \left( \partial_j \log p_\ell(s) \right) \tag{3.3}$$

for  $i, j = 1, \dots, 2N - 3$  where  $\partial_i = \partial/\partial \ell^i$ . This defines a Riemannian inner product on the tangent space of  $\mathbb{R}_{>0}^N \times \mathcal{O}_\tau$  at  $\ell$  (that is a copy of  $\mathbb{R}^{2N-3}$ ) which gives a way to measure the lengths of paths. Specifically, if  $p_\ell$  and  $p_{\ell+\delta\ell}$  lie infinitesimally close on a path, then the squared path length between them is defined to be  $\sum_{i,j} \delta \ell^i g_{ij}(\ell) \delta \ell^j$ . Standard results from Riemannian differential geometry show that if  $\ell(t)$  is a path in  $\mathbb{R}_{>0}^N \times \mathcal{O}_\tau$  then it is locally a geodesic (i.e. it minimizes path length) if it satisfies the differential equation

$$\frac{d^2 \ell^k}{dt^2} + \sum_{i,j} \Gamma_{ij}^k(\ell) \frac{d\ell^i}{dt} \frac{d\ell^j}{dt} = 0, \quad k = 1, \dots, 2N - 3 \tag{3.4}$$

where  $\Gamma_{ij}^k(\boldsymbol{\ell})$  are the Christoffel symbols

$$\Gamma_{ij}^k(\boldsymbol{\ell}) = \sum_l \frac{1}{2} g^{kl} \left( \frac{\partial g_{li}}{\partial \ell^j} + \frac{\partial g_{lj}}{\partial \ell^i} - \frac{\partial g_{ij}}{\partial \ell^l} \right).$$

The matrix  $g^{ij}$  is the inverse of  $g_{ij}$  i.e.  $\sum_k g^{ik} g_{kj} = \delta_j^i$ , where  $\delta_j^i$  is the Kronecker delta. It is important to note that the geodesic equation and loci of solutions are invariant under changes of parametrization, and so the equations can be formulated using lengths  $\ell^i$  or the weights  $\lambda^i$ . On the boundary, however, this is no longer necessarily true.

The Riemannian metric defined by the Fisher information matrix is related to the Kullback–Leibler divergence and other  $f$ -divergences by the following lemma.

**Lemma 3.1** *Suppose  $D_f$  is an  $f$ -divergence given by some convex function  $f$  with  $f(1) = 0$ , as defined by (3.1). Consider a small perturbation  $\delta\boldsymbol{\ell} = (\delta\ell^1, \dots, \delta\ell^{2N-3})$  of the edge lengths of a tree  $(\tau, \boldsymbol{\ell})$ . Then*

$$\sum_{i,j} \delta\ell^i g_{ij}(\boldsymbol{\ell}) \delta\ell^j = \frac{2}{f''(1)} D_f(p_{\boldsymbol{\ell}+\delta\boldsymbol{\ell}}; p_{\boldsymbol{\ell}}) + O(|\delta\boldsymbol{\ell}|^3) \tag{3.5}$$

where the error term consists of third-order products of the elements of  $\delta\boldsymbol{\ell}$  and

$$D_f(p_{\boldsymbol{\ell}+\delta\boldsymbol{\ell}}; p_{\boldsymbol{\ell}}) = \frac{1}{2} f''(1) \sum_s \frac{(p_{\boldsymbol{\ell}+\delta\boldsymbol{\ell}}(s) - p_{\boldsymbol{\ell}}(s))^2}{p_{\boldsymbol{\ell}}(s)} + O(|\delta\boldsymbol{\ell}|^3). \tag{3.6}$$

In other words, the norm of the perturbation, as measured with respect to the Riemannian inner product, is proportional to the  $f$ -divergence of  $p_{\boldsymbol{\ell}+\delta\boldsymbol{\ell}}$  from  $p_{\boldsymbol{\ell}}$ .

The proof is given in the Appendix. Since the lemma applies to an arbitrary  $f$ -divergence, the term on the right-hand side of Eq. (3.5) can be the Kullback–Leibler divergence or the squared Jensen–Shannon metric, for example. Lemma 3.1 gives the fundamental assumption behind the geometries we construct on  $\mathcal{W}_N$ : that distances are locally measured by the infinitesimal Kullback–Leibler divergence between probability distributions associated with trees, or equivalently, by any  $f$ -divergence. As a corollary of Lemma 3.1, it follows that the metric defined by Eq. (3.3) is positive definite (i.e. the metric is not semi-Riemannian). This is because the map from trees to distributions of characters is injective, and since  $D_f(p; q) > 0$  for all  $p \neq q$ , it follows that the right-hand side of Eq. (3.5) is strictly positive for all small non-zero perturbations  $\delta\boldsymbol{\ell}$ .

Let  $\mathcal{D}(\{0, 1\}^N)$  be the space of distributions on  $\{0, 1\}^N$ . By Lemma 2.2, the map from  $\mathcal{W}_N$  to distributions of characters determines an embedding of  $\mathcal{W}_N$  in  $\mathcal{D}(\{0, 1\}^N)$ . Given a metric  $d$  on  $\mathcal{D}(\{0, 1\}^N)$ , let  $d^*$  be the induced intrinsic metric on  $\mathcal{W}_N$ , as described in Sect. 3.1.

**Theorem 3.1** *Let  $d$  and  $d_0$  be metrics on  $\mathcal{D}(\{0, 1\}^N)$  which are the square root of an  $f$ - and  $f_0$ -divergence, respectively. Then for any  $[F], [G] \in \mathcal{W}_N$ :*

1.  $d^*([F], [G]) < \infty$  and thus  $d^*$  is well-defined.
2.  $d^*([F], [G]) = c \cdot d_0^*([F], [G])$  for some constant  $c > 0$ .
3. Any path which realizes the distance  $d^*([F], [G])$  is a solution of Eq. (3.4) at any point in the interior of a maximal orthant.

The proof is given in the Appendix. Valid choices for the metric  $d$  in Theorem 3.1 include the Jensen–Shannon metric or Hellinger metric. Theorem 3.1 establishes  $\mathcal{W}_N$  with metric  $d^*$  as a length space. Finiteness of  $d^*$  shows, for example, that points in  $\mathcal{W}_N$  corresponding to disconnected forests (or equivalently, trees with infinite edge lengths) are at a finite distance away from orthant interiors. The second assertion implies a scaling of the induced intrinsic metric under changes of the function  $f$ , which, in turn, substantiates our conclusions drawn from Lemma 3.1 that the geometry of  $\mathcal{W}_N$  induced by  $d$  is invariant under the choice of  $f$ .

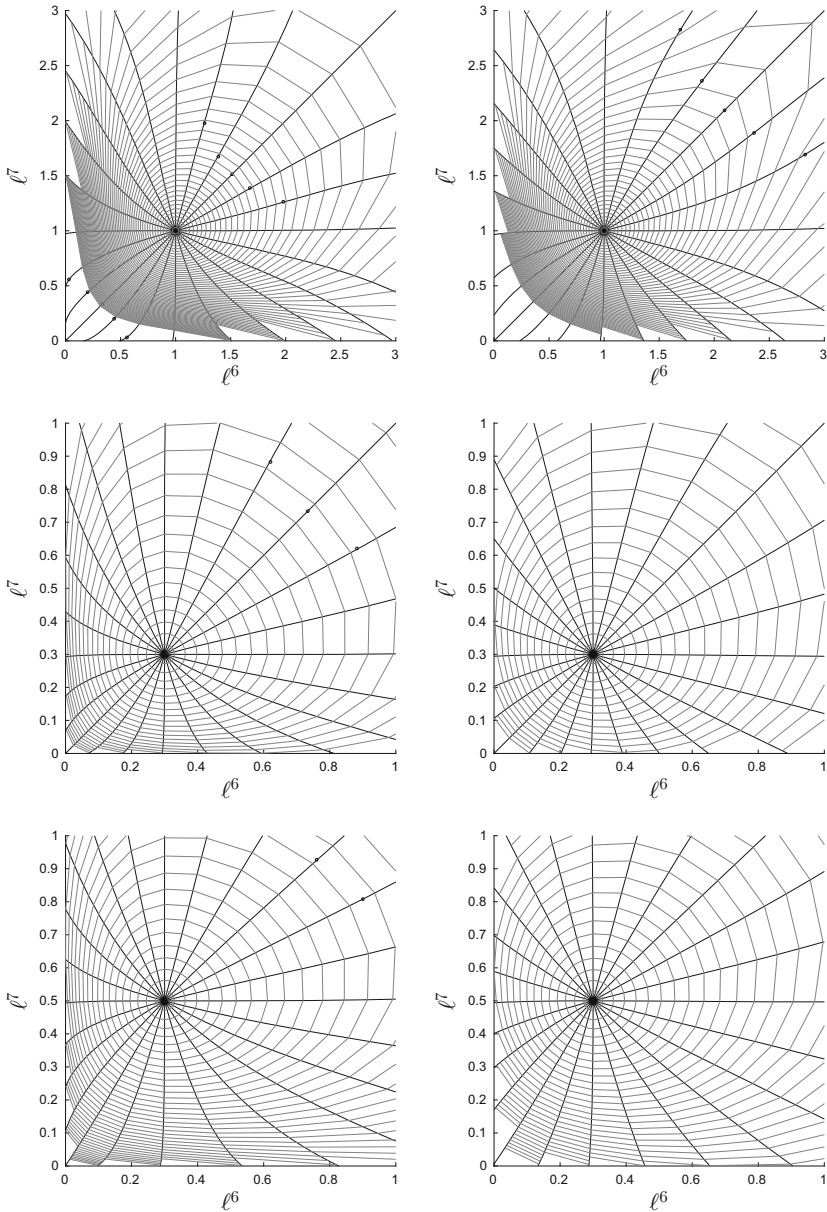
### 3.4 Numerical investigation of the geometry

The geodesic equation (3.4) can be solved numerically on the interior of any maximal orthant given some initial conditions  $\ell(0) = \ell_0$  and  $d\ell(0)/dt = \mathbf{v}_0$ . As described in Sect. 2.3, the first and second derivatives of  $p_\ell(s)$  with respect to the edge lengths  $\ell^i$  can be computed analytically. Calculation of  $g_{ij}(\ell)$  consists of a sum over all  $2^N$  possible characters involving the derivatives of  $p_\ell(s)$ . The Christoffel symbols can similarly be calculated as sums over characters. A fourth-order Runge–Kutta method was used to integrate the ODEs.

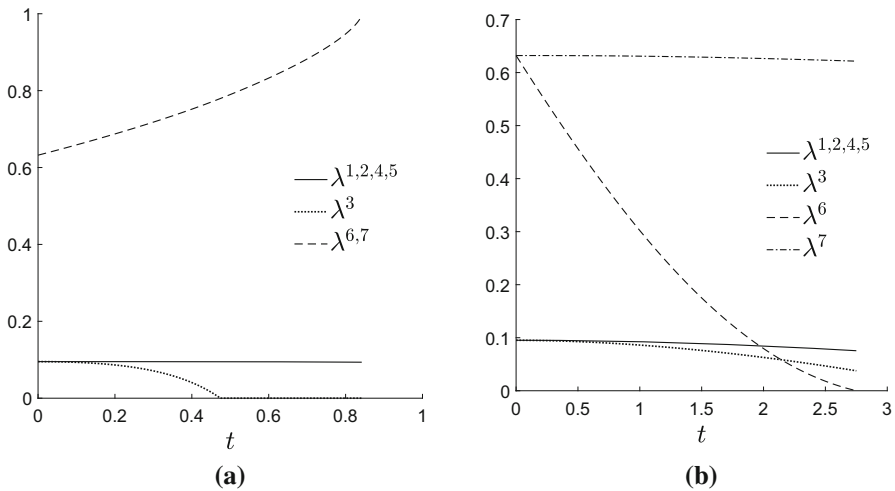
This numerical scheme was used to construct and visualize geodesics on a single orthant in  $\mathcal{W}_5$ . The particular topology and edge lengths for the orthant are represented by the Newick string  $((1 : \ell^1, 2 : \ell^2) : \ell^6, 3 : \ell^3, (4 : \ell^4, 5 : \ell^5) : \ell^7)$ . Parameters  $\ell^1, \dots, \ell^5$  are the pendant edge lengths and  $\ell^6, \ell^7$  the lengths of the two internal edges. We restricted to  $N = 5$  leaves in order to enable easy visualization of geodesics. Integration was stopped whenever any internal edge was assigned a length  $\leq 0$ , corresponding to the boundary of the orthant. If this occurred for a pendant edge, the pendant edge length was given value zero at that step, and integration was continued.

Figure 4 shows typical results. The figure shows the orthant representing the two internal edge lengths, with geodesics ‘fired’ from some fixed starting point  $\ell_0$  in 24 different directions. Also marked on the plots are contours of distance from the starting point. Each panel shows results for a different initial tree  $\ell_0$ . It is evident that the geodesics are not the same as BHV geodesics, which are straight lines radiating from the initial point with equally spaced circular contours of distance. Figure 4 shows curved geodesics with irregularly spaced contours of distance. Contours appear to stack up towards the origin and codimension-1 BHV boundaries, but are more spaced out as geodesics move out towards the boundaries at infinity. This is more obvious when initial internal edges are long (top row of Fig. 4). On the other hand, the geodesics are more similar to the BHV geodesics when internal edges are short and pendant edges are long, as in the bottom two rows of the right hand column. In all cases, in contrast to BHV tree space, geodesics seem to be slightly attracted toward the star trees. The





**Fig. 4** Solutions to the geodesic equation (black radiating curves) and contours of distance (grey) in a single maximal orthant of  $\mathcal{W}_N$ . Each panel shows the trajectories for the internal edge lengths  $\ell^6, \ell^7$ . Rows correspond to different initial sets of internal edge lengths. Columns correspond to different initial pendant edge lengths:  $\ell^i = 0.1$  for  $i = 1, \dots, 5$  on the left, and  $\ell^i = 0.5$  for  $i = 1, \dots, 5$  on the right. The initial velocity on pendant edge lengths was zero in all cases. Dots mark points at which a pendant edge was assigned length zero, and at all subsequent points, to avoid negative values. Contours near the origin in the top two plots have been removed: they stack up as the origin is approached and obscure the appearance of the geodesics



**Fig. 5** Graphs showing edge weights  $\lambda^i$  vs time along geodesics in top left panel of Fig. 4. **a** Geodesic heading in North East compass direction. **b** Geodesic heading West

pendant edges do not behave as they do in BHV tree space: they can change value even when their initial velocity is zero.

Figure 5 provides more detail for certain geodesics in Fig. 4. The graphs in the figure show each edge weight  $\lambda^i$  versus time, and the time is proportional to distance travelled. The  $\lambda$ -parametrization was used for these plots since it shows the results most clearly. Panel (a) shows that, when contours become more and more widely spaced, the boundary at infinity ( $\lambda^6 = \lambda^7 = 1$ ) can be reached in finite time, rather than asymptotically. This shows that points corresponding to trees with infinitely long edges are a finite distance away from the starting point, as established by Theorem 3.1. In the  $\ell$ -parametrization, the internal edge lengths rapidly blow up to infinity after time  $t = 0.8$ . It follows that the shortest path between two trees with finite edge lengths might involve trees with infinitely long edges. On the other hand, for some panels in Fig. 4, the contours become increasingly close as BHV boundaries are approached, but as panel (b) in Fig. 5 shows, points on BHV boundaries are in fact finitely close to orthant interiors, since the boundary is reached in finite time.

Figure 6 replots the top left and middle left panels from Fig. 4 using the  $\lambda$ -parametrization, so that the boundary at infinity corresponds to edges of the unit square with weight 1. These plots suggest that trees in which one of the two internal edges is finitely long are ‘repellant’ since the geodesics fired in the North and East compass directions end up passing through the disconnected forest with  $\lambda^6 = \lambda^7 = 1$ . Indeed, the points on the boundary with  $\lambda^6 = \lambda^7 = 1$  appear to be ‘attractive’, with geodesics pulled round to pass through these points and arriving in finite time.

While these results show how the information geometry on  $\mathcal{W}_N$  differs substantially from the BHV geometry, the method for constructing geodesics by integrating the geodesic differential equation suffers from several disadvantages. First, it requires summation over the elements of  $\{0, 1\}^N$  which makes it infeasible for large  $N$  (exponential computation time in  $N$ ). Secondly, only local geodesics are computed by

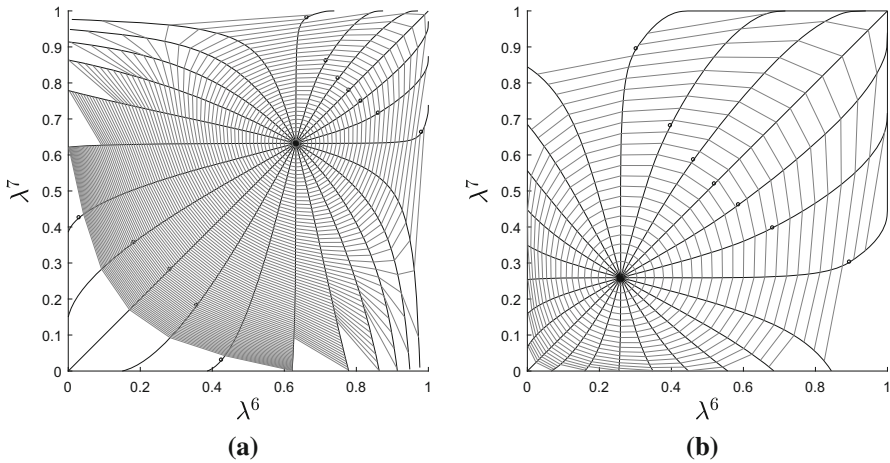


Fig. 6 Top left and middle left panels on Fig. 4 redrawn in the  $\lambda$ -parametrization

solving the initial value problem (geodesic “shooting” or “firing”) for the differential equation valid only in maximal orthants. Thirdly, for practical applications, an algorithm which takes two points in wald space and joins them by a globally shortest path is more useful, and so it is desirable to solve the boundary value problem for geodesic construction. The next section attempts to deal with some of these shortcomings.

### 4 Information geometry for a Gaussian process on trees

In this section we develop the information geometry of a continuous-valued Markov process associated to each tree which is more computationally and analytically tractable than the information geometry for the symmetric two-state Markov process. It has the advantage that the geodesic equation (3.4) can be solved numerically much faster than the corresponding equation for the symmetric two-state model, but the solutions for the two models are very similar.

#### 4.1 Definition of the Gaussian process

Our aim is to construct a Gaussian process which is a continuous-valued analogue of the symmetric two-state process by matching the moments specified in Lemma 2.1. Consider the Ornstein–Uhlenbeck process  $Z(t)$  on  $T$  which satisfies

$$Z(t_2) | Z(t_1) = z \sim N(z e^{-\ell_{t_1 t_2}}, 1 - e^{-2\ell_{t_1 t_2}}) \tag{4.1}$$

where  $\ell_{t_1 t_2}$  is the path length distance between  $t_1$  and  $t_2$  on  $T$ . The stationary distribution is the standard normal distribution  $N(0, 1)$ , and we assume the process is stationary over  $T$ . The Markov process satisfies the detailed balance equation, and so is reversible in its stationary state. As a result, realizations of the process can be simulated by fixing

an arbitrary root  $t_0 \in T$  for the tree, simulating  $Z(t_0)$  from  $N(0, 1)$  and then using Eq. (4.1) to simulate a realization  $Z(t)$  for any other  $t \in T$ . Reversibility of the process ensures the distribution obtained is invariant of the choice of root. A short calculation shows that the covariance matrix of the random variables  $Z_1, \dots, Z_N$  at the leaves of  $T$  is given by  $\text{Cov}(Z_u, Z_v) = \exp(-\ell_{uv})$ . Since  $\ell_{uu} = 0$  for all  $u = 1, \dots, N$ , this gives  $\text{Var}(Z_u) = 1$ . Similarly, the conditional distribution of  $Z(t_2)$  given  $Z(t_1) = z$  has variance  $1 - \exp(-2\ell_{t_1 t_2})$ . These moments match those in Lemma 2.1 up to the constant factor of  $1/4$ , and we will show later that this factor makes no difference to the geometry obtained. The process  $Z(t)$  can be thought of in two ways. First, it approximates the binomial random variables obtained when many independent binary characters evolve under the symmetric two-state process. Secondly, it could be regarded as an evolutionary model of a continuous trait for which the observations are standardized to have zero mean and unit variance in each population. Mean-reverting Gaussian processes like this are sometimes used to model continuous traits for which there is some constant optimal value for survival (Hansen and Martins 1996). The definition of  $Z(t)$  extends from trees to forests by taking the process to be independent on each connected component.

Given a forest  $F \in \mathcal{W}_N$ , the distribution of the random variables  $Z_1, \dots, Z_N$  at the leaves is multivariate normal with zero mean and covariance matrix  $S_F$  where

$$S_F = \left( \exp(-\ell_{uv}) \right)_{u,v=1}^N. \tag{4.2}$$

The path distance  $\ell_{uv}$  is the sum of the lengths  $\ell^e$  on edges  $e$  between  $u$  and  $v$ , and is taken to be infinite when  $u$  and  $v$  are in different components of  $F$ . This defines a map  $F \mapsto \phi_F$  from forests to multivariate normal distributions where  $\phi_F$  is the probability density function of  $N(\mathbf{0}, S_F)$ . A similar result to Lemma 2.2 holds: whenever  $F_1 \sim F_2$ , the distributions  $\phi_{F_1}$  and  $\phi_{F_2}$  are the same, so the map is well-defined on  $\mathcal{W}_N$ .

### 4.2 A Gaussian process geometry for the wald space

The information geometry of multivariate normal distributions with zero mean has been studied previously (Lenglet et al. 2006) and is analytically tractable. The theory described in Sect. 3 needs adapting to account for the change from discrete to continuous characters. The Fisher information metric of  $\mathcal{W}_N$  in Eq. (3.3) becomes an integral over  $\mathbb{R}^N$  rather than a sum, and the mass function  $p_\ell$  is replaced with the density function  $\phi_\ell$  of the Gaussian corresponding to a fully resolved tree with edge lengths  $\ell$ :

$$g_{ij}(\ell) = \int_{\mathbb{R}^N} \phi_\ell(s) \left( \partial_i \log \phi_\ell(s) \right) \left( \partial_j \log \phi_\ell(s) \right) ds. \tag{4.3}$$

The geodesic equation (3.4) remains the same. Although  $p_T(s)$  and its derivatives could be evaluated exactly for the two-state symmetric model, evaluation of the metric and Christoffel symbols required a sum over all characters. For the continuous model, the corresponding integrals have closed form, as we describe below.

Gaussian distributions with zero mean are parametrized by their covariance matrices, namely by the space of  $N \times N$  symmetric positive definite matrices, which we

will denote  $\mathcal{S}_N^+$ . The set of covariance matrices associated with forests forms a subset within  $\mathcal{S}_N^+$ , as determined by the following theorem.

- Theorem 4.1** 1. *The covariance matrix  $S_F$  associated to any  $F \in \mathcal{W}_N$ , as defined by Eq. (4.2), is positive definite so lies in  $\mathcal{S}_N^+$ , and*  
 2. *the map  $[F] \mapsto S_F$  from  $\mathcal{W}_N$  to  $\mathcal{S}_N^+$  is injective and so it determines a well-defined embedding of  $\mathcal{W}_N$  into  $\mathcal{S}_N^+$ .*

The proof is given in the Appendix.

For Gaussians with zero mean, it can be shown that the Fisher information metric at a point with covariance matrix  $S$  is

$$\langle X, Y \rangle = \frac{1}{2} \text{tr} \left( S^{-1} X S^{-1} Y \right),$$

where  $X, Y$  are matrices in the tangent space at  $S$ , i.e. symmetric matrices (Lenglet et al. 2006). This expression is obtained by evaluating the integral in Eq. (4.3). Working in some fixed maximal orthant parametrized by edge lengths  $\ell$ , let  $S_\ell$  be the corresponding covariance matrix defined in Eq. (4.2). For each edge  $e \in F$  define the *split matrix*  $\sigma^e$  by

$$\sigma_{uv}^e = \begin{cases} 1, & \text{if } e \text{ lies on the path from leaf } u \text{ to leaf } v, \text{ and} \\ 0, & \text{otherwise,} \end{cases} \tag{4.4}$$

for  $u, v = 1, \dots, N$ . Then the path length between  $u$  and  $v$  is  $\ell_{uv} = \sum_e \ell^e \sigma_{uv}^e$  where the sum is over all edges in  $F$ . Equation (4.2) becomes

$$S_F = S_\ell = \left( \prod_e \exp \left( - \ell^e \sigma_{uv}^e \right) \right)_{u,v=1}^N. \tag{4.5}$$

An entry above is zero if  $u$  and  $v$  are in different connected components, or equivalently, if they are separated by an infinitely long edge.

By differentiating Eq. (4.5), it can be seen that the tangent space at  $S_\ell$  is spanned by matrices of the form  $\sigma^e \circ S_\ell$  for each edge  $e$ , where  $\circ$  denotes the Hadamard matrix product. The Fisher information metric (4.3) for  $\mathcal{W}_N$  for a fully resolved tree becomes

$$g_{ij}(\ell) = \frac{1}{2} \text{tr} \left( S_\ell^{-1} (S_\ell \circ \sigma^i) S_\ell^{-1} (S_\ell \circ \sigma^j) \right) \tag{4.6}$$

where  $i, j = 1, \dots, 2N - 3$  index edges. Algebraic expressions for the first and second derivatives of the Fisher information metric can similarly be obtained, and hence for the Christoffel symbols. Note that scaling  $S_\ell$  by some positive constant has no effect on the metric, and so the factor of 1/4 difference between the covariance matrices obtained from the discrete process  $X(t)$  and continuous process  $Z(t)$  has no effect on the geometry.

The inner product and its derivatives can be computed in polynomial time and the paths obtained by integrating the geodesic ODE for the continuous Markov model within orthant interiors in  $\mathcal{W}_N$  resemble those for the two-state model very closely.

Namely, results for the same initial conditions as Fig. 4 were obtained but omitted, since the plots were almost indistinguishable from those for the two-state model. However, the inner products defined using the two different models are not identical: both inner products can be written down explicitly in the case  $N = 2$ , using the transition probabilities for the discrete model and Eq. (4.6) for the continuous model. The two inner products differ when the length of the single edge in the tree is small, but converge as the edge length tends to infinity.

Using Eq. (4.6) and its derivatives, we derived algebraic expressions for the Riemannian curvature tensor and the sectional curvatures at any point in  $\mathcal{W}_N$ . We implemented these expressions in R, and hence evaluated these quantities for certain trees in  $\mathcal{W}_5$ . We found that at randomly selected points in  $\mathcal{W}_5$ , and hence in all spaces with  $N \geq 5$ , the sectional curvatures had mixed signs. As a result, there is no global sign condition on curvature like that for BHV tree space, which is globally non-positively curved.

### 5 Geometry via embedding in $\mathcal{S}_N^+$

The information geometry on Gaussians with zero mean can equivalently be regarded as a geometry for the space of  $N \times N$  symmetric positive definite matrices  $\mathcal{S}_N^+$ . This is a useful viewpoint to adopt, first because it highlights the fact that the geometry we develop on  $\mathcal{W}_N$  is based entirely on the covariance between the leaves induced by the Markov process  $Z(t)$ , and secondly, because other metrics on  $\mathcal{S}_N^+$  have been studied (Dryden et al. 2009). These alternative metrics on  $\mathcal{S}_N^+$  could in turn define different metrics on  $\mathcal{W}_N$ , although they will not be considered any further in this paper. The metric on  $\mathcal{S}_N^+$  obtained from the information geometry of Gaussian distributions with zero mean will be denoted  $d_{\text{cov}}$ . The metric and its associated geodesics in  $\mathcal{S}_N^+$  can be computed in polynomial time (Lenglet et al. 2006). The main idea in this section is to use the analytically tractable geometry in  $\mathcal{S}_N^+$ , combined with a projection algorithm from the ambient space  $\mathcal{S}_N^+$  to the embedded space  $\mathcal{W}_N$ , to construct approximate geodesics in  $\mathcal{W}_N$ .

Given the embedding  $[F] \mapsto S_F$  of  $\mathcal{W}_N$  within  $\mathcal{S}_N^+$ , we can consider the intrinsic metric  $d_{\text{cov}}^*$  on  $\mathcal{W}_N$  induced by  $d_{\text{cov}}$ . By construction, the induced metric corresponds to the information geometry on  $\mathcal{W}_N$  for the continuous Markov model considered in Sect. 4. The following theorem is analogous to Theorem 3.1 for the discrete Markov substitution models. A proof is given in the appendix.

**Theorem 5.1** *For any  $[F], [G] \in \mathcal{W}_N$  the induced intrinsic metric  $d_{\text{cov}}^*([F], [G])$  is finite and therefore well-defined. Any path which realizes the distance  $d_{\text{cov}}^*([F], [G])$  is a solution of Eq. (3.4) at any point in the interior of a maximal orthant, where the Riemannian inner product is given by Eq. (4.5).*

Lenglet et al. (2006) give formulae for the distance and geodesics between pairs of points in  $\mathcal{S}_N^+$ . The distance between  $S_1, S_2 \in \mathcal{S}_N^+$  is defined by

$$d_{\text{cov}}(S_1, S_2)^2 = \frac{1}{2} \text{tr} \left( \log \left( S_1^{-1/2} S_2 S_1^{-1/2} \right)^2 \right) \tag{5.1}$$

where  $\log$  denotes the matrix logarithm.

Since the map from  $\mathcal{W}_N$  to  $\mathcal{S}_N^+$  is injective,  $d_{\text{cov}}$  pulls back to define an extrinsic metric on  $\mathcal{W}_N$ :

$$d_{\text{cov}}([F_1], [F_2]) = d_{\text{cov}}(S_{F_1}, S_{F_2}). \tag{5.2}$$

In fact, the space  $\mathcal{S}_N^+$  equipped with  $d_{\text{cov}}$  has globally non-positive curvature (Skovgaard 1984; Ballmann et al. 1985) and so there is a unique geodesic between any two points in the ambient space. The point at proportion  $t \in [0, 1]$  along the geodesic between  $S_1, S_2 \in \mathcal{S}_N^+$  is

$$\Gamma_{S_1, S_2}(t) = S_1^{1/2} \exp(tU) S_1^{1/2} \quad \text{where} \quad U = \log\left(S_1^{-1/2} S_2 S_1^{-1/2}\right). \tag{5.3}$$

Equations (5.1) to (5.3) involve eigen-decompositions of  $N \times N$  matrices, and so can be computed in  $O(N^4)$  steps.

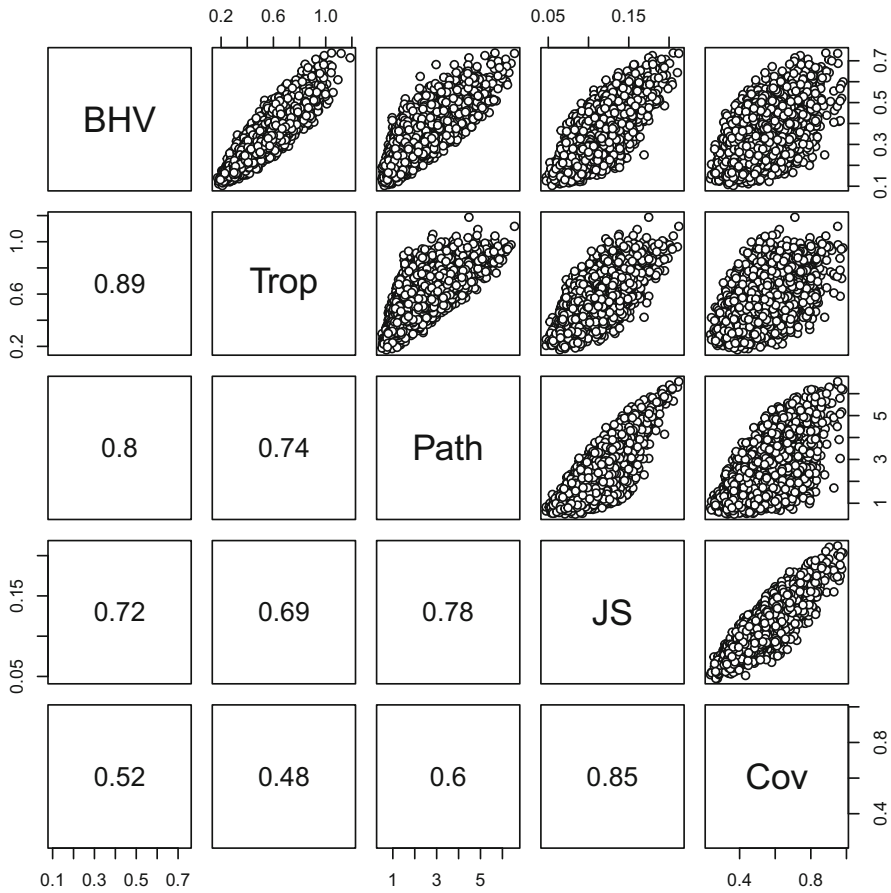
Figure 7 shows a comparison of BHV metric  $d_{\text{BHV}}$ , tropical metric, path distance metric, Jensen–Shannon metric  $d_{JS}$  and  $d_{\text{cov}}$  for every pair of trees in a sample of 100 trees obtained by bootstrap replication during maximum likelihood inference of a phylogenetic tree. The trees were inferred using the MrBayes software (Huelsenbeck and Ronquist 2001), and a sample data set of DNA from 12 primates provided with the software. The path difference metric (Steel and Penny 1993) between trees  $T, T'$  is  $(\sum_{u,v} (\ell_{uv} - \ell'_{uv})^2)^{1/2}$ . The Jensen–Shannon metric was calculated exactly by summing over all  $2^{12}$  characters for the two-state model, as described in Sect. 3.2. The covariance metric was calculated using Eqs. (5.1) and (5.2). The BHV and tropical metrics are quite closely correlated, as are the Jensen–Shannon and covariance matrices. The path difference metric has a similar correlation with the BHV, tropical and Jensen–Shannon metrics (approximately 0.7–0.8), but is relatively weakly correlated with  $d_{\text{cov}}$ . This suggests that the BHV and tropical metrics are based on features of the data which are rather distinct from those for the Jensen–Shannon and covariance metrics. The covariance metric has the advantage over the Jensen–Shannon metric of being computable in polynomial time.

### 5.1 Projection into wald space

To approximate geodesics in the extrinsic covariance metric  $d_{\text{cov}}$  of  $\mathcal{W}_N$ , we construct a projection from  $\mathcal{S}_N^+$  onto  $\mathcal{W}_N$ , that is, given  $S_0 \in \mathcal{S}_N^+$ , we aim to find an element  $[F] \in \mathcal{W}_N$  which minimizes  $d_{\text{cov}}(S_0, S_F)$ . Suppose that  $F$  is a fully resolved tree with edge lengths  $\ell$ . The expression for  $d_{\text{cov}}(S_0, S_F)^2$  can be differentiated with respect to edge lengths of  $F$  and gives

$$\partial_i d_{\text{cov}}(S_0, S_F)^2 = \text{tr}\left(\log\left(S_0^{-1/2} S_F S_0^{-1/2}\right) S_0^{1/2} S_F^{-1} (\partial_i S_F) S_0^{-1/2}\right)$$

where  $\partial_i = \partial/\partial \ell_i$  (e.g. Moakher (2005)). Moreover,  $\partial_i S_F = S_F \circ \sigma^i$  where  $\circ$  denotes the Hadamard or element-wise matrix product and  $\sigma^i$  is the split matrix associated with edge  $i$ , as defined in Eq. (4.4).



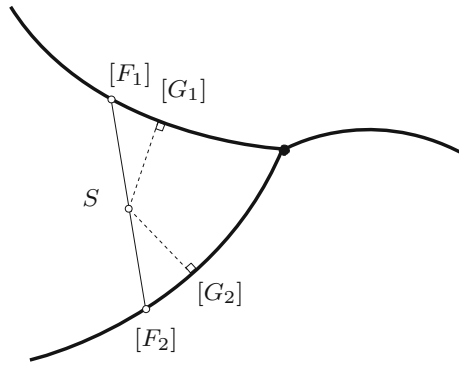
**Fig. 7** Comparison of the BHV metric  $d_{\text{BHV}}$ , tropical metric, path difference metric, Jensen–Shannon metric  $d_{\text{JS}}$  and  $d_{\text{cov}}$  for every pair from a sample of 100 trees obtained from bootstrap replicates during maximum likelihood estimation. The trees were inferred from DNA data from 12 species of primate. The correlation coefficients are shown in the bottom left panels

This analytic expression for the derivative can be used to implement a gradient descent algorithm. Within each maximal orthant  $\mathcal{O}_\tau$  the edge lengths were updated according to the rule

$$\ell_{k+1} = \ell_k - \alpha_k \nabla d_{\text{cov}}(S_0, S_{\ell_k})^2$$

where  $\ell_k$  denotes the edge lengths at iteration  $k$  and  $S_{\ell_k}$  the corresponding covariance matrix. The step size  $\alpha_k$  was determined using the Barzilai–Borwein method. Two versions of the algorithm were used. In the first, the algorithm was halted whenever an internal edge was assigned a negative length. As a result, the algorithm was constrained to lie within the orthant  $\mathcal{O}_\tau$  containing the initial tree. This algorithm was used for  $N = 5$  by running the algorithm 15 times, each time with an initial tree in one of the 15 maximal orthants of  $\mathcal{U}_5$ , and the overall tree closest to  $S_0$  found. The second version of the algorithm was able to cross codimension-1 BHV boundaries as follows.





**Fig. 8** Schematic diagram for  $N = 4$  showing three neighbouring orthants (curved heavy lines) embedded in  $S_N^+$ . The black circle is the BHV boundary between the orthants. The extrinsic geodesic between trees  $[F_1], [F_2]$  is depicted as a straight line segment. The projection of this segment consists of a path from  $[F_1]$  to  $[G_1]$  within the orthant, but then jumps to  $[G_2]$  in a different orthant. The dashed lines show the orthogonal projection of the point  $S$  along the extrinsic geodesic, and  $S$  is equidistant from  $[G_1]$  and  $[G_2]$

If an edge length was assigned a negative value, then trees in the two corresponding neighbouring orthants to  $\tau$  were considered, taking the absolute value of elements in  $\ell_{k+1}$  as edge lengths. The tree at step  $k + 1$  was taken to be whichever of these two trees was closest to  $S_0$ .

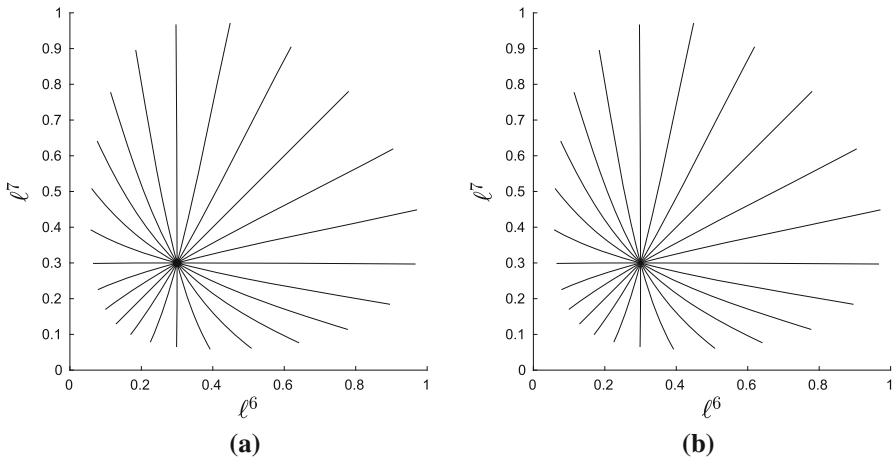
In general, the closest point in  $\mathcal{W}_N$  to a covariance matrix  $S_0 \in S_N^+$  is not necessarily unique as illustrated in Fig. 8. Moreover, the gradient descent algorithm can converge to local minima, and so the result obtained is sensitive to the tree used to initialize the algorithm.

### 5.2 Construction of geodesics in $\mathcal{W}_N$ via projection of extrinsic geodesics

Since construction of geodesics in  $S_N^+$  between any two given points and projection from  $S_N^+$  into  $\mathcal{W}_N$  can both be performed efficiently, we aim to combine these algorithms to give an efficient means of constructing geodesics within  $\mathcal{W}_N$  between any two given points. A naive approach, given  $[F_1], [F_2] \in \mathcal{W}_N$ , is to simply project the extrinsic geodesic between  $S_{F_1}$  and  $S_{F_2}$  from  $S_N^+$  into  $\mathcal{W}_N$ . This approach works for the example of the unit sphere  $S^2$  embedded within  $\mathbb{R}^3$ : the projection of the chord between two points in the sphere is a great circle between those two points. However, this approach fails for  $\mathcal{W}_N \subseteq S_N^+$  since the projected paths are often discontinuous and jump between different orthants, as illustrated in Fig. 8.

The following **recursive algorithm** for constructing an approximate geo-desic in  $\mathcal{W}_N$  is therefore proposed, which is intended to overcome this issue. Let  $t_i = i/k$  for  $i = 0, \dots, k$  and suppose  $[F_1], [F_2] \in \mathcal{W}_N$ . The algorithm outputs a sequence  $[G_0], \dots, [G_k] \in \mathcal{W}_N$  where  $[G_0] = [F_1]$  and  $[G_k] = [F_2]$ . For each iteration  $i = 1, \dots, k - 1$  of the algorithm, the following steps are performed.

1. Compute the extrinsic geodesic  $\Gamma$  from  $S_{G_{i-1}}$  to  $S_{G_k}$  using Eq. (5.3).
2. Find the point  $S \in S_N^+$  at proportion  $1/(k - i + 1)$  along  $\Gamma$ .



**Fig. 9** Comparison of paths obtained by **a** integrating the geodesic ODE for the Gaussian process model and **b** by applying the symmetrized projection method to the end points obtained in **(a)**

3. Let  $[G_i]$  be the projection of  $S$  into  $\mathcal{W}_N$ .

The idea is that at each iteration, a new extrinsic geodesic is constructed from the previous point  $[G_{i-1}]$  to the destination  $[F_2]$ , a small step is taken along that geodesic, and that point is projected into  $\mathcal{W}_N$  to give  $[G_i]$ . For the results in this paper, the projection at Step 3 was performed using the second version of the projection algorithm described in Sect. 5.1, rather than using the less efficient search over all orthants. The gradient descent for the projection to obtain  $[G_i]$  at Step 3 was initialized using the edge lengths from the forest  $[G_{i-1}]$ .

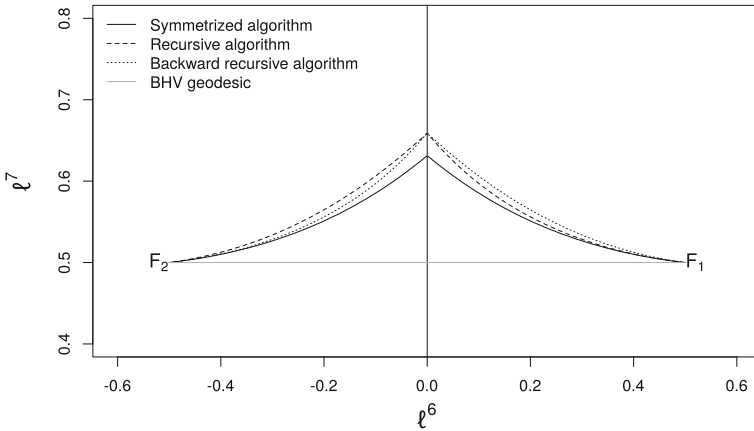
This algorithm has the disadvantage that it is not symmetric under swapping the end points  $[F_1], [F_2]$ , whereas the image of the geodesic should be invariant under this operation.

The following **symmetrized version** of the algorithm overcomes this issue.

The algorithm produces a sequence  $[G_0], \dots, [G_k], [H_k], \dots, [H_0] \in \mathcal{W}_N$  where the initial values are taken to be  $[G_0] = [F_1]$  and  $[H_0] = [F_2]$ . For each iteration  $i = 1, \dots, k - 1$  of the algorithm, the following steps are performed.

1. Compute the extrinsic geodesic  $\Gamma$  from  $S_{G_{i-1}}$  to  $S_{H_{i-1}}$  using Eq. (5.3).
2. Find the points  $R, S \in S_N^+$  at proportions  $1/(k - i + 1)$  and  $1 - 1/(k - i + 1)$  along  $\Gamma$ .
3. Let  $[G_i]$  and  $[H_i]$  be the projections of  $R$  and  $S$  into  $\mathcal{W}_N$  respectively.

The quality of the approximate geodesics produced by the symmetrized algorithm can be assessed by comparing them to geodesics in a single orthant constructed by integrating the geodesic equation as described in Sect. 4.2. Given an initial tree  $[F_1] \in \mathcal{W}_5$  and an initial velocity for  $\ell$ , the geodesic equation was integrated until the path obtained reached some specified length. The final point reached was taken to be  $[F_2]$ , and the symmetrized algorithm was then used to obtain an approximate geodesic between  $[F_1]$  and  $[F_2]$ . In all cases, the paths obtained with the two methods matched



**Fig. 10** Comparison of approximate geodesics in  $\mathcal{W}_5$  constructed between trees  $F_1$  and  $F_2$  from (5.4) in neighbouring orthants. The vertical axis  $\ell^7$  represents a codimension-1 BHV boundary between two orthants. When, due to a nearest neighbor interchange, crossing it,  $\ell^6$  tends to zero, another edge appears, with negative length corresponding to the negative values on the  $\ell^6$  axis. Three approximate geodesics are shown: (i) construction via the recursive algorithm from  $F_1$  and  $F_2$ , (ii) using the same algorithm but reversing the end-points, and (iii) construction via the symmetrized algorithm

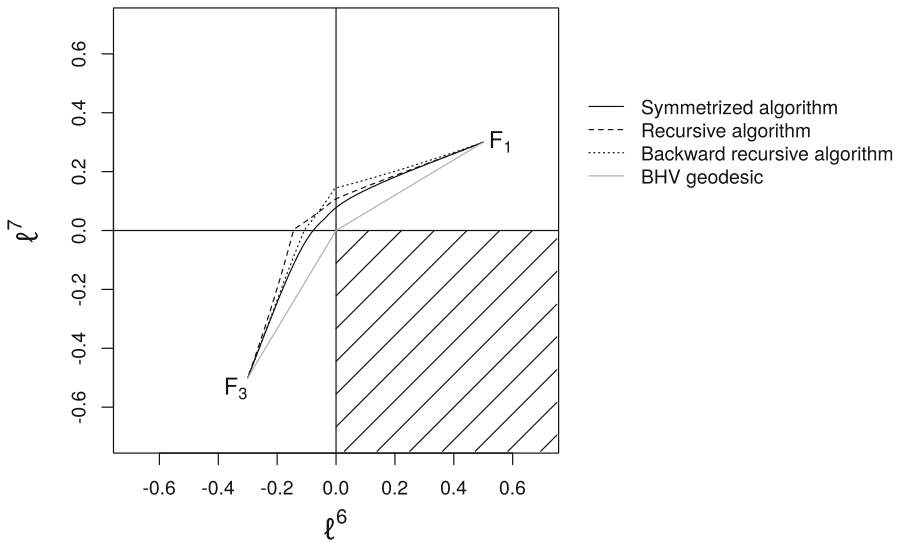
very closely, with the quality of the match improving for shorter internal edge lengths. Figure 9 shows typical results.

In contrast to the methods presented in Sects. 3 and 4, these algorithms are not based on ‘firing’ geodesics, and can produce approximate geodesics between end points in different orthants. Figures 10 and 11 show results obtained when the end points are separated by 1 or 2 nearest neighbour interchange operations respectively in  $\mathcal{W}_5$ . More precisely, we consider trees corresponding to Newick strings

$$\begin{aligned}
 F_1 &: ((1 : 0.1, 2 : 0.1) : \ell^6, 3 : 0.1, (4 : 0.1, 5 : 0.1) : \ell^7), \\
 F_2 &: ((2 : 0.1, 3 : 0.1) : \ell^6, 1 : 0.1, (4 : 0.1, 5 : 0.1) : \ell^7) \quad \text{and} \quad (5.4) \\
 F_3 &: ((2 : 0.1, 3 : 0.1) : \ell^6, 5 : 0.1, (1 : 0.1, 4 : 0.1) : \ell^7).
 \end{aligned}$$

In both figures, the approximate geodesics constructed using the recursive algorithm are not symmetric under interchange of the end points, and differ from the paths obtained using the symmetric algorithm. The lengths of the paths can be computed by using large  $k$  and summing  $d_{\text{cov}}$  between successive points in the output. In all the examples we explored, the symmetrized algorithm produced shorter paths. The approximate geodesics shown in the figures are significantly different from BHV geodesics, which consist of straight (or once broken) lines between the end points in both plots.

We apply the symmetrized algorithm to investigate the distance from trees in the interior of a maximal orthant to the star stratum on the boundary of that orthant. If  $[G_1], [G_2], \dots, [G_k] \in \mathcal{W}_N$  is an approximated geodesic between  $[G_1]$  and  $[G_k]$  computed by the symmetrized algorithm, the intrinsic distance between  $[G_1]$  and  $[G_k]$  can be approximated by  $d_{\text{cov}}^*([G_1], [G_k]) \approx \sum_{i=1}^{k-1} d_{\text{cov}}([G_i], [G_{i+1}])$ . Consider the



**Fig. 11** Comparison of approximate geodesics constructed between trees  $F_1$  and  $F_3$  from (5.4) in orthants separated by two nearest neighbour interchanges. Three neighbouring orthants in  $\mathcal{W}_5$  are shown, and the bottom right-hand orthant does not correspond to a valid tree topology. As in Fig. 10, negative values on axes correspond to negative lengths of new edges

following setup. For  $\lambda_0 \in (0, 1]$ , let  $G_1 = G(\lambda_0)$  be the forest corresponding to the Newick string  $((1 : \lambda_0, 2 : \lambda_0) : \lambda_0, (3 : \lambda_0, 4 : \lambda_0))$  in  $\lambda$ -parametrization. This is a fully resolved 4-taxon tree on which each edge has weight  $\lambda_0$ . By symmetry, the edges on the tree in the star stratum closest to  $G(\lambda_0)$  must all have equal weight  $\lambda \in (0, 1]$ . Thus, let  $G_k = F(\lambda)$  be the star tree corresponding to the Newick string  $(1 : \lambda, 2 : \lambda, 3 : \lambda, 4 : \lambda)$ , again in  $\lambda$ -parametrization. Figure 12 shows for each  $\lambda_0 \in \{0.1, 0.5, 0.9, 0.95\}$  the approximated values of  $d_{\text{cov}}^*([G(\lambda_0)], [F(\lambda)])$  against  $\lambda$ . Obviously,  $F(\lambda)$  is closest for  $\lambda$  slightly larger than  $\lambda_0$  with distance decreasing as  $\lambda_0 \rightarrow 1$ . This suggests that the star stratum is closer to the tree  $G(\lambda_0)$  than the forest consisting of 4 isolated points (obtained from  $F(\lambda)$  as  $\lambda \rightarrow 1$ ), even for  $\lambda_0$  values close to 1. Note, though, that the forest is a boundary point of the star stratum. For any  $G(\lambda_0)$  the distance to  $F(\lambda)$  tends to infinity as  $\lambda \rightarrow 0$ . Indeed,  $S_{F(0)} \notin S_4^+$  is not of full rank.

### 6 Discussion

In order to do statistical inference on data sets of phylogenetic trees one needs a structure rich enough to enable the use of geometric statistical methods. Recent research has produced geometries such as the BHV and tropical tree spaces and statistical methods adapted to these geometries. Based on a more principled set of underlying assumptions by regarding phylogenetic trees as probability models for genetic sequence data, we have developed a canonical and biologically motivated geometry on tree space by applying tools from information geometry, giving the wald space. In particular,

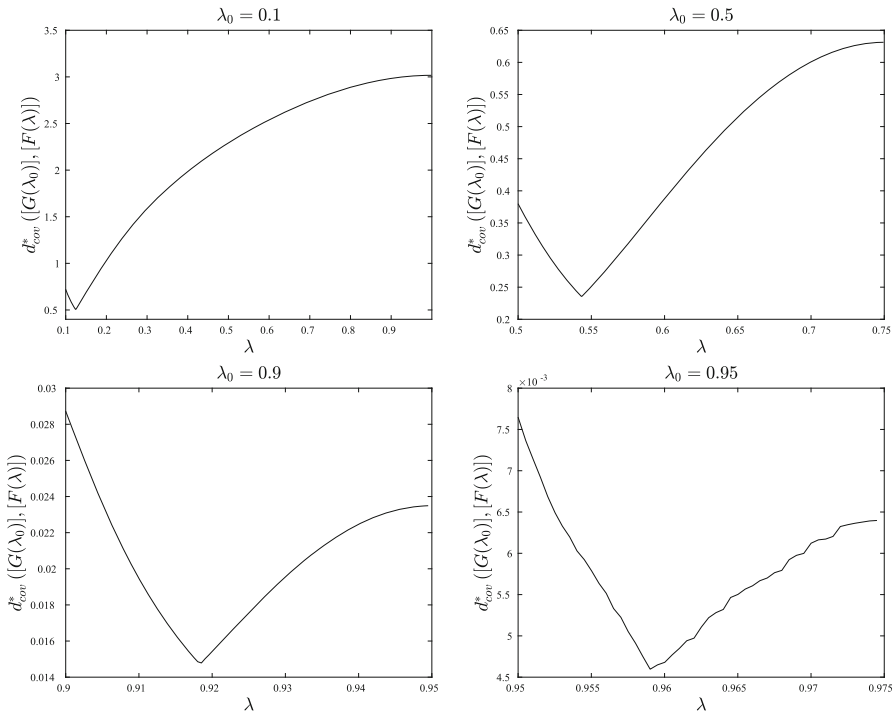


Fig. 12 Approximated distances  $\lambda \mapsto d_{cov}^*([G(\lambda_0)], [F(\lambda)])$  for different values of  $\lambda_0$

unlike previous related work (Garba et al. 2018) in which various extrinsic metrics were considered, in this paper we have focused on developing intrinsic metrics and their associated geodesics to explore and to enable accessing the geometry, for this is a key ingredient for statistical inference on non-Euclidean spaces.

There are two main difficulties with achieving our aim. First, the discrete-valued Markov process on trees with genetic alphabet  $\Omega$  characterizes trees as probability models with sample space  $\Omega^N$  where  $N$  is the number of phylogenetic taxa. Therefore, calculations of distances and construction of geodesics involve summations over  $|\Omega|^N$  terms, which becomes infeasible for large  $N$ . In order to establish computational tractability, we generalized the discrete-valued probability model to a continuous-valued Markov process in a canonical way and applied the information geometry again.

Secondly, information geometry is formulated for parametrized probability models that are a manifold, whereas tree space is a union of manifolds having different dimensions due to the orthants (representing forests with different number of edges) being glued together in a certain way. One has to be careful to compare the structure to the one defined in Moulton and Steel (2004), for example, as we are not including forests with coincident leaves and furthermore we consider a different topology induced by the Fisher information metric. We tackled this issue of not having a single connected parametrized manifold by using the continuous-valued Markov models

to embed wald space in the ambient space of symmetric positive definite matrices, which has an analytically tractable geometry and thus allows for approximation of geodesics in the embedded space  $\mathcal{W}_N$ . Our computational results show that the geometry obtained is significantly different from the BHV and tropical geometries, partly due to the inclusion of trees with infinitely long edges in wald space.

Several questions about the geometry of the wald space  $\mathcal{W}_N$  remain. While we have shown that trees with infinitely long edges are a finite distance away from other trees (Theorems 3.1 and 5.1), computational results suggest that parts of this subspace are repulsive and are avoided by geodesics (see Figs. 6 and 12). An explanation for this behaviour might be obtained via calculations or results about curvature for such points of  $\mathcal{W}_N$ , but further investigation is required. Secondly, Theorems 3.1 and 5.1 establish  $\mathcal{W}_N$  as length spaces for the two induced intrinsic metrics we study. It is desirable to strengthen these results and prove that the distance between every pair of points in the space is realized by at least one path, so that our wald space is a geodesic metric space as opposed to a length space. It appears that such a proof requires thorough analysis of the condition on edge weights which excludes trees with coincident leaves. Furthermore, the methods and results presented in Sect. 5.2 represent a first step towards the development of more sophisticated and efficient algorithms for the construction of information geodesics in wald space via the embedding in the space of covariance matrices. A more thorough evaluation of the computational cost as  $N$  increases could be carried out, and a more rigorous treatment might establish convergence properties for the symmetrized algorithm. Alternatively, existing algorithms taken from computational Riemannian geometry could be adapted to work in wald space (see Schmidt et al. (2006) for example) and might offer better performance.

The underlying motivation for this work has been to obtain a novel geometric framework for the space of phylogenetic trees which has more principled biological justification than existing geometries, and which can be used to develop statistical methods for analysing data sets of trees. Ultimately, realizing this aim is still some way off. For example, given a sample of points  $\{x_1, \dots, x_n\} \subseteq X$  in a metric space  $(X, d)$ , the Fréchet mean  $\bar{x} \in X$  is a point which minimises the sum of squared distances to the data:

$$\bar{x} = \arg \min_{x \in X} \sum_{i=1}^n d(x, x_i)^2.$$

In general, the Fréchet mean does not always exist, or it can fail to be unique, but in globally non-positively curved spaces such as  $(\mathcal{S}_N^+, d_{\text{cov}})$  and  $(\mathcal{U}_N, d_{\text{BHV}})$  there exists a unique Fréchet mean (Bridson and Haefliger 2011). Development of methods for calculation of a Fréchet mean using an intrinsic information metric in wald space seems very challenging, and the curvature calculations in Sect. 4.2 have implications for the existence and uniqueness of Fréchet means. On the other hand, given any sample of trees in  $\mathcal{W}_N$ , there is a unique intrinsic Fréchet mean in  $\mathcal{S}_N^+$  and an algorithm for computing the mean is given by Lenglet et al. (2006). Our projection algorithm could be used to project this to an extrinsic mean back into  $\mathcal{W}_N$ . Properties of the projected Fréchet mean tree could be investigated.

In comparison to the BHV and tropical metrics, the intrinsic information metrics have the advantage of interpretability in terms of genetic substitutions and the distri-

butions of characters represented by two trees. This suggests the information metrics might be better suited for statistical tasks such as hypothesis testing. In the BHV and tropical geometries, contraction and expansion of edges offer the means of moving between different topologies. In the wald space, additional topological transformations are possible via expanding edges to infinite length, and these correspond to tree bisection and reconnection (TBR) operations. Many applications in phylogenetics require searches over the space of phylogenetic trees, and movement along information geodesics in the wald space might have advantages over existing methods.

**Acknowledgements** The second and the last author express their thanks to the Oberwolfach 1804 meeting “Statistics for Data with Geometric Structure” in which wald space was first discussed. The last two authors gratefully acknowledge support from DFG GRK 2088. The last author was supported by the Niedersachsen Vorab of the Volkswagen Foundation.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

### Appendix A: Calculation of $p_T(s)$ and its derivatives

The probability  $p_T(s)$  of any binary character  $s$  on a tree  $T \in U_N$  can be computed efficiently via the following algorithm (Semple and Steel 2003), often called the *Felsenstein pruning algorithm*. First, an arbitrary internal vertex  $v_0 \in T$  is chosen and used to root the tree. The two-state symmetric model is a reversible Markov process, and so the choice of the root does not affect the value of  $p_T(s)$ . The root determines ancestral relations on the tree, and we let  $T_v$  denote the subtree of  $T$  descended from vertex  $v$ . We let  $L_v$  denote the leaves of  $T_v$ , and given a binary characer  $s$ , let  $s_v$  denote the restriction of  $s$  to  $L_v$ . Finally, we let  $p_{T_v}(s_v | \omega)$  be the probability of  $s_v$  on  $T_v$  given the letter  $\omega \in \{0, 1\}$  at  $v$ :

$$p_{T_v}(s_v | \omega) = \Pr \left( \bigcap_{u \in L_v} X(u) = s(u) \mid X(v) = \omega \right),$$

since  $s_v(u) = s(u)$  for all  $u \in L_v$ . The theorem of total probability gives

$$p_T(s) = \frac{1}{2} \sum_{\omega \in \{0,1\}} p_{T_{v_0}}(s_{v_0} | \omega). \tag{6.1}$$

For an interior vertex  $v$ , if we let  $v_i, i = 1, \dots, m$  be the vertices immediately descended from  $v$  via edges of length  $\ell^i$ , then the transition probabilities in Eq. (2.1)

give

$$p_{T_v}(s_v | \omega) = \prod_{i=1}^m \frac{1}{2} \left( (1 + e^{-\ell^i}) p_{T_{v_i}}(s_{v_i} | \omega) + (1 - e^{-\ell^i}) p_{T_{v_i}}(s_{v_i} | \bar{\omega}) \right) \tag{6.2}$$

where  $\bar{\omega} = 1 - \omega$ . For a leaf  $u$ , we have  $p_{T_u}(s_u | \omega) = p_{T_u}(s(u) | \omega) = 1$  if  $s(u) = \omega$  and zero otherwise. This and Equation (6.2) can be applied recursively to compute the terms  $p_{T_v}(s_v | \omega)$  for each vertex  $v \in T$ , starting at the leaves and working up the tree to the root  $v_0$ . Finally,  $p_T(s)$  can be computed using Equation (6.1), and it follows from the recursion that  $p_T(s)$  is a multivariate polynomial with arguments of the form  $1 + e^{-\ell^k}$  and  $1 - e^{-\ell^k}$ , where  $k$  ranges over the edges of  $T$ . The coefficients of the polynomial depend on the topology of  $T$ .

Equations (6.1) and (6.2) can also be used to differentiate  $p_T(s)$  with respect to any edge length parameter. These derivatives are required in Sect. 3. Suppose  $e$  is an edge of  $T$  and we wish to compute the derivative  $\partial p_T(s) / \partial \ell^e$ . Since we are free to choose  $v_0$ , the calculation is simplified if we let  $v_0$  be an internal vertex at one end of edge  $e$ . We can order the vertices  $v_i$  attached to  $v_0$  so that  $e = (v_0, v_1)$ . Equation (6.2) then gives

$$\begin{aligned} \frac{\partial p_{T_{v_0}}(s_{v_0} | \omega)}{\partial \ell^e} &= \frac{1}{2} e^{-\ell^e} \left( p_{T_{v_1}}(s_{v_1} | \bar{\omega}) - p_{T_{v_1}}(s_{v_1} | \omega) \right) \\ &\quad \times \prod_{i=2}^{\deg(v_0)} \frac{1}{2} \left( (1 + e^{-\ell^i}) p_{T_{v_i}}(s_{v_i} | \omega) + (1 - e^{-\ell^i}) p_{T_{v_i}}(s_{v_i} | \bar{\omega}) \right), \end{aligned}$$

where the  $p_{T_{v_i}}$  terms can be calculated recursively. Second derivatives of the mass function can be calculated analytically in a similar way.

### Appendix B: Proof of Lemma 2.2

First, suppose  $F_1 \sim F_2$ . The BHV boundary rule does not affect the distribution on characters induced by a tree, because the same distribution is obtained whether an edge of length zero is present in a tree or not. Similarly, if  $F_1, F_2$  are equal modulo an application of the boundary rule at infinity, then  $p_{F_1}(s) = p_{F_2}(s)$  since an edge with weight 1 results in independence between the letters at leaves at either side of the edge under the transition probabilities in Eq. (2.1). Specifically, if  $v_0, v_1$  are vertices at the ends of an edge  $e$  with  $\lambda^e = 1$ , then

$$X(v_1) \mid X(v_0) = \omega \sim \text{Bern}(1/2)$$

where  $\omega \in \{0, 1\}$ , so the conditional distribution of  $X(v_1)$  is the same as its marginal. The map  $[F] \mapsto p_F$  from elements of  $\mathcal{W}_N$  to distributions on characters is therefore well-defined. In fact, the work of Allman et al. (2008) shows the map is injective, and this establishes the lemma.



### Appendix C: Proof of Lemma 3.1

The Riemannian metric in Eq. (3.3) can be expanded as

$$\begin{aligned} \delta \ell^i g_{ij}(\boldsymbol{\ell}) \delta \ell^j &= \sum_s p_{\boldsymbol{\ell}}(s) \left( \delta \ell^i \frac{\partial}{\partial \ell^i} \log p_{\boldsymbol{\ell}}(s) \right) \left( \delta \ell^j \frac{\partial}{\partial \ell^j} \log p_{\boldsymbol{\ell}}(s) \right) \\ &= \sum_s p_{\boldsymbol{\ell}}(s) \left( \delta \ell^i \frac{1}{p_{\boldsymbol{\ell}}(s)} \frac{\partial p_{\boldsymbol{\ell}}(s)}{\partial \ell^i} \right) \left( \delta \ell^j \frac{1}{p_{\boldsymbol{\ell}}(s)} \frac{\partial p_{\boldsymbol{\ell}}(s)}{\partial \ell^j} \right) \\ &= \sum_s \frac{1}{p_{\boldsymbol{\ell}}(s)} \left( \delta \ell^i \frac{\partial p_{\boldsymbol{\ell}}(s)}{\partial \ell^i} \right) \left( \delta \ell^j \frac{\partial p_{\boldsymbol{\ell}}(s)}{\partial \ell^j} \right). \end{aligned}$$

The Taylor expansion of  $p_{\boldsymbol{\ell}}(s)$  is

$$p_{\boldsymbol{\ell}+\delta \boldsymbol{\ell}}(s) - p_{\boldsymbol{\ell}}(s) = \sum_i \delta \ell^i \frac{\partial p_{\boldsymbol{\ell}}(s)}{\partial \ell^i} + O(|\delta \boldsymbol{\ell}|^2).$$

Substituting this into the expression for the Riemannian metric gives

$$\begin{aligned} \sum_{i,j} \delta \ell^i g_{ij}(\boldsymbol{\ell}) \delta \ell^j &= \sum_s \frac{(p_{\boldsymbol{\ell}+\delta \boldsymbol{\ell}}(s) - p_{\boldsymbol{\ell}}(s) + O(|\delta \boldsymbol{\ell}|^2))^2}{p_{\boldsymbol{\ell}}(s)} \\ &= \sum_s \frac{(p_{\boldsymbol{\ell}+\delta \boldsymbol{\ell}}(s) - p_{\boldsymbol{\ell}}(s))^2}{p_{\boldsymbol{\ell}}(s)} + O(|\delta \boldsymbol{\ell}|^3) \end{aligned}$$

since  $p_{\boldsymbol{\ell}+\delta \boldsymbol{\ell}}(s) - p_{\boldsymbol{\ell}}(s)$  is  $O(|\delta \boldsymbol{\ell}|)$ . On the other hand, a Taylor expansion of  $f$  around 1 gives

$$\begin{aligned} D_f(p_{\boldsymbol{\ell}+\delta \boldsymbol{\ell}}; p_{\boldsymbol{\ell}}) &= \sum_s p_{\boldsymbol{\ell}}(s) f\left(\frac{p_{\boldsymbol{\ell}+\delta \boldsymbol{\ell}}(s)}{p_{\boldsymbol{\ell}}(s)}\right) \\ &= \sum_s p_{\boldsymbol{\ell}}(s) \left( f(1) + f'(1) \frac{p_{\boldsymbol{\ell}+\delta \boldsymbol{\ell}}(s) - p_{\boldsymbol{\ell}}(s)}{p_{\boldsymbol{\ell}}(s)} \right. \\ &\quad \left. + \frac{1}{2} f''(1) \left( \frac{p_{\boldsymbol{\ell}+\delta \boldsymbol{\ell}}(s) - p_{\boldsymbol{\ell}}(s)}{p_{\boldsymbol{\ell}}(s)} \right)^2 \right. \\ &\quad \left. + O\left( \left| \frac{p_{\boldsymbol{\ell}+\delta \boldsymbol{\ell}}(s) - p_{\boldsymbol{\ell}}(s)}{p_{\boldsymbol{\ell}}(s)} \right|^3 \right) \right) \\ &= f(1) + f'(1) \sum_s p_{\boldsymbol{\ell}+\delta \boldsymbol{\ell}}(s) - f'(1) \sum_s p_{\boldsymbol{\ell}}(s) \\ &\quad + \frac{1}{2} f''(1) \sum_s \frac{(p_{\boldsymbol{\ell}+\delta \boldsymbol{\ell}}(s) - p_{\boldsymbol{\ell}}(s))^2}{p_{\boldsymbol{\ell}}(s)} + O(|\delta \boldsymbol{\ell}|^3). \end{aligned}$$

The first three terms vanish since  $f(1) = 0$  and since  $\sum p_\ell(s) = 1$  for all  $\ell$ . It follows that

$$\sum_{i,j} \delta \ell^i g_{ij}(\ell) \delta \ell^j = \frac{2}{f''(1)} D_f(p_{\ell+\delta\ell}; p_\ell) + O(|\delta\ell|^3)$$

and the lemma is established.

### Appendix D: Proof of Theorem 3.1

We need to show that the induced intrinsic metric  $d^*([F_1], [F_2])$  is finite for any  $[F_1], [F_2] \in \mathcal{W}_N$ . Start by choosing the representative  $F_1 \in W_N$  for the equivalence class  $[F_1]$  to be a connected tree with edge weights  $\lambda_1$ , some elements of which might have value 1. The tree  $F_1$  can be continuously deformed within the orthant corresponding to its topology, to the star tree  $F_*$  on which all pendant edges have weight  $\lambda = 1/2$ , by changing  $\lambda$  along the obvious linear path. If the path has finite length, then the first part of the theorem has been established, since any  $[F_1]$  and  $[F_2]$  can be joined to  $[F_*]$  in this way. As shown in Remark 2.1, each  $p_\lambda(s)$  for  $\lambda$  along the path from  $F_1$  to  $F_*$  is a polynomial in  $\lambda$ . It follows that if  $\lambda$  and  $\lambda + \delta\lambda$  represent the edge weights at two nearby points on the path then

$$p_{\lambda+\delta\lambda}(s) - p_\lambda(s) = \pi_i(s, \lambda) \delta\lambda^i + O(|\delta\lambda|^2)$$

where for each character  $s$  and  $i = 1, \dots, 2N - 3$ ,  $\pi_i(s, \lambda)$  is a polynomial in  $\lambda$ . Then

$$(p_{\lambda+\delta\lambda}(s) - p_\lambda(s))^2 = (\pi_i(s, \lambda) \delta\lambda^i)^2 + O(|\delta\lambda|^3).$$

The path distance between any pair of leaves is continuous along this path, and is strictly positive since the pendant edge lengths are non-zero at all points on the path, apart from potentially at  $F_1$ . Pendant edge lengths can be zero on  $F_1$ , but by the definition of  $W_N$ , the path distance between leaves is non-zero. It follows that  $p_\lambda(s)$  is also bound away from zero. Thus there is a constant  $C(s)$  such that

$$\begin{aligned} \frac{(p_{\lambda+\delta\lambda}(s) - p_\lambda(s))^2}{p_\lambda(s)} &\leq C(s) (\pi_i(s, \lambda) \delta\lambda^i)^2 \\ &\leq C(s) \left( \sum_i \pi_i(s, \lambda)^2 \right) \|\delta\lambda\|^2 \end{aligned}$$

where the second line comes from the Cauchy-Schwarz inequality and the norm is the Euclidean norm. Since the  $\pi_i(s, \lambda)$  are polynomials in  $\lambda$  and the elements of  $\lambda$  lie between 0 and 1, the  $\pi_i(s, \lambda)$  are bounded from above and we obtain

$$\frac{(p_{\lambda+\delta\lambda}(s) - p_\lambda(s))^2}{p_\lambda(s)} \leq B(s) \|\delta\lambda\|^2$$

for some constant  $B(s)$ . Now suppose that  $D_f = d_f^2$  is a  $f$ -divergence, where  $d_f$  is a metric. Applying Eq. (3.6) from Lemma 3.1 with the  $\lambda$ -parametrization gives

$$\begin{aligned} d_f^2(p_{\lambda+\delta\lambda}, p_\lambda) &= \frac{1}{2} f''(1) \sum_s \frac{(p_{\lambda+\delta\lambda}(s) - p_\lambda(s))^2}{p_\lambda(s)} + O(|\delta\lambda|^3) \quad (6.3) \\ &\leq \frac{1}{2} f''(1) \sum_s B(s) \|\delta\lambda\|^2 + O(|\delta\lambda|^3) \\ &\leq K \|\delta\lambda\|^2 \end{aligned}$$

for some constant  $K$ . Thus the infinitesimal path length in  $\mathcal{W}_N$  as measured by the metric  $d$  is bounded by some multiple of the Euclidean path length on  $\lambda$ . The length of the linear path from  $F_1$  to  $F_*$  measured with  $d$  is therefore finite, since the Euclidean length of this path is finite, and hence  $d^*([F_1], [F_2])$  is finite.

For the second part of the theorem, suppose that  $D_{f_0} = d_{f_0}^2$  is a  $f_0$ -divergence, where  $d_{f_0}$  is a metric. Further, suppose that  $F_1$  and  $F_2$  are given by  $\lambda_1$  and  $\lambda_2$ , respectively. It suffices to consider  $\lambda_1, \lambda_2$  from the same topology and sufficiently close such that the image  $t \mapsto \lambda(t)$ ,  $\lambda(0) = \lambda_1$  and  $\lambda(1) = \lambda_2$  of the geodesic from  $p_{\lambda_1}$  to  $p_{\lambda_2}$  in the metric induced by the Riemannian metric  $g_{ij}$  lies fully in a convex  $\lambda$  coordinate patch and has finite Euclidean length there, say  $L$ . Hence, for every  $n \in \mathbb{N}$ , there are  $\delta\lambda_{(j)}$  with  $\|\delta\lambda_{(j)}\| \leq L/n$ ,  $j \in \{1, \dots, n\}$  such that  $\lambda_2 = \lambda_1 + \sum_{j=1}^n \delta\lambda_{(j)}$ . Then the second assertion of the theorem follows from (6.3), setting  $c = f''(1)/f_0''(1)$ , as  $n \rightarrow \infty$ , because

$$\left| d_f(p_{\lambda_1}, p_{\lambda_2})^2 - c \cdot d_{f_0}(p_{\lambda_1}, p_{\lambda_2})^2 \right| = \left| \sum_{j=1}^{n-1} O(|\delta\lambda_{(j)}|^3) \right| = O\left(\frac{L}{n^2}\right).$$

The second equality sign holds because the constants in the individual summands  $O(|\delta\lambda_{(j)}|^3)$  ( $1 \leq j \leq n$ ) can be bounded by the supremum of absolute values of the gradient of  $p_\lambda$  with respect to  $\lambda$ , in the coordinate patch, as can be seen from the last lines of the proof of Lemma 3.1.

The third part of the theorem, which states that minimal length paths satisfy the geodesic equation locally, is part of the standard theory for Riemannian geometry on manifolds, e.g. (Lee 1997, Sect. 4).

### Appendix E: Proof of Theorem 4.1

The theorem is trivial for  $N = 2$ , so suppose  $F \in W_N$  with  $N \geq 3$  and that the first assertion holds for all  $G \in W_{N-1}$ . The matrix  $S_F$  is not changed by inserting edges  $e$  with  $\lambda^e = 1$  to connect trees in  $F$ , or by adding edges with  $\lambda^e = 0$ , so without loss of generality we may assume  $F$  is a fully resolved tree. We may also assume there is a cherry between leaves  $N - 1$  and  $N$  since each bifurcating tree with  $N \geq 3$  has a

cherry and permuting the labels of  $F$  results in a tree with covariance  $P^T S_F P$  (with permutation matrix  $P$ ), where positive definiteness is preserved.

Let  $e_{N-1}$  and  $e_N$  be the edges incident to leaves  $N - 1$  and  $N$ , respectively. Since  $F$  is bifurcating, there is exactly one edge, say  $e_0$ , incident to  $e_{N-1}$  and  $e_N$ . Let  $S_F = (s_{uv})_{u,v=1}^N$ . The tree  $G \in W_{N-1}$  obtained by deleting  $e_N$  and leaf  $N$  and merging  $e_0$  and  $e_{N-1}$  to  $\tilde{e}$  with weight  $\lambda_{\tilde{e}} = 1 - (1 - \lambda_{e_0})(1 - \lambda_{e_{N-1}})$  has covariance  $S_G = (s_{uv})_{u,v=1}^{N-1}$ , which is by induction positive definite. Using this and Sylvester’s criterion (that a matrix is positive if and only if all principal minors have positive determinant) it suffices to show  $\det(S_F) > 0$ . We have for all  $1 \leq u \leq N - 1$  that  $s_{uN} = s_{Nu} = (1 - \lambda_{e_N})c_u$ , where

$$c_u = \begin{cases} 1 - \lambda_{e_{N-1}} & \text{when } u = N - 1, \text{ and} \\ \prod_{e \neq e_N, e_{N-1}} (1 - \lambda^e)^{\sigma_{Nu}^e} & \text{for } 1 \leq u \leq N - 2 \end{cases}$$

and  $\sigma_{Nu}^e$  is defined by Eq. (4.4). Note that for  $u, v \leq N - 1$ ,  $c_u$  and  $s_{uv}$  do not involve  $\lambda_{e_N}$ , and that  $s_{NN} = 1$ . If  $\mathfrak{S}_N$  denotes the set of permutations of  $\{1, \dots, N\}$ , then the Leibniz formula for determinants gives

$$\begin{aligned} \det(S_F) &= \left( \sum_{\substack{\tau \in \mathfrak{S}_N \\ \tau(N)=N}} \text{sgn}(\tau) \prod_{u=1}^N s_{u\tau(u)} \right) + \left( \sum_{\substack{\tau \in \mathfrak{S}_N \\ \tau(N) \neq N}} \text{sgn}(\tau) \prod_{u=1}^N s_{u\tau(u)} \right) \\ &= \det\left((s_{uv})_{u,v=1}^{N-1}\right) + (1 - \lambda_{e_N})^2 \left( \sum_{\substack{\tau \in \mathfrak{S}_N \\ \tau(N) \neq N}} \text{sgn}(\tau) c_{\tau(N)} c_{\tau^{-1}(N)} \prod_{\substack{u=1 \\ u \neq \tau^{-1}(N)} }^{N-1} s_{u\tau(u)} \right), \end{aligned}$$

so  $\det(S_F)$  is linear in  $x := (1 - \lambda_{e_N})^2$ . By symmetry of the cherry,  $\det(S_F)$  is also linear in  $y := (1 - \lambda_{e_{N-1}})^2$  as well. We write  $g(x, y) = \det(S_F)$ . For  $x = 0$ , we have  $s_{Nu} = 0$  for  $u < N$  and  $s_{NN} = 1$ , so  $g(0, y) = \det(S_F) = \det(S_G) > 0$  for all  $y \in [0, 1]$ , and similarly  $g(x, 0) > 0$  for all  $x \in [0, 1]$ . Furthermore,  $g(1, 1) = 0$ , since in that case the last two rows of  $S_F$  coincide. Since  $g$  is linear in  $x$  and in  $y$ , respectively, we have  $g(x, y) > 0$  for all  $(x, y) \in [0, 1]^2 \setminus \{(1, 1)\}$ , so that  $\det(S_F) > 0$  for all  $(\lambda_{e_{N-1}}, \lambda_{e_N}) \in [0, 1]^2 \setminus \{(0, 0)\}$ . If  $\lambda_{e_{N-1}} = \lambda_{e_N} = 0$ , we would have  $d_{N(N-1)} = 0$ , but this is not allowed by the definition of  $W_N$ .

We also need to show that the map  $[F] \mapsto S_F$  is injective on  $\mathcal{W}_N$  where  $[F]$  denotes the equivalence class of  $F \in W_N$ . This is trivial, however, since whenever  $F_1, F_2 \in W_N$  are in different equivalence classes, the matrix of distances between the leaves is different.

### Appendix F: Proof of Theorem 5.1

The proof is similar to that for Theorem 3.1, and so we give a brief sketch. We consider the same path between the trees  $[F_1], [F_*] \in \mathcal{W}_N$ , and show that each element of  $g_{ij}(\lambda)$

is bound from above along the path. Working in the  $\lambda$ -parametrization of an orthant, Eq. (4.5) becomes

$$S_\lambda = \left( \prod_e (1 - \lambda^e)^{\sigma_{uv}^e} \right)_{u,v=1}^N .$$

Each element of the matrix is therefore a polynomial in the elements of  $\lambda$ , and their derivatives with respect to  $\lambda$  are also polynomials. Recalling that the tangent space of  $\mathcal{W}_N$  at  $S_\lambda$  in a maximal orthant is spanned by  $\partial_i S_\lambda$ , where  $i \in \{1, \dots, 2N - 3\}$  ranges over the edges in that maximal orthant, Eq. (4.6) becomes

$$g_{ij}(\lambda) = \frac{1}{2} \operatorname{tr} \left( S_\lambda^{-1} (\partial_i S_\lambda) S_\lambda^{-1} (\partial_j S_\lambda) \right),$$

$i, j \in \{1, \dots, 2N - 3\}$ .

Applying the Cauchy-Schwartz inequality  $|\operatorname{tr}(A^T B)|^2 \leq \operatorname{tr}(A^T A) \operatorname{tr}(B^T B)$  gives

$$\begin{aligned} |g_{ij}(\lambda)|^2 &\leq \frac{1}{4} \operatorname{tr} \left( (\partial_i S_\lambda)^2 S_\lambda^{-2} \right) \operatorname{tr} \left( (\partial_j S_\lambda)^2 S_\lambda^{-2} \right) \\ &\leq \frac{1}{4} \operatorname{tr} \left( S_\lambda^{-4} \right) \operatorname{tr} \left( (\partial_i S_\lambda)^4 \right)^{\frac{1}{2}} \operatorname{tr} \left( (\partial_j S_\lambda)^4 \right)^{\frac{1}{2}} . \end{aligned}$$


The first term in this product is bounded on a geodesic path from  $F_1$  to  $F_*$ , since  $S_\lambda$  is positive definite and its eigenvalues are bound away from zero. The other two terms are also bounded from above, because the derivatives of  $S_\lambda$  are polynomials in  $\lambda$ . Thus  $|g_{ij}(\lambda)| \leq C$  for some constant  $C$  at all points along that path, and the same argument as for Theorem 3.1 shows that  $d_{\operatorname{cov}}^*([F_1], [F_*])$  is finite.

## References

- Adams RH, Castoe TA (2020) Probabilistic species tree distances: implementing the multispecies coalescent to compare species trees within the same model-based framework used to estimate them. *Syst Biol* 69(1):194–207
- Allen BL, Steel M (2001) Subtree transfer operations and their induced metrics on evolutionary trees. *Ann Comb* 5(1):1–15
- Allman ES, Ané C, Rhodes JA (2008) Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. *Adv Appl Probab* 40(1):229–249
- Ballmann W, Gromov M, Schroeder V (1985) Manifolds of nonpositive curvature. *Progress in mathematics*, vol 61. Birkhäuser, Basel
- Bačák M (2014) Computing medians and means in Hadamard spaces. *SIAM J Optim* 24(3):1542–1566
- Billera L, Holmes S, Vogtman K (2001) Geometry of the space of phylogenetic trees. *Adv Appl Math* 27:733–767
- Bridson MR, Haefliger A (2011) *Metric spaces of non-positive curvature*. Springer, Berlin
- Bryant D, Galtier N, Poursat M-A (2005) Likelihood calculation in molecular phylogenetics. In: Gascuel O (ed) *Mathematics of evolution and phylogeny*. Oxford University Press, Oxford, pp 33–62
- Dryden IL, Koloydenko A, Zhou D et al (2009) Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann Appl Stat* 3(3):1102–1123
- Engström A, Hersh P, Sturmfels B (2013) Toric cubes. *Rendiconti del Circolo Matematico di Palermo* 62(1):67–78

- Feragen A, Owen M, Petersen J, Wille M, Thomsen L, Dirksen A, de Bruijne M (2013) Tree-space statistics and approximations for large-scale analysis of anatomical trees. In: 23rd biennial international conference on information processing in medical imaging (IPMI)
- Garba MK (2019) Information geometry for phylogenetic trees. Ph.D. thesis, School of Mathematics, Statistics and Physics, Newcastle University
- Garba MK, Nye TMW, Boys RJ (2018) Probabilistic distances between trees. *Syst Biol* 67(2):320–327
- Gill J, Linusson S, Moulton V, Steel M (2008) A regular decomposition of the edge-product space of phylogenetic trees. *Adv Appl Math* 41(2):158–176
- Hansen TF, Martins EP (1996) Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution* 50(4):1404–1417
- Huelsenbeck JP, Ronquist F (2001) MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8):754–755
- Kim J (2000) Slicing hyperdimensional oranges: the geometry of phylogenetic estimation. *Mol Phylogenet Evol* 17(1):58–75
- Lee JM (1997) Riemannian manifolds: an introduction to curvature, vol 176. Springer, Berlin
- Lenglet C, Rousson M, Deriche R, Faugeras O (2006) Statistics on the manifold of multivariate normal distributions: theory and application to diffusion tensor MRI processing. *J Math Imaging Vis* 25(3):423–444
- Lin B, Yoshida R (2018) Tropical Fermat-Weber points. *SIAM J Discrete Math* 32(2):1229–1245
- Lin B, Monod A, Yoshida R (2018) Tropical foundations for probability and statistics on phylogenetic tree space. arXiv preprint [arXiv:1805.12400](https://arxiv.org/abs/1805.12400)
- Miller E, Owen M, Provan JS (2015) Polyhedral computational geometry for averaging metric phylogenetic trees. *Adv Appl Math* 68:51–91
- Moakher M (2005) A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM J Matrix Anal Appl* 26:735–747
- Moulton V, Steel M (2004) Peeling phylogenetic oranges. *Adv Appl Math* 33(4):710–727
- Nye TMW (2011) Principal components analysis in the space of phylogenetic trees. *Ann Stat* 39(5):2716–2739
- Nye T (2014) An algorithm for constructing principal geodesics in phylogenetic treespace. *IEEE ACM Trans Comput Biol* 11(2):304–315
- Nye TMW, Tang X, Weyenberg G, Yoshida R (2017) Principal component analysis and the locus of the Fréchet mean in the space of phylogenetic trees. *Biometrika* 104(4):901–922
- Owen M, Provan JS (2011) A fast algorithm for computing geodesic distances in tree space. *IEEE ACM Trans Comput Biol* 8(1):2–13
- Rogers JS (1997) On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Syst Biol* 46(2):354–357
- Sason I, Verdu S (2016)  $f$ -divergence inequalities. *IEEE Trans Inf Theory* 62(11):5973–6006
- Schmidt FR, Clausen M, Cremers D (2006) Shape matching by variational computation of geodesics on a manifold. In: Joint pattern recognition symposium. Springer, pp 142–151
- Semple C, Steel M (2003) Phylogenetics. Oxford lecture series in mathematics and its applications, vol 24. Oxford University Press, Oxford
- Skovgaard LT (1984) A Riemannian geometry of the multivariate normal model. *Scand J Stat* 11(4):211–223
- Speyer D, Sturmfels B (2004) The tropical Grassmannian. *Adv Geom* 4(3):389–411
- Steel MA, Penny D (1993) Distributions of tree comparison metrics—some new results. *Syst Biol* 42(2):126–141
- Willis A (2019) Confidence sets for phylogenetic trees. *J Am Stat Assoc* 114(525):235–244
- Yang Z (2006) Computational molecular evolution. Oxford University Press, Oxford
- Yoshida R, Zhang L, Zhang X (2019) Tropical principal component analysis and its application to phylogenetics. *Bull Math Biol* 81(2):568–597
- Zwiernik P, Smith JQ (2012) Tree cumulants and the geometry of binary tree models. *Bernoulli* 18(1):290–321

## Affiliations

**M. K. Garba**<sup>1,2</sup> · **T. M. W. Nye**<sup>1</sup> · **J. Lueg**<sup>3</sup> · **S. F. Huckemann**<sup>3</sup> 

M. K. Garba  
m.k.garba1@ncl.ac.uk

T. M. W. Nye  
tom.nye@ncl.ac.uk

J. Lueg  
jonas.lueg@stud.uni-goettingen.de; lueg@math.uni-goettingen.de

- <sup>1</sup> School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, UK
- <sup>2</sup> Department of Mathematical Sciences, Bayero University, Kano, Nigeria
- <sup>3</sup> Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences, Georg-August-Universität, Göttingen, Germany