Limiting behaviour of Fréchet means in the space of phylogenetic trees

D. Barden[1] · H. Le[2] · M. Owen[3]

Abstract As demonstrated in our previous work on T_4, the space of phylogenetic trees with four leaves, the topological structure of the space plays an important role in the non-classical limiting behaviour of the sample Fréchet means in T_4. Nevertheless, the techniques used in that paper cannot be adapted to analyse Fréchet means in the space T_m of phylogenetic trees with $m(\geqslant 5)$ leaves. To investigate the latter, this paper first studies the log map of T_m. Then, in terms of a modified version of this map, we characterise Fréchet means in T_m that lie in top-dimensional or co-dimension one strata. We derive the limiting distributions for the corresponding sample Fréchet means, generalising our previous results. In particular, the results show that, although they are related to the Gaussian distribution, the forms taken by the limiting distributions depend on the co-dimensions of the strata in which the Fréchet means lie.

Keywords Central limit theorem · Fréchet mean · Log map · Phylogenetic trees · Stratified manifold

✉ H. Le
  huiling.le@nottingham.ac.uk

  D. Barden
  d.barden@dpmms.cam.ac.uk

  M. Owen
  megan.owen@lehman.cuny.edu

[1] Girton College, University of Cambridge, Cambridge CB3 0JG, UK

[2] School of Mathematical Sciences, University of Nottingham, Nottingham NG7 2RD, UK

[3] Department of Mathematics and Computer Science, Lehman College, City University of New York, Bronx 10468, USA

1 Introduction

The concept of Fréchet means of random variables on a metric space is a generalisation of the least mean-square characterisation of Euclidean means: a point is a Fréchet mean of a probability measure μ on a metric space (M, d) if it minimises the Fréchet function for μ defined by

$$x \mapsto \frac{1}{2} \int_M d(x, x')^2 \mathrm{d}\mu(x'),$$

provided the integral on the right side is finite for at least one point x. Note that the factor $1/2$ will simplify some later computations. The concept of Fréchet means has recently been used in the statistical analysis of data of a non-Euclidean nature. We refer readers to Bhattacharya and Patrangenaru (2005), Bhattacharya and Patrangenaru (2014), Dryden et al. (2014), Dryden and Mardia (1998) and Kendall and Le (2011), as well as the references therein, for the relevance of, and recent developments in, the study of various aspects of Fréchet means in Riemannian manifolds. The Fréchet mean has also been studied in the space of phylogenetic trees, as motivated by Billera et al. (2001) and Holmes (2003). It was first introduced to this space independently by Bacak (2014) and Miller et al. (2015), which both gave methods for computing it. Limiting distributions of sample Fréchet means in the space of phylogenetic trees with four leaves were studied in Barden et al. (2013), and it was used to analyse tree-shaped medical imaging data in Feragen et al. (2013), while principal geodesic analysis on the space of phylogenetic trees, a related statistical issue, was studied in Nye (2011), Nye (2014), and Feragen et al. (2013).

A phylogenetic tree represents the evolutionary history of a set of organisms and is an important concept in evolutionary biology. Such a tree is a contractible graph, that is, a connected graph with no circuits, where one of its vertices of degree 1 is distinguished as the root of the tree and the other such vertices are (labelled) leaves. The space T_m of phylogenetic trees with m leaves was first introduced in Billera et al. (2001). The important feature of the space is that each point represents a tree with a particular structure and specified lengths of its edges in such a way that both the structure and the edge lengths vary continuously in a natural way throughout the space. The space is constructed by identifying faces of a disjoint union of Euclidean orthants, each corresponding to a different tree structure. In particular, it is a topologically stratified space and also a $CAT(0)$, i.e. globally non-positively curved, space (cf. Bridson and Haefliger 1999). A detailed account of the underlying geometry of tree spaces can be found in Billera et al. (2001) and a brief summary can be found in the Appendix to Barden et al. (2013).

As demonstrated in Basrak (2010) and Hotz et al. (2013) for T_3 and in Barden et al. (2013) for T_4, the global, as well as the local, topological structure of the space of phylogenetic trees plays an important role in the limiting behaviour of sample Fréchet means. These results imply that the known results (cf. Kendall and Le 2011) on the limiting behaviour of sample Fréchet means in Riemannian manifolds cannot be applied directly. Moreover, due to the increasing complexity of the structure of T_m as m increases, the techniques used in Barden et al. (2013) for T_4 could not be

adapted to derive the limiting behaviour of sample Fréchet means in T_m for general m. For example, although the natural isometric embedding of T_4 is in 10-dimensional Euclidean space $\mathbb{R}^{10}$, it is intrinsically 2-dimensional, being constructed from 15 quadrants identified three at a time along their common axes. This made it possible in Barden et al. (2013), following Billera et al. (2001), to represent T_4 as a union of certain quadrants embedded in $\mathbb{R}^3$ in such a way that it was possible to visualise the geodesics explicitly. That is, naturally, not possible for $m > 4$. The need to describe geodesics explicitly arises as follows. In a complete manifold of non-positive curvature, the global minimum of a Fréchet function would be characterised by the vanishing of its derivative. In tree space, as in general stratified spaces, such derivatives do not exist at non-manifold points. However, directional derivatives for a Fréchet function, which serve our purpose, do exist at all points and for all tangential directions. They are defined via the log map, which is a generalisation of the inverse of the exponential map of Riemannian manifolds, and is expressed in terms of the lengths and initial tangent vectors of unit speed geodesics.

In this paper, we derive the expression for the log map using the geometric structure of geodesics in T_m obtained in Owen (2011) and Owen and Provan (2011). As a result, we are able to establish a central limit theorem for *iid* random variables having probability measure μ that has its Fréchet mean lying in a top-dimensional stratum. In this case, our central limit theorem shows that sample means behave in a predictable way, similar to the case for arbitrary Euclidean data although with a more complex covariance structure. In particular, our result says that, as the size of a sample increases, the mean of that sample approaches the true mean of the underlying distribution in a quantifiable way. This means that we can use the mean of sample data as an approximation of the true mean as we would do with Euclidean data. This opens the door for work on confidence intervals, such as Willis (2016). Moreover, we take advantage of the special structure of tree space in the neighbourhood of a stratum of co-dimension one to obtain the corresponding results when the Fréchet mean of μ lies in such a stratum. In particular, we show that, in this case, the limiting distribution can take one of the three possible forms, distinguished by the nature of its support. Unlike the Euclidean case, the limiting distributions in both cases here are expressed in terms of the log map at the Fréchet mean of μ. This is similar to the central limit theorem for sample Fréchet means on Riemannian manifolds (cf. Kendall and Le 2011). Although it may appear non-intuitive, it allows us to use the standard results on Euclidean space. For example, in the top-dimensional case, the limiting distribution is a Gaussian distribution and so some classical hypothesis tests can be carried out in a similar fashion to hypothesis tests for data lying in a Riemannian manifold as demonstrated in Dryden and Mardia (1998) for the statistical analysis of shape. However, in the case of co-dimension one, the limiting distribution is non-standard and so the classical hypothesis tests cannot easily be modified to apply. Further investigation is required and we aim to pursue this, as well as the applications of the results to phylogenetic trees, in future papers.

The remainder of the paper is organised as follows. To obtain the directional derivatives of a Fréchet function, we need an explicit expression for the log map that is amenable to calculation. This in turn requires a detailed analysis of the geodesics which we carry out in the next section using results from Owen (2011) and

Owen and Provan (2011). The resulting expression (6) for the log map in Theorem 1 and its modification (9) are then used in the following two sections which study the limiting distributions for sample Fréchet means in T_m; Sect. 3 concentrates on the case when the Fréchet means lie in the top-dimensional strata, while Sect. 4 deals with the case when they lie in the strata of co-dimension one. In the final section, we discuss some of the problems involved in generalising our results to the case that the Fréchet means lie in strata of arbitrary co-dimension.

2 The log map on a top-dimensional stratum

The log map is the generalisation of $\exp^{-1}$, the inverse of the exponential map on a Riemannian manifold. For a tree T^* in T_m the log map, $\log_{T^*}$, at T^* takes the form

$$\log_{T^*}(T) = d(T^*, T)\, \mathbf{v}(T) \tag{1}$$

as T varies, where $\mathbf{v}(T)$ is a unit vector at T^* along the geodesic from T^* to T and $d(T^*, T)$ is the distance between T^* and T along that geodesic. This is well defined since T_m is a globally non-positively curved space, or $CAT(0)$-space (cf. Bridson and Haefliger 1999), and so this geodesic is unique.

Note that for data in Euclidean space, since tangent vectors at different points may be identified by parallel translation. $\log_{T^*}(T)$ would be represented by the difference $T - T^*$ of the two position vectors. Then, $\log_T(T^*)$ would be its negative $T^* - T$. However, although for convenience we shall embed T_m in a Euclidean space, that embedding is not unique and there is no canonical relation between $\log_{T^*}(T)$, which is a vector tangent to T_m at T^*, and $\log_T(T^*)$ which is tangent at T. The log map can be thought of as a projection from the more complicated tree space onto a simpler Euclidean space, or more generally onto a Euclidean cone, that has minimal distortion around T^*, with the amount of possible distortion increasing as trees get further from T^*. Specifically, the log map is a bijection for trees in the orthant of T^*. For trees outside of this orthant, multiple trees can be mapped to the same Euclidean point, including two trees with no edges in common. Thus, we generally cannot use information about the position of a point on the tangent space to say something about the inverse tree, and thus the log map is primarily a mathematical tool rather than something that has biological meaning.

To analyse this log map further, we first recall some relevant aspects of the structure of trees and tree spaces. Apart from the roots and leaves of a tree, which are the vertices of degree 1 mentioned above, there are no vertices of degree two and the remaining vertices, of degree at least 3, are called internal. An edge is called internal if both its vertices are. A tree with m labelled leaves and unspecified internal edge lengths determines a combinatorial type. Then, T_m is a stratified space with a stratum for each such type: a given type with k ($\leqslant m - 2$) internal edges determines a stratum with k positive parameters ranging over the points of an open k-dimensional Euclidean orthant, each point representing the tree with those specific parameters as the lengths of its internal edges. Note that, for this paper, we shall only consider the internal edges
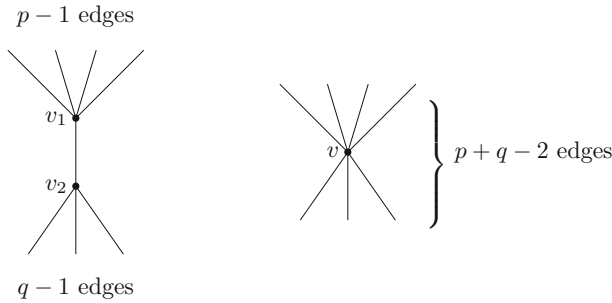
Fig. 1 The edge between vertices v_1 and v_2 shrinks to 0 to form a vertex of degree $p + q + 2$

of a tree. So by 'edge' we always mean 'internal edge' and, to simplify the notation, we consider T_{m+2}, rather than T_m.

The metric on T_{m+2} is induced by regarding the identification of a stratum τ with a Euclidean orthant $\mathcal{O}$ as an isometry. Then each face, or boundary orthant of co-dimension one, of $\mathcal{O}$ is identified with a boundary stratum σ of τ. A tree of type σ is obtained from a tree of type τ by coalescing the vertices v_1 and v_2 of degree p and q of the edge whose parameter has become zero, to form a new vertex v of degree $p + q - 2$. See Fig. 1.

We are particularly interested in the top-dimensional strata. These are formed by binary trees, in which all internal vertices have degree 3. A binary tree with $m + 2$ leaves has $m + 1$ internal vertices and m internal edges so that the corresponding stratum has dimension m. There are $(2m + 1)!!$ such strata in T_{m+2} (cf. Schroder 1870). For these strata, the boundary relation results in two adjacent vertices of degree 3 coalescing to form a vertex of degree 4. Since each vertex of degree 4 can be formed 3 different ways, each stratum of co-dimension one is a component of the boundary of three different top-dimensional strata. Figure 2 shows an example of these strata in T_4.

If a tree T^* lies in a top-dimensional stratum of T_{m+2}, since such a stratum can be identified with an orthant $\mathcal{O}$ in $\mathbb{R}^m$, we may identify the tangent space to T_{m+2} at T^* with $\mathbb{R}^m$. Then, for each point $T \in T_{m+2}$, the geodesic from T^* to T in T_{m+2} will start with a linear segment in $\mathcal{O}$, which determines an initial *unit* tangent vector $\boldsymbol{v}(T) \in \mathbb{R}^m$ at T^*. Thus, we may identify the image of the log map defined in (1) as the vector $d(T^*, T)\boldsymbol{v}(T)$ in $\mathbb{R}^m$.

For example, the space T_3 of trees with three leaves is the 'spider': three half Euclidean lines joined at their origins. Denoting the length of the edge e of T by $|e|_T$, then $d(T^*, T) = ||e|_{T^*} - |e|_T|$, if T^* and T lie in the same orthant of T_3, and $d(T^*, T) = |e|_{T^*} + |e|_T$, otherwise. Thus, the log map for T_3 can be expressed explicitly as:

$$\log_{T^*}(T) = \begin{cases} (|e|_T - |e|_{T^*})\boldsymbol{e} & \text{if } T \text{ and } T^* \text{ are in the same orthant;} \\ -(|e|_T + |e|_{T^*})\boldsymbol{e} & \text{otherwise,} \end{cases}$$

Fig. 2 Three adjacent
top-dimensional strata in T_4 and
their shared co-dimension one
stratum. A sample tree is shown
for each stratum, and the axes
are labelled by the
corresponding edge-type
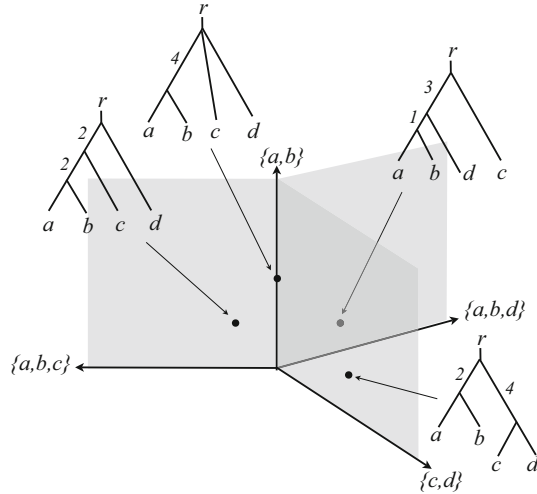
where e is the canonical unit vector determining the orthant in which T^* lies. Note
that we abuse notation by calling the (single) internal edge e in all 3 trees, despite
these edges dividing the leaves in different ways. The explicit expression for the log
map for the space T_4 of trees with four leaves is already much more complicated than
this and was derived in Barden et al. (2013).

To obtain the expression for the log map at T^* for the space T_{m+2} of trees with
$m + 2$ ($m > 2$) leaves, we first summarise without proofs the description, given in
Billera et al. (2001), Owen (2011), Owen and Provan (2011) and Vogtmann (2007),
of the geodesic between two given trees in T_{m+2}.

When an (internal) edge is removed from a tree, it splits the set of the leaves plus
the root into two disjoint subsets, each having at least two members, and we identify
the edges from different trees that induce the same split. Each edge has a 'type' that
is specified by the subset of the corresponding split that does not contain the root. For
example, in the tree in Fig. 3a, the edge labelled x_3 has the edge-type $\{a, b\}$, while the
edge labelled x_1 has the edge-type $\{a, b, c, d\}$. There are

$$M = 2^{m+2} - m - 4 \tag{2}$$

possible edge-types. Two edge-types are called *compatible* if they can occur in the same
tree, and T_{m+2} may be identified with a certain subset of $\mathbb{R}^M$, each possible edge-type
being identified with a positive semi-axis in $\mathbb{R}^M$. To make this identification explicit,
we choose a canonical order of the edges by first ordering the leaves and then taking
the induced lexicographic ordering of the sets of (ordered) leaves that determine the
edges. Then, if Σ is a set of mutually compatible edge-types and $\mathcal{O}(\Sigma)$ is the orthant
spanned by the corresponding semi-axes in $\mathbb{R}^M$, each point of $\mathcal{O}(\Sigma)$ represents a
tree with the combinatorial type determined by Σ and T_{m+2} is the union of all such
orthants.

Fig. 3 Trees, carrier, and isometric embedding for Example 1. **a** Tree T^*. **b** Tree T. **c** The geodesic between the trees corresponding to $\boldsymbol{u}^*$ and $\boldsymbol{u}$ is marked with the *dashed line*. The $-x_1, -x_2, x_3$ octant does not exist in tree space, but the $-x_2$, x_3 quadrant does, so the geodesic is restricted to lying in the *grey* area. It bends at the points p and q. **d** The isometric embedding of the *grey* area in **c** into V^2. Intuitively, this corresponds to "unfolding" the bends. $\boldsymbol{u}^*$ and $\boldsymbol{u}$ are mapped to $\boldsymbol{v}^*$ and $\boldsymbol{v}$. The Euclidean geodesic between $\boldsymbol{v}^*$ and $\boldsymbol{v}$ in V^2 is contained in the *grey* area, and thus can be mapped back onto the geodesic in tree space
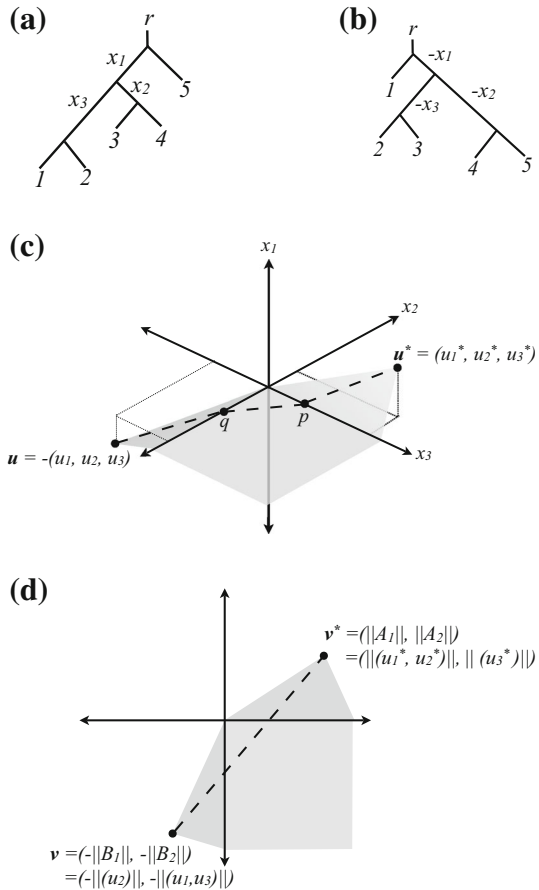
For a set of edges A in a tree T, define $\|A\|_T = \sqrt{\sum_{e \in A} |e|_T^2}$ and write $|A|$ for the number of edges in A. For two given trees T^* and T, let E^* and E be their respective edge sets, or sets of non-trivial splits. Assume first that T^* and T have no common edge, i.e. $E^* \cap E = \emptyset$. Then, the geodesic from T^* to T can be determined as follows.

Lemma 1 *Let T^* and T be two trees with no common edges, lying in top-dimensional strata of $\boldsymbol{T}_{m+2}$. Then, there is an integer k, $1 \leqslant k \leqslant m$, and a pair $(\mathcal{A}, \mathcal{B})$ of partitions $\mathcal{A} = (A_1, \ldots, A_k)$ of E^* and $\mathcal{B} = (B_1, \ldots, B_k)$ of E, all subsets A_i and B_j being non-empty, such that*

(P1) *for each $i > j$, the union $A_i \cup B_j$ is a set of mutually compatible edges;*

(P2) $\dfrac{\|A_1\|_{T^*}}{\|B_1\|_T} \leqslant \dfrac{\|A_2\|_{T^*}}{\|B_2\|_T} \leqslant \cdots \leqslant \dfrac{\|A_k\|_{T^*}}{\|B_k\|_T};$

(P3) *for all (A_i, B_i), there are no non-trivial partitions $C_1 \cup C_2$ of A_i and $D_1 \cup D_2$ of B_i such that $C_2 \cup D_1$ is a set of mutually compatible edges and $\dfrac{\|C_1\|_{T^*}}{\|D_1\|_T} < \dfrac{\|C_2\|_{T^*}}{\|D_2\|_T}.$*

The geodesic is the shortest path through the sequence of orthants $\mathcal{C} = (\mathcal{O}_0, \ldots, \mathcal{O}_k)$ *where*

$$\mathcal{O}_i = \mathcal{O}(B_1 \cup \cdots \cup B_i \cup A_{i+1} \cup \cdots \cup A_k) \tag{3}$$

and has length $\|(\|A_1\|_{T^*} + \|B_1\|_T, \|A_2\|_{T^*} + \|B_2\|_T, \ldots, \|A_k\|_{T^*} + \|B_k\|_T)\|$.

Note that (3) implies that $\mathcal{O}_0 = \mathcal{O}(E^*)$ is the orthant in which T^* lies and that T is in $\mathcal{O}_k$. These results, developed from Vogtmann (2007), are given in this form, though not in a single lemma, in Owen and Provan (2011, section 2.3), where the properties (P1), (P2) and (P3) are stated in identical terms. The edge set for $\mathcal{O}_i$ is denoted by $\mathcal{E}^i$ in the statement of Theorem 2.4 there and the formula for the length of the geodesic is equation (1) in that statement.

Following Vogtmann (2007) and Owen and Provan (2011), respectively, we call the orthant sequence $\mathcal{C}$ the *carrier* of the geodesic, and the pair of partitions $(\mathcal{A}, \mathcal{B})$ the *support* of the geodesic. In general, the integer k and the support $(\mathcal{A}, \mathcal{B})$ need not be unique. However, they are unique if all the inequalities in (P2) are strict (Owen and Provan 2011, Remark, p.7) and, in this case, we shall refer to the carrier and support as the *minimal* ones.

Under the above assumption, the integer k appearing in Lemma 1 is the number of times that the geodesic includes a segment in the interior of one orthant followed by a segment in the interior of a neighbouring orthant. Hence, the constraints $1 \leqslant k \leqslant m$: $k = 1$ implies that the geodesic goes through the cone point and $k = m$ that it passes through a sequence of top-dimensional orthants.

We can now give an isometric embedding $\tilde{\mathcal{C}}$ in $\mathbb{R}^m$ of $\mathcal{C} \subseteq \mathbb{R}^M$ with T^* mapped to $\boldsymbol{u}^* = (u_1^*, \ldots, u_m^*)$ in the positive orthant, where the $u_i^* > 0$ represent the lengths of the edges of T^*, and with T mapped to $\boldsymbol{u} = -(u_1, \ldots, u_m)$ in the negative orthant, where the $u_i > 0$ are the lengths of the edges of T. Let $(t_1^*, \ldots, t_m^*)$ be the coordinates of T^* ordered by the canonical ordering given just before Lemma 1 that embeds $\boldsymbol{T}_{m+2}$ in $\mathbb{R}^M$. Then, we can reorder the coordinates u_i^* such that the edges in A_1 correspond to the first $|A_1|$ positive semi-axes in $\mathbb{R}^m$, the edges in A_2 correspond to the next $|A_2|$ positive semi-axes in $\mathbb{R}^m$, etc., while the edges in B_1 correspond to the first $|B_1|$ negative semi-axes in $\mathbb{R}^m$, the edges in B_2 correspond to the next $|B_2|$ negative semi-axes in $\mathbb{R}^m$, etc. By (P1), the edge sets $B_1, \ldots, B_i, A_{i+1}, \ldots, A_k$ are mutually compatible for all $0 \leqslant i \leqslant k$, implying that the images of these edges in $\mathbb{R}^m$ are mutually orthogonal, and so they determine an isometric embedding of $\mathcal{O}_i$, defined by (3), and hence the required isometric embedding $\tilde{\mathcal{C}}$ of $\mathcal{C}$. Let π be the inverse of the permutation of the coordinates described above, so that

$$\pi : \boldsymbol{u}^* = (u_1^*, \ldots, u_m^*) \mapsto \boldsymbol{t}^* = (t_1^*, \ldots, t_m^*). \tag{4}$$

Example 1 Figure 3c shows the embedded geodesic and minimal carrier between the trees T^* and T (see Fig. 3a, b), which correspond to the points $\boldsymbol{u}^*$ and $\boldsymbol{u}$, respectively. The minimal support consists of $A_1 = \{u_1^*, u_2^*\}$, $A_2 = \{u_3^*\}$, $B_1 = \{-u_2\}$, and $B_2 = \{-u_1, -u_3\}$. For convenience, π is the identity permutation in this case. The minimal carrier consists of the all positive octant determined by $x_1 > 0$, $x_2 > 0$ and

$x_3 > 0$; the 2-dimensional quadrant formed by the positive x_3 and negative x_2 axes; and the all negative octant.

For any $1 \leqslant l \leqslant m$, let V^l be the subspace of $\mathbb{R}^l$ that is the union of the (closed) orthants $\mathcal{P}_i$, $i = 0, \ldots, l$, where

$$\mathcal{P}_i = \{(x_1, \ldots, x_l) \in \mathbb{R}^l \mid x_j \leqslant 0 \quad \text{for } j \leqslant i \quad \text{and} \quad x_j \geqslant 0 \text{ for } j > i\}.$$

For the given T^*, T, and corresponding k from Lemma 1, there are $k + 1$ orthants in the carrier of the geodesic between T^* and T. If $k = m$ (the intrinsic dimension of T_{m+2}), then the carrier $\mathcal{C}$ is isometric to $\tilde{\mathcal{C}} = V^m$, with $\mathcal{O}_i$ coinciding with $\mathcal{P}_i$ and with the geodesic from u^* to u being a straight line contained in $\tilde{\mathcal{C}}$. Otherwise if $k < m$, the space $\tilde{\mathcal{C}}$ is strictly contained in V^m, and some of the top-dimensional orthants of V^m may not correspond to orthants in tree space. Additionally, the geodesic between u^* and u in $\tilde{\mathcal{C}}$ will bend at certain orthant boundaries within the ambient space V^m. We now give an isometric embedding onto V^k of a subspace of $\tilde{\mathcal{C}}$ containing the geodesic in V^m such that the image geodesic is a straight line.

The geodesic between u^* and u passes through k orthant boundaries. At the ith orthant boundary, the edges in A_i, which have been shrinking in length since the geodesic started at u^*, simultaneously reach length 0, and the edges in B_i simultaneously appear in the tree with length 0 and start to grow in length. The length of each edge in A_i changes linearly as we move along the geodesic, and thus since these lengths all reach 0 at the same point, the ratios of these lengths to each other remain the same along the geodesic. An analogous statement can be made for the lengths of the edges in B_i (cf. Owen 2011, Corollary 4.3). The basic idea behind the embedding into V^k is that because the lengths of the edges in A_i, for any i, are all linearly dependent on each other, we can represent those edges in V^k using only one dimension, and analogously for the edges in B_i.

More specifically, for $1 \leqslant i \leqslant k$, let

$$v_i^* = (u_{|A_1|+\cdots+|A_{i-1}|+1}^*, \ldots, u_{|A_1|+\cdots+|A_{i-1}|+|A_i|}^*),$$

be the projection of u^* on the orthant $\mathcal{O}(A_i)$. That is, the coordinates of v_i^* are the lengths of the edges in A_i, ordered as chosen above. Similarly, let

$$v_i = (u_{|B_1|+\cdots+|B_{i-1}|+1}, \ldots, u_{|B_1|+\cdots+|B_{i-1}|+|B_i|}),$$

so that the coordinates of v_i are the lengths of the edges in B_i, in that order. Then, the geodesic between T^* and T in $\mathcal{C}$ is piece-wise linearly isometric with the Euclidean geodesic between the vectors

$$v^* = (\|v_1^*\|, \ldots, \|v_k^*\|) = (\|A_1\|_{T^*}, \ldots, \|A_k\|_{T^*})$$

and

$$v = (-\|v_1\|, \ldots, -\|v_k\|) = (-\|B_1\|_T, \ldots, -\|B_k\|_T)$$

in V^k and, hence, in $\mathbb{R}^k$. In particular, the Euclidean distance between these two Euclidean points is the same as the distance between T^* and T in $\mathcal{C}$. Thus, we have the following result, the essence of which appears in Owen (2011, Theorem 4.10), to which we refer readers for more detailed proof.

Lemma 2 *For any given T^* and T in $\mathbf{T}_{m+2}$ with no common edge and with T^* lying in a top-dimensional stratum, there is an integer k, $1 \leqslant k \leqslant m$, for which there are two vectors $\mathbf{v}^*, \mathbf{v} \in \mathbb{R}^k$, depending on both T^* and T, such that the geodesic between T^* and T is homeomorphic and piece-wise linearly isometric, with the (straight) Euclidean geodesic between $\mathbf{v}^*$ and $\mathbf{v}$, where $\mathbf{v}^*$ lies in the positive orthant of $\mathbb{R}^k$ and $\mathbf{v}$ in the closure of the negative orthant.*

For Example 1, $k = 2$, and thus the grey area shown in Fig. 3c is isometrically mapped to V^2, as shown in Fig. 3d.

In the general case where T^* and T have a common edge, say e, this common edge determines, for each of the two trees, two quotient trees T_i^* and T_i, $i = 1, 2$, described as follows (cf. Owen 2011; Vogtmann 2007). The trees T_1^* and T_1 are obtained by replacing the subtree 'below' e with a single new leaf, so that e becomes an external edge. These two replaced subtrees form the trees T_2^* and T_2, with the 'upper' vertex of the edge e becoming the new root. Then, the geodesic $\gamma(t)$ between T^* and T is isometric with $(\gamma_e(t), \gamma_1(t), \gamma_2(t))$, where γ_e is the linear path from $|e|_{T^*}$ to $|e|_T$ and γ_i is the geodesic from T_i^* to T_i in the corresponding tree space. For this, we treat $\mathbf{T}_1$ and $\mathbf{T}_2$, the spaces of trees with no internal edges, as single points, so that any geodesic in them is a constant path. Assuming that T_i^* and T_i have no common edge for $i = 1$ or 2, we may obtain, as above, a straightened image of each geodesic γ_i in V^{k_i} with T_i^* represented in the positive orthant and T_i in the negative one. Combining these with the geodesic γ_e, which is already a straight linear segment, we have an isometric representation of γ as a straight linear segment in $\mathbb{R}_+ \times V^{k_1} \times V^{k_2}$. In this case, the sequence of strata containing the tree space geodesic between T^* and T is contained in the product of the carriers for the relevant quotient trees, together with an additional factor for the common edge. For example, if $0 < t_1 < t_2 < t_3 < t_4 < 1$ and the geodesic γ_1 spends $[0, t_2]$ in orthant $\mathcal{O}_1$, $[t_2, t_3]$ in orthant $\mathcal{O}_2$, $[t_3, 1]$ in orthant $\mathcal{O}_3$, while the geodesic γ_2 spends $[0, t_1]$ in orthant $\mathcal{P}_1$, $[t_1, t_4]$ in orthant $\mathcal{P}_2$, $[t_4, 1]$ in orthant $\mathcal{P}_3$, then the carrier for the product geodesic would be the sub-sequence

$$\mathcal{O}_1 \times \mathcal{P}_1, \mathcal{O}_1 \times \mathcal{P}_2, \mathcal{O}_2 \times \mathcal{P}_2, \mathcal{O}_3 \times \mathcal{P}_2, \mathcal{O}_3 \times \mathcal{P}_3$$

of the full lexicographically ordered sequence of nine products.

If T^* and T have more than one common edge, then either T_1^* and T_1, or T_2^* and T_2, will have a common edge and we may repeat the process. Having done so as often as necessary, we arrive at a sequence of orthants determined by the non-common edges of T^* and T. These we relabel $\mathcal{O}_0$ to $\mathcal{O}_k$ as in Lemma 1. If $\mathcal{O}_{-1}$ is the orthant determined by the axes corresponding to the common edges of T^* and T, then the sequence

$$\mathcal{O}_{-1} \times \mathcal{O}_0, \mathcal{O}_{-1} \times \mathcal{O}_1, \ldots, \mathcal{O}_{-1} \times \mathcal{O}_k$$

is the carrier of the geodesic from T^* to T. Similarly, the support for the tree space geodesic between T^* and T is found by interleaving the partitions in the supports of the relevant quotient trees, so that property (P2) is satisfied in the combined support. The resulting partitions $\mathcal{A}$ and $\mathcal{B}$ are then preceded by the set $A_0 = B_0$ of axes corresponding to the common edges so that $\mathcal{O}_{-1} = \mathcal{O}(A_0) = \mathcal{O}(B_0)$, with the convention that the corresponding ratio is $-\|A_0\|_{T^*}/\|B_0\|_T$, and (3) is modified to

$$\mathcal{O}_i = \mathcal{O}(B_0 \cup B_1 \cup \cdots \cup B_i \cup A_{i+1} \cup \cdots \cup A_k).$$

In this generalised context, the value $k = 0$ is now possible, implying that all edges are common to T^* and T. In other words, they lie in the same orthant. Note that this presentation differs slightly from that in Section 1.2 of Miller et al. (2015) in that, by collecting all the common edges in a single member $A_0 = B_0$ of the support, we are implicitly suppressing the axiom (P3) for that set. Note that the maximum value of the number k, which is determined by the non-common edges of T^* and T, is $m - |A_0|$ in the general case.

Definition 1 We call k, the number of changes of orthant in the unique minimal carrier of the geodesic from T^* to T, the carrier number $k(T^*, T)$ of T^* and T.

Clearly, $k(T^*, T) = k(T, T^*)$.

The minimal carrier and support determine the corresponding $\boldsymbol{u}^*$, $\boldsymbol{v}_i^*$, $\boldsymbol{v}^*$ and $\boldsymbol{v}$ in a similar manner to the special case where there is no common edge between T^* and T given in Lemma 2, modified to account for the common edges. For this, the first $|A_0|$ coordinates of $\boldsymbol{u}^*$ will be the $(A_0)_{T^*}$ and those of $\boldsymbol{u}$ will be $+(B_0)_T$; for $k = k(T^*, T)$ and $1 \leqslant i \leqslant k$,

$$\boldsymbol{v}_i^* = \left(u_{|A_0|+|A_1|+\cdots+|A_{i-1}|+1}^*, \ldots, u_{|A_0|+|A_1|+\cdots+|A_{i-1}|+|A_i|}^* \right) \tag{5}$$

and $\boldsymbol{v}_i$ is modified similarly; and $\boldsymbol{v}^*$ and $\boldsymbol{v}$ have additional first coordinates $\boldsymbol{v}_0^* = (A_0)_{T^*}$ and $\boldsymbol{v}_0 = (B_0)_T$, respectively. Then, with this modification, the geodesic between T^* and T is homeomorphic and piece-wise linearly isometric, with the (straight) Euclidean geodesic between $\boldsymbol{v}^*$ and $\boldsymbol{v}$, where $\boldsymbol{v}^*$ lies in the positive orthant of $\mathbb{R}_+^{|A_0|} \times \mathbb{R}^k$. This generalisation of Lemma 2 to the general case was obtained, with different notation, in Billera et al. (2001), Owen (2011) and Vogtmann (2007). Then, the log map as defined by (1) can be expressed using these vectors as follows.

Theorem 1 *Fix T^* in a top-dimensional stratum of $\boldsymbol{T}_{m+2}$ with coordinates $\boldsymbol{t}^* = (t_1^*, \ldots, t_m^*)$, where the ordering of the coordinates is that induced by the canonical ordering for $\mathbb{R}^M$. For $T \in \boldsymbol{T}_{m+2}$, there are vectors $\boldsymbol{v}^*$ and $\boldsymbol{v}$ in $\mathbb{R}_+^{|A_0|} \times \mathbb{R}^k$, where $|A_0|$ is the number of common edges of T^* and T, k is the carrier number $k(T^*, T)$ and $\boldsymbol{v}^*$ lies in the positive orthant of the corresponding space, and a linear map ρ such that*

$$\log_{T^*}(T) = \rho(\boldsymbol{v} - \boldsymbol{v}^*) = \rho(\boldsymbol{v}) - \boldsymbol{t}^*. \tag{6}$$

Proof Let $\boldsymbol{v}^*$ have $(|A_0| + i)$th coordinate $\|\boldsymbol{v}_i^*\|$, $i = 1, \ldots, k$, with the additional initial coordinates $\boldsymbol{v}_0^*$ when T^* and T have common edges, where $\boldsymbol{v}_i^*$ are as defined by (5), and $\boldsymbol{v}$ be determined similarly. The piece-wise linear isometry that straightens the geodesic from T^* to T, given in Lemma 2 for the special case as well as the above for the general case, has an inverse on the positive orthant in $\mathbb{R}_+^{|A_0|} \times \mathbb{R}^k$. This inverse is given by

$$\chi : \boldsymbol{e}_i \mapsto \frac{1}{\|\boldsymbol{v}_i^*\|} \boldsymbol{v}_i^* \qquad 1 \leqslant i \leqslant k, \tag{7}$$

and the identity on the $|A_0|$ initial coordinates, where $\boldsymbol{e}_i$ is the $(|A_0|+i)$th standard basis vector in $\mathbb{R}_+^{|A_0|} \times \mathbb{R}^k$. Note that, being a linear map, when $A_0 = \emptyset$, $\chi((x_1, \ldots, x_k)) = \sum_{i=1}^k x_i \frac{1}{\|\boldsymbol{v}_i^*\|} \boldsymbol{v}_i^*$, where $(x_1, \ldots, x_k) = \sum_{i=1}^k x_i \boldsymbol{e}_i \in \mathbb{R}^k$. Although it is not expressed precisely as it is here, the idea for a more detailed derivation of this in this case is captured in Theorem 4.4 in Owen (2011), where χ is denoted by g_0.

Since $\boldsymbol{v}^* = \boldsymbol{v}_0^* + \sum_{i=1}^k \|\boldsymbol{v}_i^*\| \boldsymbol{e}_i$, we have that $\chi(\boldsymbol{v}^*) = \boldsymbol{u}^*$ and that χ maps the initial segment of the straight geodesic in $\mathbb{R}_+^{|A_0|} \times \mathbb{R}^k$, together with its initial tangent vector $\boldsymbol{v} - \boldsymbol{v}^*$, onto those of the geodesic in V^m. The permutation π, which maps the positive orthant in V^m into $\boldsymbol{T}_m \subset \mathbb{R}^M$ where M is defined by (2), is also an isometry preserving the initial segments of the geodesics. It follows that

$$\log_{T^*}(T) = \pi \circ \chi(\boldsymbol{v} - \boldsymbol{v}^*). \tag{8}$$

Noting that the maps π and χ are linear and $\pi \circ \chi(\boldsymbol{v}^*) = t^*$, the required result follows by taking $\rho = \pi \circ \chi$. □

Figure 4 shows the log map for the tree T^* for Example 1.

Although $\boldsymbol{T}_{m+2}$ is $CAT(0)$, $\log_{T^*}$ is not a one-to-one map. In particular, if T_1 and T_2 are two different trees such that $k(T^*, T_1) = k(T^*, T_2)$ is not maximal, then it is possible that $\log_{T^*}(T_1) = \log_{T^*}(T_2)$, as observed in the case of $\boldsymbol{T}_4$ in Barden et al. (2013). As another example, consider two trees $T^*, T \in \boldsymbol{T}_{m+2}$ with no common edges such that the geodesic between them passes through the cone point with a given length l. Then for any other tree T' with a geodesic to T^* of length l, passing through the cone point, we also have that $\log_{T^*}(T) = \log_{T^*}(T')$.

Fig. 4 The log map for tree T^* in Example 1. The vector between $\boldsymbol{u}^*$ and $log_{T^*}(T)$ is shown as a *dashed line*. It coincides with the geodesic between T^* and T in the starting orthant, but then continues into the ambient space, while the geodesic must bend to remain in the tree space
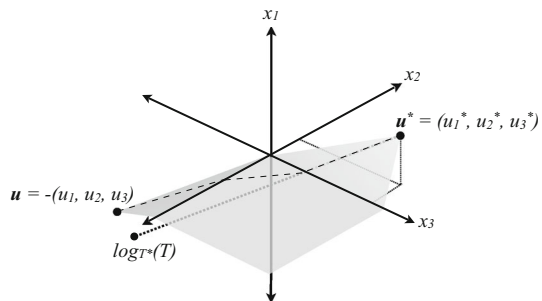
Recalling that each component of v_i^* and v_i is, respectively, the length of an edge in A_i and B_i then, with some ambiguity in the ordering of the edges of T^*, another equivalent way to express $\log_{T^*}$ is

$$\log_{T^*}(T) = \{\bar{B}_0 - \bar{A}_0\} - \sum_{j=1}^{k} \frac{\|B_j\|_T + \|A_j\|_{T^*}}{\|A_j\|_{T^*}} \bar{A}_j$$

where $\bar{A}_j = (e_{T^*})_{e \in A_j}$. To derive the limiting distribution of sample Fréchet means, the ordering must be kept explicit and independent of T. Hence, we have to use the expression for the log map given by (8), even though it is not as transparent as this one.

Note also that, although the definitions for both π and χ implicitly depend on the ordering we chose for the coordinates of u^*, the composition $\pi \circ \chi$ is independent of that choice, and so the log map is well defined, as long as we chose the same ordering for u^* for both π and χ.

The minimal carrier that determines the maps π and χ as well as the vectors v^* and v depends on both T^* and T, although we have suppressed that dependence in the notation. However, there are only finitely many choices for the carrier number $k(T^*, T)$ and the minimal support when T^* is fixed and T varies within a given stratum of T_{m+2}. In particular, if $k(T^*, T)$ remains constant in a neighbourhood of (T^*, T), then π and χ do not change for small enough changes in T^* and T. It follows that there are only finitely many possibilities for the form (6) that $\pi \circ \chi$ takes when T varies in T_{m+2}. Here, by form, we mean the algebraic expression of $\log_{T^*}$ as a map. That is, by '$\log_{T^*}(T_1)$ and $\log_{T^*}(T_2)$ taking the same form', we mean that they can be obtained using a single algebraic expression for $\log_{T^*}$. Since the permutation π returns all the axes to their canonical order, this expression is determined by the partition $\mathcal{A}$ of the edges of T^*, with the subsets of non-common edges possibly permuted. For example, in the case of T_4, $\log_{T^*}$ only takes two possible forms, depending on whether the geodesic from T^* to T passes through the cone point or not where the cone point, the origin in $\mathbb{R}^M$, represents the tree whose two edges have zero length. The two corresponding subsets of T_4 are, respectively, indicated by the unions of light and dark grey regions in Figure 3 of Barden et al. (2013) when T^* is the tree corresponding to (x_i, x_j). The different possibilities for the form (6) give rise to a *polyhedral subdivision* of tree space T_{m+2}, defined as follows.

Definition 2 For a fixed T^* lying in a top-dimensional stratum of T_{m+2}, the polyhedral subdivision of tree space T_{m+2}, with respect to T^*, is determined by the possible forms that $\log_{T^*}$ can take: each polyhedron of the subdivision is the closure of the set of trees T that have a particular form for $\log_{T^*}(T)$. We shall call each such top-dimensional polyhedron a maximal cell of the polyhedral subdivision and let $\mathcal{D}_{T^*}$ be the subset of T_{m+2} consisting of all trees that lie on the boundaries of maximal cells determined by the polyhedral subdivision with respect to T^*.

Note that, if the geodesics to T_1 and T_2 from T^* pass through the same sequence of strata, then $\log_{T^*}(T_1)$ and $\log_{T^*}(T_2)$ take the same form. However, the converse is not always true. For example, it is possible that T_1 and T_2 lie in different strata, but in the

same maximal cell. Hence, the definition of polyhedral subdivision of T_{m+2} defined here is similar to, but coarser than, the concept of 'vistal polyhedral subdivision' given in section 3 of Miller et al. (2015). This is due to the fact that, while $\mathcal{A}$ and $\mathcal{B}$ in the minimal support play a symmetric role for the geodesic between T^* and T, their roles in the log map $\log_{T^*}$ are asymmetric. When T varies, as long as the corresponding partition $\mathcal{A}$ either is unchanged or, at most, its subsets corresponding to the non-common edges are permuted, the algebraic expression for $\log_{T^*}$ remains the same.

This polyhedral subdivision varies continuously with respect to T^*. If T lies in the interior of a maximal cell of the subdivision and T^*, itself in a top-dimensional (open) stratum, varies in a small enough neighbourhood, then the support for T^* and T is unique. Then, the derivative of the log map will be well defined.

When T lies on the boundary of a maximal cell of the subdivision, but not on a stratum boundary, the possible supports for T^* and T are those determined by the polyhedra to which that boundary belongs. However, all these supports give rise to the same geodesic between T^* and T, as they must, since T_{m+2} is a $CAT(0)$-space, and among them will be the minimal support that we are assuming for our analysis. Moreover, in this case, there is at least one non-minimal support for T^* and T with the property that, for the corresponding $\boldsymbol{v}^*$ and $\boldsymbol{v}$, $\|\boldsymbol{v}_i^*\|/\|\boldsymbol{v}_i\| = \|\boldsymbol{v}_{i+1}^*\|/\|\boldsymbol{v}_{i+1}\|$ for some $i \geqslant 1$.

Recall that from Definition 1 that the carrier number counts the number of orthants that the geodesic from T^* to T meets in a linear segment of positive length. It will become clear later that the set of trees T for which, for a given T^*, the carrier number $k(T^*, T)$ is less than its possible maximum $m - |A_0|$, where $|A_0|$ is the number of common edges of T^* and T, plays a role that distinguishes the limiting distributions of sample Fréchet means in the tree spaces from those in Euclidean space. Hence, we introduce the following definition.

Definition 3 A point $T \in T_{m+2}$ is called singular, with respect to a tree T^* lying in a top-dimensional stratum, if the carrier number $k(T^*, T)$ of T^* and T is less than $m - |A_0|$. The set of such singular points will be denoted by $\mathcal{S}_{T^*}$.

The following result describes the image, under $\log_{T^*}$, in the tangent space at T^* of the set $\mathcal{S}_{T^*}$: although $\mathcal{S}_{T^*}$ may be rather complex, its image is relatively simple.

Corollary 1 *If $T^* \in T_{m+2}$ lies in a top-dimensional stratum, then the image, under $\log_{T^*}$, of the set $\mathcal{S}_{T^*}$ of the singular points with respect to T^* is contained in the union of the hyperplanes $x_i t_j^* = x_j t_i^*$, $1 \leqslant i \neq j \leqslant m$, in $\mathbb{R}^m$.*

Proof The number of orthants in the minimal carrier of the geodesic from T^* to T is less than $m - |A_0|$ if and only if the dimension j_i of some vector $\boldsymbol{v}_i^*$ is greater than one for $i \geqslant 1$. Then, χ maps the line determined by $\boldsymbol{e}_i$ in $\mathbb{R}_+^{|A_0|} \times \mathbb{R}^k$ into the subspace of $\mathbb{R}^m$ that is the intersection of the co-dimension one hyperplanes $x_{i'} u_{j'}^* = x_{j'} u_{i'}^*$ in $\mathbb{R}^m$, where $j_1 + \cdots + j_{i-1} < i' \neq j' \leqslant j_1 + \cdots + j_i$ and where the ordering of the coordinates $u_{i'}^*$, and hence of the $x_{i'}$, is as in the minimal carrier. Then, applying the permutation π and using the same notation for the permuted $\boldsymbol{x}$-coordinates, the result follows. $\square$
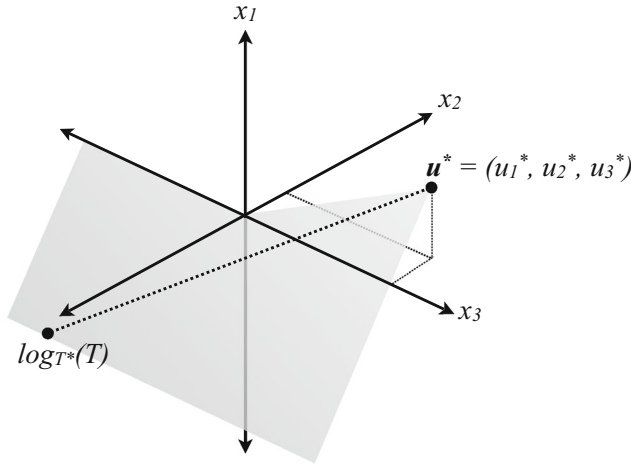
Fig. 5 The *grey* area is part of the hyperplane $x_1 \cdot u_2^* = x_2 \cdot u_1^*$, which contains some of the singular points for the log map log_{T^*} for Example 1

For example, see Fig. 5 for an illustration of one of the hyperplanes for Example 1.

To describe the limiting behaviour of sample Fréchet means, it will be more convenient to have a modified version of the log map, Φ_{T^*}, at T^* defined by

$$\Phi_{T^*}(T) = \log_{T^*}(T) + t^*. \tag{9}$$

In the present context, where T^* lies in a top-dimensional stratum, $\Phi_{T^*}(T) = \pi \circ \chi(v)$.

Note that, when T^* lies in a top-dimensional stratum, the map corresponding to Φ_{T^*} here obtained in Barden et al. (2013) in the case of T_4 was expressed as the composition of a similarly defined map on Q_5, a simpler auxiliary stratified space, with a map from Q_5 to T_4. Instead of the log map, that map on Q_5 was expressed in terms of the gradient of the squared distance function. The relationship between the latter and the log map shows that the resulting expression in Barden et al. (2013) is equivalent to the one defined here. The derivation of Φ_{T^*} from $\log_{T^*}$ implicitly requires that the tangent space to T_{m+2} at T^*, in which the image of $\log_{T^*}$ lies, be translated to the parallel copy $\mathbb{R}^m$ at the origin, in which it makes sense to add the coordinate vector t^*. As a result, for all $\widetilde{T}^*$ in the same stratum as T^*, the image of $\Phi_{\widetilde{T}^*}$ will lie in this same subspace $\mathbb{R}^m$.

3 Fréchet means on a top-dimensional stratum

Let μ be a probability measure on T_{m+2} and assume that the Fréchet function for μ is finite. The space T_{m+2} being $CAT(0)$ implies that the Fréchet function for μ is strictly convex so that, in particular, the Fréchet mean of μ is unique when it exists. In this section, we consider the case when this mean, denoted by T^*, lies in a top-

dimensional stratum. For this, as in the previous section, we identify any tree $\widetilde{T}^*$ in the stratum of $\boldsymbol{T}_{m+2}$ in which T^* lies with the point in the positive orthant of $\mathbb{R}^m$ having the lengths of the internal edges of $\widetilde{T}^*$ as coordinates in the canonical order. In particular, $T^* = (t_1^*, \ldots, t_m^*)$.

First, we use the log map to give a necessary and sufficient condition for T^* to be the Fréchet mean of μ as follows, generalising the characterisation of Fréchet means on complete and connected Riemannian manifolds of non-negative curvature. In particular, it shows that, when T is a random variable on $\boldsymbol{T}_{m+2}$ with distribution μ and T^* is in a top-dimensional stratum, then T^* is the Fréchet mean of μ if and only if T^* is the Euclidean mean of the Euclidean random variable $\Phi_{T^*}(T)$.

Lemma 3 *Assume that the Fréchet mean T^* of μ lies in a top-dimensional stratum. Then, T^* is characterised by the following condition:*

$$\int_{\boldsymbol{T}_{m+2}} \Phi_{T^*}(T)\, \mathrm{d}\mu(T) = T^*. \tag{10}$$

Proof It can be checked that, since T^* lies in a top-dimensional stratum, the squared distance $d(T^*, T)^2$ is differentiable at T^* and its gradient at T^* is $-2\log_{T^*}(T)$. Thus, as discussed in Barden et al. (2013) T^*, lying in a top-dimensional stratum, is the Fréchet mean of a given probability measure μ on $\boldsymbol{T}_{m+2}$ if and only if

$$\int_{\boldsymbol{T}_{m+2}} \log_{T^*}(T)\, \mathrm{d}\mu(T) = 0.$$

Then, the required result follows by re-expressing the above condition for T^* to be the Fréchet mean of μ in terms of Φ_{T^*} given by (9). □

The derivation of the central limit theorem for Fréchet means in $\boldsymbol{T}_{m+2}$ requires the study of the change of Φ_{T^*} as T^* changes with T remaining fixed. For this we recall that, for a fixed T, the minimal support for the geodesic between T^* and T determines a particular maximal cell, in which T lies, of the polyhedral subdivision with respect to T^*. When the minimal support for the geodesic between $\widetilde{T}^*$ and T is the same as that for T^* and T, we shall say that the two resulting maximal cells *correspond to each other*. We have the following result on the derivative of Φ_{T^*} with respect to T^*, noting that the derivative of the map

$$(x_1, \ldots, x_l) \mapsto \frac{1}{\|(x_1, \ldots, x_l)\|}(x_1, \ldots, x_l)$$

is

$$M^{\dagger}_{(x_1,\ldots,x_l)} = \frac{1}{\|(x_1, \ldots, x_l)\|} I_l - \frac{1}{\|(x_1, \ldots, x_l)\|^3} \begin{pmatrix} x_1 \\ \vdots \\ x_l \end{pmatrix} (x_1 \cdots x_l),$$

where, in particular, when $l = 1$, $M^{\dagger}_{x_1} = 0$.

Lemma 4 *Assume that $T^* \in \mathbf{T}_{m+2}$ lies in a top-dimensional stratum. Then, for any fixed $T \in \mathbf{T}_{m+2}$ lying in the interior of a maximal cell of the polyhedral subdivision with respect to T^*, $\Phi_{T^*}(T)$ is differentiable with respect to T^*. Moreover, for such T, if $\mathbf{v}_i^*$ is as defined in (5) prior to Theorem 1 and $\mathbf{v}_i$ defined analogously for $i = 1, \ldots, k$, where $k = k(T^*, T)$, then the derivative of $\Phi_{T^*}(T)$ at T^*, with respect to T^*, is given by*

$$M_{T^*}(T) = P_{T^*, T}^\top \operatorname{diag}\left\{ v_1 M_{\mathbf{v}_1^*}^\dagger, \ldots, v_k M_{\mathbf{v}_k^*}^\dagger \right\} P_{T^*, T} \tag{11}$$

where $v_i = \|\mathbf{v}_i\|$ and $P_{T^*, T}$ denotes the matrix representing the permutation π defined by (4).

Note that, for the sub-matrix $v_i M_{\mathbf{v}_i^*}^\dagger$ to be non-zero, $\mathbf{v}_i^*$ must be at least 2-dimensional and, by definition, v_i is non-positive so that T must lie in $\mathcal{S}_{T^*}$. In particular, if $k(T^*, T) = m - |A_0|$, in other words, if the geodesic between T^* and T is 'straight', then the derivative of $\Phi_{T^*}(T)$ at T^* is zero. This could be seen directly: since, in that case, the tree space geodesic between T^* and T would be a Euclidean geodesic between them. Then, $\log_{T^*}(T) = \mathbf{t} - \mathbf{t}^*$ so that $\Phi_{T^*}(T) = \mathbf{t}$ independent of $\mathbf{t}^*$.

Proof By the discussion preceding the lemma, the edges common to T^* and T will make no contribution to the derivative. Since the polyhedral subdivision is continuous with respect to T^*, it is sufficient to show that, when $\tilde{T}^*$ is sufficiently close to T^*, so that in particular T^* and $\tilde{T}^*$ lie in the same top stratum and T lies in the interior of the corresponding maximal cells of the polyhedral subdivisions with respect to T^* and $\tilde{T}^*$, we have

$$\begin{aligned} &\Phi_{\tilde{T}^*}(T) - \Phi_{T^*}(T) \\ &\approx (\tilde{T}^* - T^*)\, P_{T^*, T}^\top \operatorname{diag}\left\{ v_1 M_{\mathbf{v}_1^*}^\dagger, \ldots, v_k M_{\mathbf{v}_k^*}^\dagger \right\} P_{T^*, T} \\ &\quad + \|T\|\, o(\|\tilde{T}^* - T^*\|). \end{aligned} \tag{12}$$

To show (12), it is sufficient to assume that T^* and T have no common edge. Moreover, since $\pi_{T^*, T}$, and so $P_{T^*, T}$, is a linear map, its derivative is identical with itself. Hence, by applying the appropriate permutation to re-order the $\tilde{\mathbf{u}}^*$ and $\mathbf{u}$ corresponding to $\tilde{T}^*$ and T when necessary, it is sufficient to show that

$$\begin{aligned} &\{\Phi_{\tilde{T}^*}(T) - \Phi_{T^*}(T)\} P_{T^*, T}^\top \\ &\approx (\tilde{\mathbf{u}}^* - \mathbf{u}^*)\, \operatorname{diag}\left\{ v_1 M_{\mathbf{v}_1^*}^\dagger, \ldots, v_k M_{\mathbf{v}_k^*}^\dagger \right\} + \|T\|\, o(\|\tilde{T}^* - T^*\|). \end{aligned}$$

Since T lies in the interior of a maximal cell of the polyhedral subdivision of $\mathbf{T}_{m+2}$ with respect to T^*, then $v_i^*/v_i > v_{i+1}^*/v_{i+1}$ for all i, where all v_i are negative. By continuity, all these strict inequalities hold when v_i^* is replaced by $\tilde{v}_i^*$ if $\tilde{T}^*$ is sufficiently close to T^*. Hence, T lies in the interior of a maximal cell of the polyhedral subdivision of $\mathbf{T}_{m+2}$ with respect to $\tilde{T}^*$. Thus, the only difference between

the expressions for $\Phi_{T^*}(T)$ and $\Phi_{\widetilde{T}^*}(T)$ is that $\boldsymbol{v}^*$ and $\boldsymbol{v}_i^*$ in the former are replaced by $\tilde{\boldsymbol{v}}^*$ and $\tilde{\boldsymbol{v}}_i^*$, respectively, in the latter. It follows that, in this case, the difference $\{\Phi_{\widetilde{T}^*}(T) - \Phi_{T^*}(T)\}P_{T^*, T}^{\top}$ can be expressed as

$$\left(v_1\left(\frac{1}{\|\tilde{\boldsymbol{v}}_1^*\|}\tilde{\boldsymbol{v}}_1^* - \frac{1}{\|\boldsymbol{v}_1^*\|}\boldsymbol{v}_1^*\right), \ldots, v_k\left(\frac{1}{\|\tilde{\boldsymbol{v}}_k^*\|}\tilde{\boldsymbol{v}}_k^* - \frac{1}{\|\boldsymbol{v}_k^*\|}\boldsymbol{v}_k^*\right)\right).$$

The required result follows by applying the first-order Taylor expansion to each sub-vector component and using the formula preceding the statement of the Lemma. $\qquad\square$

If T lies on the boundary of a maximal cell of the polyhedral subdivision of $\boldsymbol{T}_{m+2}$ with respect to T^*, each choice of maximal cell of the polyhedral subdivision with respect to T^* will determine a support for the geodesic from T^* to T. If we restrict the neighbouring $\widetilde{T}^*$ to move from T^* in a direction such that T lies in the corresponding maximal cell of the polyhedral subdivision with respect to $\widetilde{T}^*$, then the argument in the proof for Lemma 4 still holds. Thus, $\Phi_{T^*}(T)$ will have all directional derivatives, at T^*, with respect to T^* having similar forms to that given in Lemma 4. However, some different directions will require different choices of maximal cell of the polyhedral subdivision with respect to T^* in which T lies. Thus, the directional derivative will have different forms and $\Phi_{T^*}(T)$ will not be differentiable.

Lemma 4 enables us to obtain the limiting distribution of the sample Fréchet means of a sequence of *iid* random variables on $\boldsymbol{T}_{m+2}$ when the Fréchet mean of the underlying probability measure lies in a top-dimensional stratum as follows, recalling that $\mathcal{D}_{T^*}$, defined in Definition 2, is the subset of $\boldsymbol{T}_{m+2}$ consisting of all trees that lie on the boundaries of maximal cells determined by the polyhedral subdivision with respect to T^*. On one hand, the result shows that, in this case, the limiting distribution, being a Gaussian distribution, bears a certain similarity to that of the sample means of Euclidean random variables. On the other hand, recalling that the derivative of $\Phi_{T^*}(T)$ at T^* is zero if $T \notin \mathcal{S}_{T^*}$, it also shows that the role played by $\mathcal{S}_{T^*}$ in the limiting behaviour of the sample Fréchet means is reflected in the covariance structure of the Gaussian distribution, departing from the limiting distribution of the sample means of Euclidean random variables.

Theorem 2 *Let μ be a probability measure on $\boldsymbol{T}_{m+2}$ with finite Fréchet function and with Fréchet mean T^* lying in a top-dimensional stratum. Assume that $\mu(\mathcal{D}_{T^*}) = 0$. Suppose that $\{T_i : i \geqslant 1\}$ is a sequence of iid random variables in $\boldsymbol{T}_{m+2}$ with probability measure μ and denote by $\hat{T}_n$ the sample Fréchet mean of $T_1, \ldots, T_n$. Then,*

$$\sqrt{n}(\hat{T}_n - T^*) \xrightarrow{d} N(0, A^{\top}VA), \qquad as\ n \to \infty,$$

where V is the covariance matrix of the random variable $\log_{T^*}(T_1)$, or equivalently that of $\Phi_{T^*}(T_1)$, and

$$A = \{I - E[M_{T^*}(T_1)]\}^{-1}, \tag{13}$$

assuming that this inverse exists, and where $M_{T^*}(T)$ is the $m \times m$ matrix defined by (11).

Proof The main argument underlying the proof is similar to that of the proof in Barden et al. (2013) for T_4, i.e. to express the difference between the Fréchet mean of the underlying probability measure and the sample Fréchet means in terms of the difference $\Phi_{\hat{T}^*}(T_i) - \Phi_{T^*}(T_i)$. However, the proof in Barden et al. (2013) relies on an explicit embedding that is only valid for T_4. As a consequence of Lemma 4, we can now achieve this for any tree space.

Since $\hat{T}_n$ is the Fréchet sample mean of $T_1, \ldots, T_n$, then for sufficiently large n, $\hat{T}_n$ will be close to T^* a.s. (cf. Ziezold 1977) and, in particular, lie in the same stratum as T^*. Thus, the above results (10) and (12) give

$$\sqrt{n}(\hat{T}_n - T^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\Phi_{\hat{T}_n}(T_i) - T^*\}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\Phi_{T^*}(T_i) - T^*\} + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\Phi_{\hat{T}_n}(T_i) - \Phi_{T^*}(T_i)\}$$

$$\approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\Phi_{T^*}(T_i) - T^*\} + \sqrt{n}(\hat{T}_n - T^*)\frac{1}{n} \sum_{i=1}^{n} M_{T^*}(T_i)$$

$$+ o(\|\hat{T}_n - T^*\|)\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \|T_i\|.$$

Hence,

$$\sqrt{n}(\hat{T}_n - T^*)\left\{I - \frac{1}{n}\sum_{i=1}^{n} M_{T^*}(T_i)\right\}$$

$$\approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\Phi_{T^*}(T_i) - T^*\} + o(\|\hat{T}_n - T^*\|)\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \|T_i\|.$$

Since $\{\Phi_{T^*}(T_i) : i \geqslant 1\}$ is a sequence of *iid* random variables in $\mathbb{R}^m$ with mean T^* and $\{M_{T^*}(T_i) : i \geqslant 1\}$ is a sequence of *iid* random matrices, the following theorem follows from the standard Euclidean result as in Barden et al. (2013). □

Recalling that $M_{T^*}(T_1) = 0$ for T_1 not lying in the singularity set of $\log_{T^*}$, we see that the contribution to $E[M_{T^*}(T_1)]$ consists of all singular points of $\log_{T^*}$. For $m = 1$, i.e. the case for T_3, the only possible choice for k is $k = 1 = m$ which implies that $M_{T^*}(T) \equiv 0$, so that the above result for this special case is the same as that obtained in Hotz et al. (2013). For $m = 2$, i.e. the case for T_4, the only possible case for T lying in the singularity set of $\log_{T^*}$ is when $k = 1$, which corresponds to the geodesic between T^* and T passing through the origin and $\Phi_{T^*}(T) = -\|T\|\frac{1}{\sqrt{(t_1^*)^2+(t_2^*)^2}}(t_1^*, t_2^*)$. Then, the corresponding $M_{T^*}(T)$ has the expression

$$M_{T^*}(T) = -\|T\| \frac{1}{\|T^*\|^3} \begin{pmatrix} -t_2^* \\ t_1^* \end{pmatrix} \left(-t_2^* \; t_1^* \right),$$

so that the above result for this case recovers that in Barden et al. (2013).

Note that μ induces, by $\log_{T^*}$, a probability distribution μ' on the tangent space of T_{m+2} at T^*. Then, the sample Fréchet means of μ' are the standard Euclidean means

$$\frac{1}{n} \sum_{i=1}^{n} \log_{T^*}(T_i) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \Phi_{T^*}(T_i) - T^* \right\},$$

so that the rescaled sample Fréchet means have the limiting distribution $N(0, V)$. However, the sample Fréchet means of μ' are generally different from $\log_{T^*}(\hat{T}_n)$, the log images of the sample Fréchet means of μ, and there is no closed expression for the relationship between the two.

It is also interesting to compare the result of Theorem 2 with the limiting distributions for the sample Fréchet means on Riemannian manifolds obtained in Kendall and Le (2011). Both limiting distributions take a similar form, with the role played by curvature in the case of manifolds being replaced here by the global topological structure of the tree space.

4 Fréchet means on a stratum of co-dimension one

A stratum $\mathcal{O}(\Sigma)$ of co-dimension one corresponding to the set Σ of mutually compatible edge-types arises as a boundary face of a top-dimensional stratum when one, and only one, internal edge of the latter is given length zero so that its two vertices are coalesced to form a new vertex of valency four. The four incident edges determine disjoint subsets A, B, C, X of leaves and root, where X contains the root. Then, an additional internal edge may be introduced to Σ, namely α, β or γ that correspond, respectively, to the sets of leaves $A \cup B, A \cup C$ or $B \cup C$. This gives top-dimensional strata $\mathcal{O}(\Sigma \cup \alpha), \mathcal{O}(\Sigma \cup \beta)$ or $\mathcal{O}(\Sigma \cup \gamma)$, all of whose boundaries contain the stratum $\mathcal{O}(\Sigma)$. Moreover, these are the only such top-dimensional strata. For example, in Fig. 2, the leaves and root subsets are $A = \{a, b\}$, $B = \{c\}$, $C = \{d\}$, and $X = \{r\}$, while the sets of edge-types are $\Sigma = \{\{a, b\}\}$, $\alpha = \{\{c, d\}\}$, $\beta = \{\{a, b, d\}\}$ and $\gamma = \{\{a, b, c\}\}$.

If $A > B > C$ is the canonical order of the sets of leaves, then $\alpha < \beta < \gamma$ is the induced order of the edges and corresponding semi-axes and, if we write the coordinates of a tree T^* in $\mathcal{O}(\Sigma)$ as $(t_2^*, \dots, t_m^*)$, we can write the coordinates of trees in the neighbouring orthants as $(t_\alpha^*, t_\beta^*, t_\gamma^*, t_2^*, \dots, t_m^*)$ where precisely two of t_α^*, t_β^* and t_γ^* are zero, since the remaining $m - 1$ edge-types are common to all the trees involved in these three orthants and their common boundary component. Note however that, although the coordinates $(t_\alpha^*, t_\beta^*, t_\gamma^*)$ and $(t_2^*, \dots, t_m^*)$ can be chosen in canonical order, the resulting sequence $(t_\alpha^*, t_\beta^*, t_\gamma^*, t_2^*, \dots, t_m^*)$ will not in general be in canonical order.

It is clear now that the tree space T_{m+2} is not locally a manifold at any tree in the strata of co-dimension one. However, the stratification enables us to define, at a tree in

a stratum of positive co-dimension, its tangent cone (cf. Bridson and Haefliger 1999) to consist of all initial tangent vectors of smooth curves *starting* from that tree. Then, the tangent cone to T_{m+2} at a tree in a stratum of co-dimension one is an open book (cf. Hotz et al. 2013) with three pages extending each of the three strata and with the stratum of co-dimension one in which the tree lies being extended to form its spine.

The definition of the log map (1) applies equally to a tree T^* in a stratum σ of co-dimension one: if the geodesic from T^* to T passes through one of the three strata whose boundary includes σ, the unit vector component of $\log_{T^*}(T)$ is taken in the same direction in the page of the tangent book that corresponds to that stratum. The scalar component of the log map is still the distance between the trees. Similarly, the definition (9) for Φ_{T^*} remains valid in this case.

From now on, we assume that T^* lies in a stratum $\mathcal{O}(\Sigma)$ of co-dimension one. Although the squared distance $d(T^*, T)^2$ is no longer differentiable at T^*, it has directional derivatives along all possible directions. Hence, the condition for T^* to be the Fréchet mean of a probability measure μ on T_{m+2}, i.e. the condition for T^* to satisfy

$$\int_{T_{m+2}} d(T^*, T)^2 \, \mathrm{d}\mu(T) < \int_{T_{m+2}} d(T', T)^2 \, \mathrm{d}\mu(T) \qquad \text{for any } T' \neq T^*,$$

becomes that the Fréchet function for μ has, at T^*, non-negative directional derivatives along all possible directions. To investigate the latter condition, we label the three strata joined at the stratum $\mathcal{O}(\Sigma)$, of co-dimension one, in which T^* lies as the α-, β- and γ-strata and denote by $\log_{T^*}^{\alpha}$, $\log_{T^*}^{\beta}$ and $\log_{T^*}^{\gamma}$, respectively, the modifications of the map $\log_{T^*}$ that agree with $\log_{T^*}$ on the domains for which the image lies in the pages of the tangent book tangent to the α-, β- and γ-strata, respectively, and are zero elsewhere. That is, for example,

$$\log_{T^*}^{\alpha}(T) = \begin{cases} \log_{T^*}(T) & \text{if } T \text{ is such that } \log_{T^*}(T) \text{ lies in the page of the} \\ & \text{tangent book tangent to the } \alpha\text{-orthant} \\ 0 & \text{otherwise.} \end{cases}$$

Write e_α, e_β and e_γ for the outward unit vectors in the tangent book at T^* lying in the page tangent to the α-, β- and γ-strata, respectively, and orthogonal to its spine, and define

$$I_i = \int_{T_{m+2}} \langle \log_{T^*}^i(T), e_i \rangle \, \mathrm{d}\mu(T), \qquad i = \alpha, \beta, \gamma.$$

We also define $\log_{T^*}^s$ to be the modification of $\log_{T^*}$ with respect to the spine of the tangent book, the tangent space to $\mathcal{O}(\Sigma)$, analogous to the above $\log_{T^*}^i$. Then, we have the following characterisation of T^* in a stratum of co-dimension one to be the Fréchet mean of μ, in terms of the derivatives of the Fréchet function along the three directions orthogonal to the tangent space to $\mathcal{O}(\Sigma)$, as well as the Euclidean mean of $\log_{T^*}^s(T)$, where T is a random variable on T_{m+2} with distribution μ.

Lemma 5 *With the notation and definition above, a given tree T^* in a stratum $\mathcal{O}(\Sigma)$ of co-dimension one is the Fréchet mean of a given probability measure μ on $\boldsymbol{T}_{m+2}$ if and only if*

$$I_\alpha \leqslant I_\beta + I_\gamma, \qquad I_\beta \leqslant I_\gamma + I_\alpha, \qquad I_\gamma \leqslant I_\alpha + I_\beta \tag{14}$$

and

$$\int_{\boldsymbol{T}_{m+2}} \log^s_{T^*}(T)\, \mathrm{d}\mu(T) = 0. \tag{15}$$

Proof Recall that since $T^* \in \mathcal{O}(\Sigma)$, the condition for T^* to be the Fréchet mean of a probability measure μ on $\boldsymbol{T}_{m+2}$ is that the Fréchet function for μ has, at T^*, non-negative directional derivatives along all possible directions.

For any vector $\boldsymbol{w}$ at T^* which is tangent to $\mathcal{O}(\Sigma)$, the non-negativity of the directional derivative along $\boldsymbol{w}$ can be expressed as $\int_{\boldsymbol{T}_{m+2}} \langle \log^s_{T^*}(T), \boldsymbol{w} \rangle\, \mathrm{d}\mu(T) \leqslant 0$. Since $-\boldsymbol{w}$ also tangent to $\mathcal{O}(\Sigma)$ at T^*, this inequality must be an equality for all such $\boldsymbol{w}$, which gives (15). Hence, by linearity, the non-negativity of directional derivatives, of the Fréchet function for μ, at T^* along all possible directions may be characterised by requiring the non-negativity of the directional derivatives along the $\boldsymbol{e}_\alpha$, $\boldsymbol{e}_\beta$ and $\boldsymbol{e}_\gamma$ directions, together with (15). However, analogously to the deduction in Barden et al. (2013), it can be checked that the requirement for the directional derivative along each of the $\boldsymbol{e}_\alpha$, $\boldsymbol{e}_\beta$ and $\boldsymbol{e}_\gamma$ directions to be non-negative is, respectively, equivalent to each of the inequalities (14). □

To see the relation between the inequalities (14) and the asymptotic behaviour of sample Fréchet means, we will use a folding map F_α (cf. Hotz et al. 2013) that operates on the tangent book at T^*. The map F_α folds the two pages that are tangent to the β- and γ-strata onto each other, so that they form the complement in $\mathbb{R}^m$ of the closure of the page tangent to the α-stratum. Define F_β and F_γ similarly. Then, $F_\alpha \circ \log_{T^*}$ maps $\boldsymbol{T}_{m+2}$ to $\mathbb{R}^m$ and, in fact, is the limit of $\log_{\widetilde{T}^*}$ when $\widetilde{T}^*$ tends to T^* from the α-stratum. In addition, we modify the definition (7) of $\chi_{T^*, T}(\boldsymbol{e}_i)$ to be $\pi^{-1}_{T^*, T}(\boldsymbol{e}_\alpha)$ when, and only when, the $\boldsymbol{v}^*_i$ in (7) contains t^*_α and is 1-dimensional. With this modification and by noting that the argument leading to Lemma 2, as well as its result, still hold when T^* lies in a stratum of co-dimension one, the results of Theorem 1 and Lemma 4 can be extended to obtain the expression for $F_\alpha \circ \log_{T^*}$ and its derivative, and the analogues with β or γ replacing α, when the necessary care is taken of which stratum is to contain the initial geodesic. Moreover,

$$\int_{\boldsymbol{T}_{m+2}} \langle F_\alpha \circ \log_{T^*}(T), \boldsymbol{e}_\alpha \rangle\, \mathrm{d}\mu(T) = I_\alpha - I_\beta - I_\gamma. \tag{16}$$

These observations lead to the following lemma which extends the results obtained in Hotz et al. (2013) for open books and in Barden et al. (2013) for $\boldsymbol{T}_4$ and relates the Fréchet means of large samples avoiding a stratum to the strict-positivity of the derivative of the Fréchet function along the corresponding orthogonal direction to the tangent space to $\mathcal{O}(\Sigma)$.

Lemma 6 *Let T^* be the Fréchet mean of a given probability measure μ on $\boldsymbol{T}_{m+2}$, and lie in a stratum $\mathcal{O}(\Sigma)$ of co-dimension one. Assume that $\mu(\mathcal{D}_{T^*}) = 0$, where $\mathcal{D}_{T^*}$ is defined in Definition 2, and that, at T^*, $I_\alpha < I_\beta + I_\gamma$. If $\{T_i : i \geqslant 1\}$ is a sequence of iid random variables in $\boldsymbol{T}_{m+2}$ with probability measure μ then, for all sufficiently random large n, the sample Fréchet mean $\hat{T}_n$ of $T_1, \ldots, T_n$ cannot lie in the α-stratum.*

Proof Since $\hat{T}_n$ converges to T^* a.s. as n tends to infinity (cf. Ziezold 1977) we only need to show that, for all sufficiently large n, $\hat{T}_n$ cannot lie in the neighbourhood of T^*, restricted to the α-stratum.

Consider the probability measure μ_α induced from μ by $F_\alpha \circ \log_{T^*}$ on the Euclidean space. Then, under the given conditions, it follows from (16) that the Euclidean mean of μ_α lies on the open half of the Euclidean space complement to the page tangent to the α-stratum (cf. also Hotz et al. (2013)). Thus, for all sufficiently large n, the Euclidean mean of the induced random variables $F_\alpha \circ \log_{T^*}(T_1), \ldots, F_\alpha \circ \log_{T^*}(T_n)$,

$$\hat{T}_n^\alpha = \frac{1}{n} \sum_{i=1}^n F_\alpha \circ \log_{T^*}(T_i),$$

does not lie in the closed half of this Euclidean space where the page tangent to the α-stratum lies. This implies that, for all sufficiently large n,

$$\langle \hat{T}_n^\alpha, \boldsymbol{e}_\alpha \rangle < 0. \tag{17}$$

If it were possible that, for arbitrarily large n, $\hat{T}_n$ lies in the α-stratum, we could obtain a contradiction. Firstly, noting the observations prior to the lemma and following the arguments of the proof for Lemma 4, for all sufficiently large n, we have

$$\frac{1}{n} \sum_{i=1}^n \Phi_{\hat{T}_n}(T_i) = \frac{1}{n} \sum_{i=1}^n F_\alpha \circ \Phi_{T^*}(T_i) + (\hat{T}_n - T^*) \frac{1}{n} \sum_{i=1}^n M_{T^*}(T_i) \\ + o(\|\hat{T}_n - T^*\|) \frac{1}{n} \sum_{i=1}^n \|T_i\|, \tag{18}$$

where $M_{T^*}(T)$ is given by (11) and $F_\alpha \circ \Phi_{T^*} = F_\alpha \circ \log_{T^*} + T^*$. However, on the one hand, since $\frac{1}{n} \sum_{i=1}^n \Phi_{\hat{T}_n}(T_i) = \hat{T}_n$ and since $\hat{T}_n$ lies in the α-stratum, $\langle \hat{T}_n, \boldsymbol{e}_\alpha \rangle > 0$, so that

$$\left\langle \frac{1}{n} \sum_{i=1}^n \Phi_{\hat{T}_n}(T_i), \boldsymbol{e}_\alpha \right\rangle > 0. \tag{19}$$

While, on the other hand, it follows from $\langle T^*, \boldsymbol{e}_\alpha \rangle = 0$ and from (17) that

$$\left\langle \frac{1}{n} \sum_{i=1}^n F_\alpha \circ \Phi_{T^*}(T_i), \boldsymbol{e}_\alpha \right\rangle = \langle \hat{T}_n^\alpha, \boldsymbol{e}_\alpha \rangle < 0. \tag{20}$$

It can also be checked that

$$M_{T^*}(T_i)\, \boldsymbol{e}_\alpha = \frac{v_i^\alpha}{\|\boldsymbol{v}_{i,s}^*\|} \boldsymbol{e}_\alpha,$$

where $v_i^\alpha = v_{i,s}$, if t_α corresponds to a coordinate of $\boldsymbol{v}_{i,s}^*$ and if the dimension of $\boldsymbol{v}_{i,s}^*$ is greater than one, and $v_i^\alpha = 0$ otherwise. Then, since $v_i^\alpha \leqslant 0$, for each i

$$\langle (\hat{T}_n - T^*)\, M_{T^*}(T_i),\, \boldsymbol{e}_\alpha \rangle = v_i^\alpha \langle \hat{T}_n, \boldsymbol{e}_\alpha \rangle \leqslant 0. \tag{21}$$

Equations (20) and (21) together imply that, for all sufficiently large n, the e_α-component of the right hand side of (18) is negative, which contradicts (19). $\quad\square$

With the result of Lemma 6, we now have the limiting distribution of the sample Fréchet means on $\boldsymbol{T}_{m+2}$ given by the next theorem, which is the generalisation of the result for $\boldsymbol{T}_4$ given in Theorem 2 in Barden et al. (2013). In particular, it shows that the limiting distribution can take any of four possible forms, all related to a Gaussian distribution, depending on the number of the strictly positive derivatives of the Fréchet function along the three directions orthogonal to the tangent space to $\mathcal{O}(\Sigma)$. For clarity, we have assumed in the following that the coordinates $(t_i, t_2, \ldots, t_m)$, $i = \alpha, \beta, \gamma$, discussed at the beginning of the section are all in the canonical order, so that they give the coordinates for trees in each of the three strata. Otherwise, a further permutation of the coordinates, which we have suppressed, will be necessary to bring them into canonical order and so to validate the result.

Theorem 3 *Let T^* in a stratum $\mathcal{O}(\Sigma)$ of co-dimension one be the Fréchet mean of a given probability measure μ on $\boldsymbol{T}_{m+2}$. Assume that $\mu(\mathcal{D}_{T^*}) = 0$, where $\mathcal{D}_{T^*}$ is defined in Definition 2. Let further $\{T_i \,:\, i \geqslant 1\}$ be a sequence of iid random variables in $\boldsymbol{T}_{m+2}$ with probability measure μ and write $\hat{T}_n$ for the sample Fréchet mean of $T_1, \ldots, T_n$.*

(a) *If all three inequalities in (14) are strict then, for all sufficiently large n, $\hat{T}_n$ will lie in the stratum $\mathcal{O}(\Sigma)$ and the sequence $\sqrt{n}\{(\hat{t}_2^n, \ldots, \hat{t}_m^n) - (t_2^*, \ldots, t_m^*)\}$ of the coordinates of $\sqrt{n}\{\hat{T}_n - T^*\}$ on the spine will converge in distribution to $N(0, A_s^\top V_s A_s)$ as $n \to \infty$, where V_s is the covariance matrix of the random variable $\log_{T^*}^s(T_1)$, $A_s = P_s^\top A P_s$, P_s is the projection matrix to the subspace of $\mathbb{R}^m$ with the first coordinate removed and A is as given in (13).*

(b) *If the first inequality in (14) is an equality and the other two are strict then, for all sufficiently large n, $\hat{T}_n$ will lie in the α-stratum and*

$$\sqrt{n}\{\hat{T}_n - T^*\} \overset{d}{\longrightarrow} (\max\{0, \eta_1\}, \eta_2, \ldots, \eta_m), \quad as\ n \to \infty,$$

where $(\eta_1, \ldots, \eta_m) \sim N(0, A^\top V A)$, V is the covariance matrix of $F_\alpha \circ \log_{T^*}(T_1)$ and A is as in (13) with $t_1^* = 0$.

(c) *If the first two inequalities in (14) are equalities and the third is strict then, for all sufficiently large n, $\hat{T}_n$ will lie either in the α-stratum or in the β-stratum and the limiting distribution of $\sqrt{n}\{\hat{T}_n - T^*\}$, as $n \to \infty$, will take the same form as that*

of $(\eta_1, \ldots, \eta_m)$ *above, where the coordinates of* $\hat{T}_n$ *are taken as* $(\hat{t}_\alpha^n, \hat{t}_2^n, \ldots, \hat{t}_m^n)$, *respectively* $(-\hat{t}_\beta^n, \hat{t}_2^n, \ldots, \hat{t}_m^n)$, *if* $\hat{T}_n$ *is in the* α*-stratum, respectively the* β*-stratum.*

(d) *If all the equalities in* (14) *are actually equalities, then we have the same result as in* (a).

Proof (a) By Lemma 6, when n is sufficiently large, $\hat{T}_n$ must lie in the stratum $\mathcal{O}(\Sigma)$ of co-dimension one so that it has zero first coordinate, i.e. $\hat{T}_n = (0, \hat{t}_2^n, \ldots, \hat{t}_m^n)$. Noting that $F_\alpha \circ \log_{\hat{T}_n}^s = \log_{\hat{T}_n}^s$, the result (15) of Lemma 5 shows that $\hat{t}_i^n$, $i = 2, \ldots, m$, are the respective coordinates of $\frac{1}{n} \sum_{i=1}^n F_\alpha \circ \Phi_{\hat{T}_n}(T_i)$, the sample Euclidean mean of $F_\alpha \circ \Phi_{\hat{T}_n}(T_1), \ldots, F_\alpha \circ \Phi_{\hat{T}_n}(T_n)$. Then, a modification of the proof of Theorem 2 to restrict it to the relevant coordinates of $\{F_\alpha \circ \Phi_{\hat{T}_n}(T_i) : i \geqslant 1\}$ gives the required limiting distribution of $\sqrt{n}\{(\hat{t}_2^n, \ldots, \hat{t}_m^n) - (t_2^*, \ldots, t_m^*)\}$.

(b) We deduce from the assumed strict inequalities, from (15) and (16) and from Lemma 6 that T^* is the Euclidean mean of $F_\alpha \circ \Phi_{T^*}(T_1)$ and that, when n is sufficiently large, $\hat{T}_n$ can only lie in the closure of the α-stratum, so that it has coordinates $\hat{T}_n = (\hat{t}_\alpha^n, \hat{t}_2^n, \cdots \hat{t}_m^n)$.

Write

$$\tilde{F}_\alpha \circ \Phi_{\hat{T}_n}(T) = \begin{cases} \Phi_{\hat{T}_n}(T) & \text{if } \hat{t}_\alpha^n > 0 \\ F_\alpha \circ \Phi_{\hat{T}_n}(T) & \text{if } \hat{t}_\alpha^n = 0. \end{cases} \tag{22}$$

Then, $\tilde{F}_\alpha \circ \Phi_{\hat{T}_n}(T)$ lies in $\mathbb{R}^m$ and, by (15), $\hat{t}_j^n$, $j = 2, \ldots, m$, are the respective coordinates of $\frac{1}{n} \sum_{i=1}^n \tilde{F}_\alpha \circ \Phi_{\hat{T}_n}(T_i)$. To see relationship between $\frac{1}{n} \sum_{i=1}^n \tilde{F}_\alpha \circ \Phi_{\hat{T}_n}(T_i)$ and $\hat{t}_\alpha^n$, we note that, if $\hat{t}_\alpha^n > 0$,

$$\frac{1}{n} \sum_{i=1}^n \tilde{F}_\alpha \circ \Phi_{\hat{T}_n}(T_i) = \frac{1}{n} \sum_{i=1}^n \Phi_{\hat{T}_n}(T_i) = \hat{T}_n, \tag{23}$$

where the first equality follows from the definition of $\tilde{F}_\alpha \circ \Phi_{\hat{T}_n}(T)$ and the second follows from Lemma 3 as $\hat{T}_n$ lies in a top-dimensional stratum. Hence,

$$\left\langle \frac{1}{n} \sum_{i=1}^n \tilde{F}_\alpha \circ \Phi_{\hat{T}_n}(T_i), \, e_\alpha \right\rangle = \langle \hat{T}_n, e_\alpha \rangle = \hat{t}_\alpha^n.$$

On the other hand, if $\hat{t}_n^\alpha = 0$, then $\hat{T}_n$ lies in $\mathcal{O}(\Sigma)$ and

$$\frac{1}{n} \sum_{i=1}^n \tilde{F}_\alpha \circ \Phi_{\hat{T}_n}(T_i) = \frac{1}{n} \sum_{i=1}^n F_\alpha \circ \Phi_{\hat{T}_n}(T_i).$$

Applying Lemma 5 and (16) to the empirical distribution centred on $T_1, \ldots, T_n$ with equal weights $1/n$, we also have

$$\left\langle \frac{1}{n} \sum_{i=1}^{n} \tilde{F}_\alpha \circ \Phi_{\hat{T}_n}(T_i), \, \boldsymbol{e}_\alpha \right\rangle \leqslant 0.$$

Thus,

$$\hat{t}_\alpha^n = \max \left\{ 0, \, \left\langle \frac{1}{n} \sum_{i=1}^{n} \tilde{F}_\alpha \circ \Phi_{\hat{T}_n}(T_i), \, \boldsymbol{e}_\alpha \right\rangle \right\}.$$

Now, similarly to the proofs of Theorem 2 and Lemma 6, the observations prior to Lemma 6 imply that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \tilde{F}_\alpha \circ \Phi_{\hat{T}_n}(T_i) - T^* \right\}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ F_\alpha \circ \Phi_{T^*}(T_i) - T^* \right\} + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \tilde{F}_\alpha \circ \Phi_{\hat{T}_n}(T_i) - F_\alpha \circ \Phi_{T^*}(T_i) \right\}$$

$$\approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ F_\alpha \circ \Phi_{T^*}(T_i) - T^* \right\} + \frac{1}{\sqrt{n}} (\hat{T}_n - T^*) \sum_{i=1}^{n} M_{T^*}(T_i) \tag{24}$$

$$+ o(\|\hat{T}_n - T^*\|) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \|T_i\|,$$

where $M_{T^*}(T)$ is given by (11). Since the first coordinate of T^* is zero, so too are the entries, except for the diagonal one, in the first row and column of $M_{T^*}(T)$ and so also are the corresponding entries in the matrix A. Moreover, noting the comments following Lemma 4 and the definition of $M^\dagger$ prior to that lemma, we see that the first diagonal entry of $M_{T^*}(T)$ is always non-positive. Thus, the first diagonal entry of A must be positive, so that this is also the case for $\{I - \frac{1}{n} \sum_{i=1}^{n} M_{T^*}(T_i)\}^{-1}$, when n is sufficiently large.

Thus, when $\hat{t}_\alpha^n > 0$, it follows from (23) and (24) that

$$\sqrt{n}(\hat{T}_n - T^*) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ F_\alpha \circ \Phi_{T^*}(T_i) - T^* \right\} \left\{ I - \frac{1}{n} \sum_{i=1}^{n} M_{T^*}(T_i) \right\}^{-1}$$

$$+ o(\|\hat{T}_n - T^*\|) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \|T_i\|.$$

In particular, for all sufficiently large n, the first coordinate of the random vector given by the first term on the right is positive. When $\hat{t}_\alpha^n = 0$, the above approximation

still holds except for the first coordinate. In that case, $\langle (T_n \overset{\wedge}{-} T^*) M_{T^*}(T_i), \boldsymbol{e}_\alpha \rangle = 0$, following from the form of $M_{T^*}(T)$ noted above, and by (24), for sufficiently large n,

$$\left\langle \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ F_\alpha \circ \Phi_{T^*}(T_i) - T^* \right\}, \boldsymbol{e}_\alpha \right\rangle \leqslant 0$$

up to higher order terms, which is equivalent to

$$\left\langle \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ F_\alpha \circ \Phi_{T^*}(T_i) - T^* \right\} \left\{ I - \frac{1}{n} \sum_{i=1}^{n} M_{T^*}(T_i) \right\}^{-1}, \boldsymbol{e}_\alpha \right\rangle \leqslant 0$$

up to higher order terms. Hence, for sufficiently large n, we have

$$\sqrt{n}(\hat{T}_n - T^*) \approx (\max\{0, \eta_1^n\}, \eta_2^n, \ldots, \eta_m^n),$$

where

$$(\eta_1^n, \eta_2^n, \ldots, \eta_m^n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ F_\alpha \circ \Phi_{T^*}(T_i) - T^* \right\} \left\{ I - \frac{1}{n} \sum_{i=1}^{n} M_{T^*}(T_i) \right\}^{-1},$$

so that the required result follows from a similar argument to that of the proof for Theorem 2.

(c) In this case, it follows from Lemma 5 that T^* is the Euclidean mean both of $F_\alpha \circ \Phi_{T^*}(T_1)$ and of $F_\beta \circ \Phi_{T^*}(T_1)$. Moreover, the integral I_γ becomes zero and so, since the integrand is non-negative, the support of the measure on the tangent book at T^* induced by μ is contained in the union of the leaves tangent to the α- and β-strata together with the spine.

It is now more convenient to represent the union of the α- and β-strata by coordinates in the two orthants $\{(t_1, \ldots, t_m) : t_2, \ldots, t_m \geqslant 0\}$ of $\mathbb{R}^m$. For this, we map:

$$(t_\alpha, t_2, \ldots, t_m) \mapsto (t_\alpha, t_2, \ldots, t_m) \text{ and } (t_\beta, t_2, \ldots, t_m) \mapsto (t_\beta, t_2, \ldots, t_m)R,$$

where $R = \text{diag}\{-1, I_{m-1}\}$. Similarly, we define maps $\tilde{\Phi}_{(t_1, \ldots, t_m)}(T)$ to accord with this by $\tilde{\Phi}_{(-t_\beta, t_2, \ldots, t_m)}(T) = \Phi_{(t_\beta, t_2, \ldots, t_m)}(T)R$, while $\tilde{\Phi}_{(t_\alpha, t_2, \ldots, t_m)} = \Phi_{(t_\alpha, t_2, \ldots, t_m)}$. Since $\Phi_{(0_\alpha, t_2, \ldots, t_m)}(T) = \Phi_{(0_\beta, t_2, \ldots, t_m)}(T)R$, the map $\tilde{\Phi}$ is indeed a.s. well defined for points $(0, t_2, \ldots, t_m)$. Clearly,

$$\tilde{\Phi}_{(t_1, t_2 \ldots, t_m)}(T) = \begin{cases} \tilde{F}_\alpha \circ \Phi_{(t_1, t_2, \ldots, t_m)}(T) & \text{if } t_1 \geqslant 0 \\ \tilde{F}_\beta \circ \Phi_{(-t_1, t_2, \ldots, t_m)}(T)R & \text{if } t_1 \leqslant 0 \end{cases}$$

where $\tilde{F}_\alpha$, similarly $\tilde{F}_\beta$, is defined by (22). Under this new coordinate system, since $F_\alpha \circ \Phi_{T^*}(T_1) = F_\beta \circ \Phi_{T^*}(T_1) R$ a.s., we have in particular that

$$T^* = \int_{T_{m+2}} \tilde{\Phi}_{(0, t_2^*, \ldots, t_m^*)}(T) \, d\mu(T). \tag{25}$$

By Lemma 6, the given assumption also implies that, for sufficiently large n, $\hat{T}_n$ will a.s. lie either in the α-stratum or in the β-stratum. If $\hat{T}_n$ lies in the α-stratum, then $t_\alpha^n > 0$ and

$$(t_n^{\hat{\alpha}}, t_n^{\hat{2}}, \ldots, t_m^{\hat{n}}) = \frac{1}{n} \sum_{i=1}^{n} \Phi_{\hat{T}_n}(T_i) = \frac{1}{n} \sum_{i=1}^{n} \tilde{\Phi}_{(t_n^{\hat{\alpha}}, t_n^{\hat{2}}, \ldots, t_n^{\hat{m}})}(T_i) \tag{26}$$

and, if $\hat{T}_n$ lies in the β-stratum with (original) coordinates $\hat{T}_n = (-t_\beta^{\hat{n}}, -t_2^{\hat{n}}, \ldots, t_m^{\hat{n}})$, then

$$(-t_\beta^{\hat{n}}, -t_2^{\hat{n}}, \ldots, -t_m^{\hat{n}}) = \frac{1}{n} \sum_{i=1}^{n} \Phi_{\hat{T}_n}(T_i) R = \frac{1}{n} \sum_{i=1}^{n} \tilde{\Phi}_{(-t_\beta^{\hat{n}}, -t_2^{\hat{n}}, \ldots, -t_m^{\hat{n}})}(T_i). \tag{27}$$

If $\hat{T}_n$ lies on the stratum $\mathcal{O}(\Sigma)$ of co-dimension one then, by applying the argument in (b) to both $t_\alpha^{\hat{n}} = 0$ and $t_\beta^{\hat{n}} = 0$, we also have

$$(0, -t_2^{\hat{n}}, \ldots, -t_m^{\hat{n}}) = \frac{1}{n} \sum_{i=1}^{n} \Phi_{(0_\alpha, -t_2^{\hat{n}}, \ldots, -t_m^{\hat{n}})}(T_i) = \frac{1}{n} \sum_{i=1}^{n} \tilde{\Phi}_{(0, t_2^{\hat{n}}, \ldots, t_m^{\hat{n}})}(T_i) \quad \text{a.s..} \tag{28}$$

Recalling that, under the new coordinate system,

$$\hat{T}_n \equiv \begin{cases} (t_n^{\hat{\alpha}}, t_n^{\hat{2}}, \ldots, t_n^{\hat{m}}) & \text{if } \hat{T}_n \text{ is in the } \alpha\text{-stratum} \\ (-t_\beta^{\hat{n}}, -t_2^{\hat{n}}, \ldots, -t_m^{\hat{n}}) & \text{if } \hat{T}_n \text{ is in the } \beta\text{-stratum} \end{cases}$$

we have by (25), (26), (27) and (28) that, in terms of the new coordinates,

$$\sqrt{n}\{\hat{T}_n - T^*\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \tilde{\Phi}_{T^*}(T_i) - (0, t_2^*, \ldots, t_m^*) \right\}$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \tilde{\Phi}_{\hat{T}_n}(T_i) - \tilde{\Phi}_{T^*}(T_i) \right\}.$$

Hence, since (24) still holds under this new coordinate system when $\tilde{F}_\alpha \circ \Phi_{\hat{T}_n}$ and $F_\alpha \circ \Phi_{T^*}$ there are replaced by $\tilde{\Phi}_{\hat{T}_n}$ and $\tilde{\Phi}_{T^*}$, respectively, a similar argument to that

of the proof for Theorem 2 shows that the central limit theorem now takes the required form.

(d) Noting that all integrands in (14) are non-negative, the three equalities will together imply that $\log_{T^*}(T_1) = \log_{T^*}^s(T_1)$ a.s., so that for $i = \alpha, \beta, \gamma$

$$\left\langle \sum_{i=1}^n \left\{ F_i \circ \Phi_{T^*}(T_i) - T^* \right\}, \, \boldsymbol{e}_i \right\rangle = 0 \quad \text{a.s..} \tag{29}$$

On the other hand, if it were possible that, for arbitrarily large n, $\hat{T}_n$ lies in one of the α- β- or γ-strata, say the α-stratum, then $\langle \hat{T}_n - T^*, \boldsymbol{e}_\alpha \rangle > 0$. On the other hand, since

$$\hat{T}_n - T^* = \frac{1}{n} \sum_{i=1}^n \left\{ \Phi_{\hat{T}_n}(T_i) - T^* \right\}$$

$$\approx \frac{1}{n} \sum_{i=1}^n \left\{ F_\alpha \circ \Phi_{T^*}(T_i) - T^* \right\} + \frac{1}{n} (\hat{T}_n - T^*) \sum_{i=1}^n M_{T^*}(T_i),$$

and since, as noted in (b), the first diagonal element of $M_{T^*}(T_i)$ is non-positive and the remaining entries in the first row and column of $M_{T^*}(T_i)$ are all zero, we have by (29) that

$$\langle \hat{T}_n - T^*, \, \boldsymbol{e}_\alpha \rangle \approx \left\langle \frac{1}{n} (\hat{T}_n - T^*) \sum_{i=1}^n M_{T^*}(T_i), \, \boldsymbol{e}_\alpha \right\rangle \leqslant 0.$$

This contradiction implies that, for all sufficiently large n, $\hat{T}_n$ must lie in the stratum $\mathcal{O}(\Sigma)$ of co-dimension one. Thus, the argument for (a) implies that, when the inequalities in (14) are all equalities, the central limit theorem for the sample Fréchet means takes the same form as that when the three inequalities are all strict. □

Similar to the note at the end of the previous section, one can also consider the distribution μ', induced by $\log_{T^*}$ from μ on the tangent book of $\boldsymbol{T}_{m+2}$ at T^*. Then, one can apply the result of Hotz et al. (2013) to obtain the limiting distribution of the sample Fréchet means of μ'. Again, although the limiting distribution obtained in this way retains the local topological feature of the space, the influence of the global topological structure is lost. More importantly, since there is no clear relationship between the sample Fréchet means of μ and μ', the limiting distribution for the former cannot be easily deduced from that for the latter.

5 Strata of higher co-dimension

The structure of tree space in the neighbourhood of a stratum of higher co-dimension is basically similar to, but in detail rather more complex than, that of a stratum of co-dimension one. For example, a stratum σ of co-dimension l, where $2 \leqslant l \leqslant m$,

corresponds to a set of $m - l$ mutually compatible edge-types. It arises as a boundary $(m - l)$-dimensional face of a stratum τ of co-dimension l', where $0 \leqslant l' < l$ and when the internal edges of the trees in σ are a particular subset of $m - l$ of the internal edges of the trees in τ. For this situation, we say that σ bounds τ and τ co-bounds σ.

Recall from the previous section that the tangent cone to T_{m+2} at a tree T in σ consists of all initial tangent vectors to smooth curves starting from T, the smoothness only being one-sided at T. For simplicity assume, without loss of generality, that under the isometric embedding of T_{m+2} in $\mathbb{R}^M$ all trees in σ have zero for their first l coordinates. Then, the tangent cone at T has a stratification analogous to that of T_{m+2} itself in the neighbourhood of T: for each stratum τ of co-dimension l' that co-bounds σ in T_{m+2} there is a stratum $(\mathbb{R}_\tau^{l-l'})_+ \times \mathbb{R}^{m-l}$ in the tangent cone at T, which may be identified with a subset of the full tangent space of $\mathbb{R}^M$ at T, where $\mathbb{R}^{m-l}$ is the (full) tangent subspace to σ at T and $(\mathbb{R}_\tau^{l-l'})_+$ is the orthant determined by the edge-types that have positive length in τ but zero length in σ. For example, the cone point in T_4 is a stratum of co-dimension two. Its tangent cone can be identified with T_4 itself. This rather involved structure of the tangent cone results in a much more complicated description of the log map and, consequently, of its behaviour. Nevertheless, with the above conventions, it is possible to generalise our expression for the log map to this wider context and hence to obtain analogues of Theorem 1 as well as Lemma 4. These results can then be used to describe, in a fashion similar to those of Lemmas 5 and 6, certain limiting behaviour of sample Fréchet means when their limit lies in a stratum of higher co-dimension. For example, the limiting behaviour of sample Fréchet means in T_4 when the true Fréchet mean lies at the cone point has been studied in Barden et al. (2013). The picture given there is incomplete and, although those results can be further refined and improved, it is clear that a complete description of the limiting behaviour of sample Fréchet means in the wider context is still a challenge and the global topological structure of the space will play a crucial role.

References

Bacak, M. (2014). Computing medians and means in hadamard spaces. *SIAM Journal on Optimization*, *24*, 1542–1566.

Barden, D., Le, H., Owen, M. (2013). Central limit theorems for Fréchet means in the space of phylogenetic trees. *Electronic Journal of Probability*, **18**(25).

Basrak, B. (2010). Limit theorems for the inductive mean on metric trees. *Journal of Applied Probability*, *47*, 1136–1149.

Bhattacharya, R., Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds-II. *Annals of Statistics*, *33*, 1225–1259.

Bhattacharya, R., Patrangenaru, V. (2014). Statistics on manifolds and landmarks based image analysis: A nonparametric theory with applications. *Journal of Statistical Planning and Inference*, *145*, 1–22.

Billera, L., Holmes, S., Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, *27*, 733–767.

Bridson, M., Haefliger, A. (1999). *Metric Spaces of Non-positive Curvature*. Berlin: Springer.

Dryden, I., Mardia, K. (1998). *Statistical Shape Analysis*. Chichester: Wiley.

Dryden, I., Le, H., Preston, S., Wood, A. (2014). Mean shapes, projections and intrinsic limiting distributions. *Journal of Statistical Planning and Inference*, **145**, 25–32.

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 *Information Processing in Medical Imaging,* 23*rd International Conference, IPMI* (pp. 74–85)

Holmes, S. (2003). Statistics for phylogenetic trees. *Theoretical Population Biology*, *63*, 17–32.

Hotz, T., Huckemann, S., Le, H., Marron, J., Mattingly, J., Miller, E., et al. (2013). Sticky central limit theorems on open books. *Annals of Applied Probability*, *23*, 2238–2258.

Kendall, W., Le, H. (2011). Limit theorems for empirical fréchet means of independent and non-identically distributed manifold-valued random variables. *Brazilian Journal of Probability and Statistics*, *25*, 323–352.

Miller, E., Owen, M., Provan, S. (2015). Polyhedral computational geometry for averaging metric phylogenetic trees. *Advances in Applied Mathematics*, *68*, 51–91.

Nye, T. (2011). Principal components analysis in the space of phylogenetic trees. *Annals of Statistics*, *39*, 2716–2739.

Nye, T. (2014). An algorithm for constructing principal geodesics in phylogenetic treespace. *Transactions on Computational Biology and Bioinformatics*, *11*, 304–315.

Owen, M. (2011). Computing geodesic distances in tree space. *SIAM Journal on Discrete Mathematics*, *25*, 1506–1529.

Owen, M., Provan, J. (2011). A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *8*, 2–13.

Schroder, E. (1870). Vier combinatorische probleme. *Zeitschrift fur Mathematik und Physik*, *15*, 361–376.

Vogtmann, K. (2007). Geodesics in the space of trees. www.math.cornell.edu/~vogtmann/papers/TreeGeodesicss/index.html.

Willis, A. (2016). Confidence sets for phylogenetic trees. arXiv:1607.08288v1 [stat.ME].

Ziezold, H. (1977). On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. In *Transactions of the seventh Prague Conference on Information Theory, Statistical Decision Functions and Random Processes A* (pp. 591–602).