

Geometry-driven Bayesian Inference for Ultrametric Covariance Matrices

Tsung-Hung Yao*

Zhenke Wu†

Karthik Bharath‡

Veerabhadran Baladandayuthapani†

Abstract

Ultrametric matrices arise as covariance matrices in latent tree models for multivariate data with hierarchically correlated components. As a parameter space in a model, the set of ultrametric matrices is neither convex nor a smooth manifold, and focus in literature has hitherto mainly been restricted to estimation through projections and relaxation-based techniques. Leveraging the link between an ultrametric matrix and a rooted tree, we equip the set of ultrametric matrices with a convenient geometry based on the well-known geometry of phylogenetic trees, whose attractive properties (e.g. unique geodesics and Fréchet means) the set of ultrametric matrices inherits. This results in a novel representation of an ultrametric matrix by coordinates of the tree space, which we then use to define a class of Markovian and consistent prior distributions on the set of ultrametric matrices in a Bayesian model, and develop an efficient algorithm to sample from the posterior distribution that generates updates by making intrinsic local moves along geodesics within the set of ultrametric matrices. In simulation studies, our proposed algorithm restores the underlying matrices with posterior samples that recover the tree topology with a high frequency of true topology and generate element-wise credible intervals with a high nominal coverage rate. We use the proposed algorithm on the pre-clinical cancer data to investigate the mechanism similarity by constructing the underlying treatment tree and identify treatments with high mechanism similarity also target correlated pathways in biological literature.

1 Introduction

Structured covariance and inverse covariance matrices abound in statistical models for data exhibiting specific forms of dependencies (Lauritzen, 1996). The focus of this work is on Gaussian models on trees parameterized by a class of structure covariance matrices known as ultrametric matrices. The class constitutes a special case of covariance graph models first studied by Cox and Wermuth (1993) with additional constraints on the covariance matrix, and are intimately related to multivariate totally positive of order two distributions (Karlin and Rinott, 1983; Lauritzen et al., 2019). Gaussian models parameterized by ultrametric matrices are also referred to as Gaussian tree models (Felsenstein, 2003) and latent tree models (Wang et al., 2008; Choi et al., 2011), and are widely used in various applications, including psychology (Lauritzen et al., 2019), cancer biology (Yao et al., 2023), and finance (Agrawal et al., 2020); see Zwiernik (2016) for a comprehensive survey of latent tree models. Ultrametric matrices first appeared in the potential theory of finite state Markov chains (Dellacherie et al., 2014), and have also appeared in the literature as tree-structured matrices (McCullagh, 2006; Bravo et al., 2009).

*Department of Biostatistics, The University of Texas MD Anderson Cancer Center, USA

†Department of Biostatistics, University of Michigan, USA

‡School of Mathematical Sciences, University of Nottingham, UK

The set \mathcal{U}_p of $p \times p$ ultrametric matrices is characterized by semialgebraic inequalities, and, aided by techniques from combinatorics and polyhedral geometry, there is a substantial body of work that focuses on their estimation (e.g., Zwiernik et al., 2017; Sturmfels et al., 2021). Work on developing inferential methods for them, however, has not experienced the same level of progress. Central to the set of challenges that are stymieing progress for inference is the non-trivial geometry of \mathcal{U}_p . This can be intuited through the link between an ultrametric matrix $\Sigma^T \in \mathcal{U}_p$ and a rooted tree T with p leaves: the tree T possesses both topological (graph structure) and geometric (edge lengths) information that is encoded in Σ^T , and varying T by altering both sources of information varies the structure of Σ^T (e.g., pattern of zeros; magnitude of entries) in a non-trivial manner.

For ultrametric matrices that are *correlation* matrices with entries in $[0, 1]$, a subset of \mathcal{U}_p , the inequalities in Definition 1 engender an embedding within symmetric matrices through their link to tree-based metrics on $\{1, 2, \dots, p\}$ (Buneman, 1974), leading to the embedded geometry of the edge-product space or phylogenetic oranges (Kim, 2000; Moulton and Steel, 2004). The recently developed Wald space for phylogenetic trees by Garba et al. (2021) and Lueg et al. (2022) is built on this embedded geometry. The embedded geometry of ultrametric correlation matrices was used Shiers et al. (2016) to assess adequacy of a Gaussian tree model for data, however with an Inverse Wishart prior with support on all positive definite matrices. Based on their work, Leung and Drton (2018) developed a Bootstrap-based procedure for testing for a more general class of latent tree models. We are unaware of a similar embedding for \mathcal{U}_p based on its semialgebraic characterisation.

There are two important desiderata for sampling-based inference for Bayesian models with parameter space \mathcal{U}_p that possesses non-trivial geometry: (i) definition of and sampling from a flexible prior distribution on \mathcal{U}_p ; (ii) efficient exploration of \mathcal{U}_p by algorithms for posterior sampling that use the *intrinsic geometry* of \mathcal{U}_p , without requiring expensive projections onto \mathcal{U}_p from a larger space. An intrinsic metric geometry of \mathcal{U}_p that provides tools (e.g., geodesics, distances) to characterise and quantify how changes in the tree T manifests in corresponding changes in Σ^T , and vice versa, will constitute a significant step towards designing inferential methods that satisfy (i) and (ii). Such a geometric framework is unavailable in the literature. The goal of this paper is to fill this gap, and demonstrate the framework's benefits in performing inference for ultrametric matrices in Bayesian Gaussian latent tree models.

Our approach is based on pulling back the intrinsic *stratified geometry* of the well-studied phylogenetic treespace (Billera et al., 2001) onto \mathcal{U}_p . The CAT(0) geometry of the treespace, the useful space for the development of several statistical tools for tree-structured data, is then inherited by \mathcal{U}_p . The result is a novel representation of an ultrametric matrix Σ^T as a linear combination of a set of basis matrices, which characterizes the set of possible topologies for the tree T . The representation complements the well-known decomposition of Σ^T by Nabben and Varga (1994) in the literature on ultrametric matrices, and may be of independent interest. The coordinates thus obtained for Σ^T elucidate on how topology of the tree T is encoded via choice of the stratum and edge lengths determine its location within a stratum. Shortest paths, or geodesics, on \mathcal{U}_p between two ultrametric matrices then signify geometric and topological deformations between the corresponding trees. Of key inferential consequence is the fact that the geometry enables us to:

- (i) specify a class of Markovian and consistent prior distributions on \mathcal{U}_p that allows for decoupled

control over induced distributions on tree topology and edge lengths;

- (ii) design an efficient algorithm for posterior inference on \mathcal{U}_p in a latent tree model that makes updates using local moves along geodesics within \mathcal{U}_p , without requiring any projection or relaxation;
- (iii) compute statistical summaries (e.g., sample Fréchet mean) from posterior samples.

Since a tree is also a graph, we note that the class of Gaussian models parameterized by covariance matrices that are ultrametric is subsumed under the class of Gaussian covariance graph models, where (in)dependence between nodes are specified through zeros of the covariance matrix and not its inverse. General conjugate and non-conjugate Bayesian methods developed (Dawid and Lauritzen, 1993; Letac and Massam, 2007; Khare and Rajaratnam, 2011) for such models based on prior distributions on structured covariance matrices with pre-specified zeros can in principle be adapted to the covariances in \mathcal{U}_p . However, information on how topology and geometry of the tree T manifests in Σ^T is not transparent, and it is unclear how one extracts benefits such as (i)-(iii) above from such general frameworks for graphs.

The rest of the paper is organized as follows. Section 2.1 reviews ultrametric matrices and properties of \mathcal{U}_p , and section 2.2 reviews the phylogenetic treespace of Billera et al. (2001). Section 3 develops the intrinsic geometry of \mathcal{U}_p (Theorem 1) and the ensuing representation of an ultrametric matrix (Corollary 1), and discusses how the stratified geometry of \mathcal{U}_p with identified boundaries manifests in the ultrametric matrices (Proposition 1 and Corollary 2). In Section 4, we construct a Markovian and consistent prior on \mathcal{U}_p . Section 5 delineates an algorithm (Algorithm 1) that draws posterior covariances effectively via the intrinsic geometry of \mathcal{U}_p . In Section 6, we present results from extensive simulations that assess quality of recovery of an ultrametric matrix covariance in Gaussian and misspecified models, along with uncertainty quantification. In Section 7, we demonstrate utility of the developed Bayesian algorithm on pre-clinical data obtained from studies on cancer treatment. Section 8 offers concluding remarks and future directions. General purpose code in R with packages and datasets for the proposed method is available at <https://github.com/bayesrx/ultrametricMat>.

2 Preliminaries

In this Section, we introduce and briefly review properties of ultrametric matrices, and the challenges involved in their estimation and inference. One of the main objectives of our work is to equip the set of ultrametric matrices with a geometry that enables sampling-based inference, and our approach is based on establishing a bijection with the phylogenetic treespace by Billera et al. (2001), henceforth known as the BHV space. We briefly review the BHV space and its salient properties.

2.1 Ultrametric matrices

Let \mathcal{S}_p be the set of $p \times p$ real symmetric matrices and $L = \{0, 1, 2, \dots, p\}$.

Definition 1. A matrix $\Sigma^T \in \mathcal{S}_p$ is *strictly ultrametric* if

- (i) every element of Σ^T is non-negative;
- (ii) for all $i \in L$, $\Sigma_{ii}^T > \max\{\Sigma_{ij}^T : j \in L \setminus \{i\}\}$;

(iii) for all $i, j, k \in L$

$$\Sigma_{ij}^T \geq \min\{\Sigma_{ik}^T, \Sigma_{kj}^T\}. \quad (1)$$

Matrices satisfying the definition are referred to as ultrametric since reverse of the inequality in (1) signifies ultrametricity of a metric once elements Σ_{ij}^T are interpreted as the distance between leaves i and j of tree T , defined to be the sum of edge lengths from the root to the most recent common ancestor of i and j . Property (ii) in the definition ensures that the diagonal element is the largest in each row and column, and that such Σ^T are nonsingular (Nabben and Varga, 1994). We denote by \mathcal{U}_p the set of $p \times p$ strictly ultrametric matrices and by $\mathcal{S}_p^+ := \{\Sigma \in \mathcal{S}_p : x^\top \Sigma x > 0, x \in \mathbb{R}^p\}$ the convex cone¹ of symmetric positive definite matrices in $\mathbb{R}^{p \times p}$; note that $\mathcal{U}_p \subset \mathcal{S}_p^+$.

If property (ii) in the definition above is relaxed to “ \geq ”, then Σ^T is referred to an ultrametric matrix (Dellacherie et al., 2014). Denote by \mathcal{U}_p^* the set of ultrametric matrices, which contains singular matrices that lie on the boundary of the non-negative cone $\{\Sigma \in \mathcal{S}_p : x^\top \Sigma x \geq 0, x \in \mathbb{R}^p\}$. We primarily focus on the set \mathcal{U}_p of strictly ultrametric matrices in this work, and sometimes refer to them as ultrametric when the context is clear.

Matrices in \mathcal{U}_p are inverses of the so-called M -matrices (Ostrowski, 1937), matrices with non-positive off-diagonal and positive diagonal entries that are used to model conditional dependencies in Gaussian models (Karlin and Rinott, 1980, 1983). Nabben and Varga (1994) determined the following decomposition of an ultrametric matrix in \mathcal{U}_p^* when studying its relationship with a particular M -matrix:

$$\Sigma^T = \sum_{j=1}^{2p-1} d_j v_j v_j^\top = V D V^\top, \quad (2)$$

where $\{v_j\}$ are p -dimensional binary vectors with values in $\{0, 1\}$ with $v_1 = \mathbf{1}$, the vector of ones, and $\{d_j\}$ are non-negative real numbers. The matrix $V = (v_1, \dots, v_{2p-1})^\top$ is known as the basis matrix of Σ^T with the *partition property* (Bravo et al., 2009) that associates with a rooted tree structure (Nabben and Varga, 1994). Starting from the root of the tree pertaining to v_1 , it encodes a recursive partition of $\{1, \dots, p\}$ such that every column v_i with more than one non-zero elements there exists two other columns v_j, v_k such that $v_i = v_j + v_k$. The matrix V determines topology of the tree T , while the diagonal matrix D with entries d_j stores the edge lengths. The matrix Σ^T is made strictly ultrametric by enforcing d_1 and the d_j 's pertaining to leaf edge lengths to be strictly positive.

The set of M -matrices is a convex subset of \mathcal{S}_p^+ and this was used by Slawski and Hein (2015) to compute the maximum likelihood estimate (MLE) of the (inverse) covariance matrix in a Gaussian graphical model. However, existing approaches to estimation on \mathcal{U}_p are dominated by optimization methods based on projections or relaxations, since the inequalities in Definition 1 are non-convex. For example, Lauritzen et al. (2019) and Agrawal et al. (2020) handle the inequalities by considering a dual problem. On the other hand, ultrametric matrices are a special case of linear covariance matrices introduced by Anderson (1970) and occur as covariance matrices of Brownian tree models introduced in phylogenetics (Felsenstein, 1973), and more generally as covariance matrices of marginal likelihoods in generative models for hierarchically correlated multivariate data (Neal, 2003; Yao et al., 2023). As a

¹We use A^T with a roman letter “ T ” as superscript to highlight dependence of the matrix A on the tree T , and distinguish this from A^\top to denote its transpose.

special case of linear covariance matrices, the MLE of Σ^T can be computed given the basis matrix V (Zwiernik et al., 2017); this corresponds to fixing the topology of the tree T apriori, and the corresponding set of ultrametric matrices can be identified with a simplicial cone within a spectrahedron in \mathcal{S}_p (Sturmfel et al., 2021). However, the basis matrix V is part of the parameter space and needs to be estimated.

Despite being a subset of the convex set of inverse M -matrices, the set \mathcal{U}_p is neither convex nor a smooth manifold (McCullagh, 2006). However, to carry out sampling-based inference on \mathcal{U}_p under a Bayesian setting, it is important to enable algorithms to make *intrinsic* local moves that do not leave the parameter space. Extrinsic geometries for \mathcal{U}_p based on embeddings into \mathcal{S}_p or \mathcal{S}_p^+ require projections to ensure that samples always assume values in \mathcal{U}_p . Computing such projections can be computationally expensive. An intrinsic metric geometry of \mathcal{U}_p that facilitates development of such sampling algorithms is currently unavailable, and we aim to develop an algorithm that leverages the geometry to draw posterior samples efficiently.

2.2 Geometry of the BHV space for trees

Tree structure. Every ultrametric matrix $\Sigma^T \in \mathcal{U}_p$ encodes a unique rooted tree T on p leaves, or $p+1$ leaves if the root is counted as one, with its branching structure and edge lengths. It is possible to endow a geometry on \mathcal{U}_p via one on the set of rooted trees, and one such geometry is available through that of the BHV space.

Consider the set of acyclic graphs T known as *trees* with a unique vertex of degree one referred to as the *root*. Nodes with degree one are referred to as *leaves* and all other nodes have degree greater than two and are known as *internal nodes*. The root is thus viewed as a leaf node, and a statistical rational for this is provided in Remark 3. We consider trees T on $p+1$ leaves labeled $L = \{0, 1, \dots, p\}$, where 0 labels the root. Vertices are connected by edges from the set \mathcal{E}_T , which is the union of the set \mathcal{E}_T^I of edges connecting internal vertices with the set \mathcal{E}_T^L of edges connecting internal vertices to the $p+1$ leaves. *Resolved* trees T are those with internal vertices of degree three and $2p-1$ edges in \mathcal{E}_T , while *unresolved* trees are trees T with fewer than $2p-1$ edges and containing internal vertices of degree four or higher. The vector $\mathcal{L}_T \in \mathbb{R}_{>0}^{2p-1}$ stores lengths of edges in \mathcal{E}_T . Thus every tree has an edge of positive length connected to the root and p leaf edges of positive length.

Split and compatibility. The topology of a tree T is characterized in the connectivity between its internal edges in \mathcal{E}_T^I , encoded in the set of partitions into two of $L = \{0, 1, \dots, p\}$ called *splits* pertaining to each internal edge in \mathcal{E}_T^I . Precisely, each edge $e \in \mathcal{E}_T^I$ uniquely determines a split $L = A \cup A^c$ upon its removal from a tree T , where A contains leaves on the descendant subtree of e and its complement $A^c = L - A$ contains the rest of the leaves; thus the edge emanating from the root 0 determines the split $L = \{0\} \cup \{1, \dots, p\}$. Denote by e_A the corresponding edge with length $|e_A|$. For example, $e_{L \setminus \{0\}}$ is the root edge, and $|e_{L \setminus \{0\}}|$ is the root edge length. The set $A \subset L$ identifies a split $L = A \cup A^c$, and we use split to refer to A or the edge e_A interchangeably; context will disambiguate the two.

Arbitrary collections of splits do not characterize a valid tree topology, but only a collection of *compatible* ones do: two distinct edges e_{A_1} and e_{A_2} are compatible if exactly one of $A_1 \cap A_2, A_1 \cap A_2^c, A_1^c \cap A_2$ from the associated splits is empty set (Semple et al., 2003, Theorem 3.1.4) (upon noting $A_1^c \cap A_2^c$ always

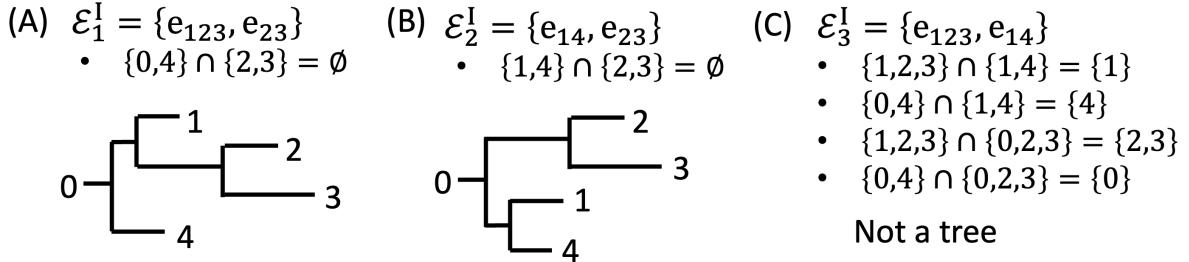


Figure 1: Three examples for demonstrating the compatibility of internal edge sets for $p = 4$ with the corresponding tree topology. Two compatible sets are shown in Panel (A) and (B) with the concordant tree topology. Panel (C) exhibits incompatible edge set that characterizes no tree topology.

contains the root); identical edges are by default compatible. Again, we interchangeably refer to compatibility of splits A_1 and A_2 to sometimes mean compatibility of the edges e_{A_1} and e_{A_2} , and this extends to a collection $\{A_1, \dots, A_k\}$ of subsets of L , e.g., we say $\{e_{A_1}, e_{A_2}\}$ is compatible with $\{e_{A_3}, e_{A_4}\}$ if any edge from the first set and any edge from the second set are compatible. Leaf edges $e_A \in \mathcal{E}_T^L$, including the one from the root, associated with singleton splits $A \subset L$ are compatible with all internal edges in \mathcal{E}_T^I , and thus do not contribute to the topology of T . A compatible edge set \mathcal{E}_T thus fully characterizes the topology of a tree T . There are $(2p - 3)!!$ distinct topologies on fully resolved trees on $p + 1$ leaves where the root is a leaf node labeled 0.

Three internal edge sets with $p = 4$ are shown in Figure 1. Panel (A) and (B) demonstrate two compatible internal edge sets. Two trees with different topology are characterized based on the corresponding compatible internal edge set. Contrarily, Panel (C) shows an incompatible edge set for none of the intersect is empty ($\{1, 2, 3\} \cup \{1, 4\}, \{0, 4\} \cup \{1, 4\}, \{1, 2, 3\} \cup \{0, 2, 3\}$ and $\{0, 4\} \cup \{0, 2, 3\}$), which implies no tree topology can be characterized by the edge set.

Geometry of BHV space. The BHV space \mathcal{T}_{p+1}^I parameterizes the space of such rooted, labeled, resolved and unresolved trees T on $p + 1$ leaves and prescribes a continuous geometry based on the lengths $|e_A|$ of internal edges $e_A \in \mathcal{E}_T^I$, where A is associated with a split of L . A fully resolved topology is parameterized by $\mathbb{R}_{>0}^{p-2}$, where each axis corresponds to one of the $p - 2$ internal splits that characterize the topology and the coordinates encode the corresponding lengths of the internal edges. The boundary of $\mathbb{R}_{>0}^{p-2}$ consists of unresolved trees with internal nodes of degree greater than 3, obtained by shrinking the internal edges to zero. Each of the $(2p - 3)!!$ topologies is identified with a copy of $\mathbb{R}_{\geq 0}^{p-2}$, known as an *orthant*, and the BHV space \mathcal{T}_{p+1}^I is defined by the $(2p - 3)!!$ orthants glued isometrically along their common boundaries comprising unresolved trees. Panel (A) of Figure 3 illustrates that two neighboring orthants share a common edge in \mathcal{T}_5^I . By accounting for lengths of $p + 1$ leaf edges, space of rooted, labeled trees on $p + 1$ leaves is the product space

$$\mathcal{T}_{p+1} = \mathcal{T}_{p+1}^I \times \mathbb{R}_{>0}^{p+1},$$

where we do not allow for zero-length leaf edges. A tree T is thus characterized by the pair $(\mathcal{E}_T, \mathcal{L}_T)$ of a compatible edge set and vector of edge lengths. The (common) origin in the BHV space \mathcal{T}_{p+1}^I consists

of trees with $p + 1$ leaf edges of positive lengths and no internal edges.

The distance $d_{\text{BHV}}(T_1, T_2)$ between two trees T_1 and T_2 on $p + 1$ leaves is defined to be the infimum of lengths of paths between T_1 and T_2 in \mathcal{T}_{p+1}^I , which are straight lines within each orthant. It is known that the space $(\mathcal{T}_p^I, d_{\text{BHV}})$ is a geodesic metric space (Billera et al., 2001). A natural distance metric on \mathcal{T}_{p+1} then is

$$d_{\text{tree}}(T_1, T_2) := d_{\text{BHV}}(T_1, T_2) + \|x - y\|_2,$$

where $x, y \in \mathbb{R}_{>0}^{p+1}$ are the vectors of leaf edge lengths in T_1 and T_2 , respectively.

3 An intrinsic stratified geometry of ultrametric matrices

Armed with the requisite constructs from the BHV space, we present a key result that equips the set \mathcal{U}_p of strictly ultrametric matrices with a convenient geometry for statistical modelling.

Theorem 1. *There is a homeomorphism $\Phi : \mathcal{U}_p \rightarrow \mathcal{T}_{p+1}$ such that when quipped with the metric*

$$d(\Sigma_1^T, \Sigma_2^T) := d_{\text{tree}}(\Phi(\Sigma_1^T), \Phi(\Sigma_2^T)), \quad (3)$$

the space (\mathcal{U}_p, d) is a CAT(0) stratified geodesic metric space.

Remark 1. An important consequence of the bijection Φ is that the preimage of *every* tree in \mathcal{T}_{p+1} , including those on the boundary with one or more internal edges of zero length, is a positive definite strictly ultrametric matrix (see Lemma 1 in the Appendix). It is possible to establish a similar bijection between the larger set \mathcal{U}_p^* , containing positive semi-definite ultrametric matrices, and a suitably modified \mathcal{T}_{p+1} , but we focus on \mathcal{U}_p in this work.

Let $\Psi := \Phi^{-1}$ be the inverse of Φ . Theorem 1 enables us to pullback the CAT(0) geometry from the tree space \mathcal{T}_{p+1} onto the set \mathcal{U}_p of ultrametric matrices using Ψ . CAT(0) spaces are spaces with non-positive curvature in the Alexandrov sense, and corresponding geometry equips \mathcal{U}_p with following properties that are particularly important in the sequel:

- (i) There is a unique geodesic between any two ultrametric matrices that lies within \mathcal{U}_p , which enables us to develop a sampling algorithm that makes intrinsic local moves along straight lines in \mathcal{U}_p ;
- (ii) Fréchet means of probability measures exist and are unique; this enables computation of an estimate of the posterior mean using samples from the posterior distribution on \mathcal{U}_p .

For details on CAT(0) spaces, we refer to Bridson and Haefliger (1999). Additionally, the CAT(0) structure of \mathcal{U}_p opens up the possibility of using the plethora of statistical tools for the BHV space \mathcal{T}_{p+1}^I (e.g., means and variances (Brown and Owen, 2020), principal component analysis (Nye, 2011)) and transfer results on to \mathcal{U}_p with Ψ ; most relevant to our work is computation of the geodesic distance d on \mathcal{U}_p using the polynomial-time algorithm for computing d_{BHV} on \mathcal{T}_{p+1}^I (Owen and Provan, 2011) and fast computation of the Fréchet mean (Miller et al., 2015).

The proof of Theorem 1 uses the decomposition of Σ^T in (2) to provide a constructive definition of Φ as

$$\Phi(\Sigma^T) = (\Phi_V(V), \Phi_D(D)), \quad \Sigma^T \in \mathcal{U}_p, \quad (4)$$

based on $V \mapsto \Phi_V(V) = \mathcal{E}_T$ that maps a V with the partition property to a compatible edge set \mathcal{E}_T and $D \mapsto \Phi_D(D) = \mathcal{L}_T$ that maps D to a vector $\mathcal{L}_T \in \mathbb{R}_{>0}^{2p-1}$. An algorithmic definition of Φ can be obtained upon observing a key property of strictly ultrametric matrices in \mathcal{U}_p that can be gleaned from an alternative decomposition to (2) obtained by the inverse Ψ of Φ in the following result.

Corollary 1. *For every edge e_A in a collection \mathcal{E}_T of compatible edges/splits corresponding to a tree T , there is a unique $p \times p$ symmetric binary matrix E_A containing ones at all pairs of indices obtained from A (the set of leaves on the descendent subtree of the edge e_A) and zeros elsewhere, such that the ultrametric matrix Σ^T can be expressed in coordinates $(\mathcal{E}_T, \mathcal{L}_T)$ via the decomposition*

$$\Sigma^T = \Psi((\mathcal{E}_T, \mathcal{L}_T)) := \sum_{e_A \in \mathcal{E}_T} |e_A| E_A. \quad (5)$$

The topology of T is characterized through the pattern of zeroes in the matrix

$$\tilde{\Sigma}^T := \Sigma^T - |e_{L \setminus \{0\}}| \mathbf{1} \mathbf{1}^\top,$$

where $e_{L \setminus \{0\}}$ represents the root edge.

From the above Corollary, we observe that Ψ can be expressed component-wise such that $\Psi_V(\mathcal{E}_T) = \{E_A\}$ and $\Psi_D(\mathcal{L}_T) = (|e_A|, \dots, |e_{A_{2p-1}}|)^\top$, from which the corresponding basis matrix V with partition property and D are uniquely obtained. Proof of the decomposition (5) follows from the proof of Theorem 1. The claim concerning topology of T is proved upon noting that an element Σ_{ij}^T of Σ^T is additive with respect to the sequence of edge lengths from the root to the most recent common ancestor of leaves i and j , and this ensures that $|e_{L \setminus \{0\}}|$ is the smallest possible value in Σ^T . In fact, the topology of not just T but also any of its subtrees $T_A, A \subset L$ is encoded in the pattern of zeros of $\Sigma^{T_A} - |e_{A \setminus \{0\}}| \mathbf{1}_A \mathbf{1}_A^\top$, where Σ^{T_A} is the ultrametric matrix corresponding to the tree T_A and $\mathbf{1}_A$ is the vector of ones of dimension $|A|$. This is evident from the role of the rank-one binary matrices in the decomposition (5).

Returning to how Φ works on Σ^T , consider the following procedure. Given Σ^T , compute $\tilde{\Sigma}^T$, whose zero entries divide it into multiple submatrices representing different subtrees. Recursively, repeat this process on each submatrix until all submatrices reduce to either 2×2 diagonal matrix or a scalar, corresponding, respectively, to two leaves or a single leaf of T . At each stage of the recursion, add to the resulting submatrix a matrix of ones of appropriate dimension scaled by the minimum value that was subtracted. This value is the length of the edge emanating from the root of the corresponding subtree. The map Φ reduces to Φ_V when *every* coefficient in the decomposition (5) is set to one and T reduces to a combinatorial tree with unit edge lengths. In this case the smallest element in Σ^T is necessarily one.

Figure 2 illustrates how Φ works on a five-dimensional Σ^T . Given Σ^T in Panel (A), we compute $\tilde{\Sigma}^T$ with the minimum element c_0 , which corresponds to the single edge from the root with the edge length c_0 . The resulting $\tilde{\Sigma}^T$ is divided by zero elements into two submatrices in blue and yellow boxes. These submatrices represent corresponding descendant subtrees shown in blue and yellow triangles in Panel (A). For each submatrix, we repeat the same process until all submatrices reduce to either 2×2 diagonal matrix or a scalar. Consequently, Panel (B) and (C) represent the result of using the same process for two submatrices from (A). Specifically, blue triangle in (A) becomes a subtree with the single edge with length a , which is the minimum element in the submatrix highlighted by blue box in (A). Similarly, Panel

(C) resolves the yellow triangle into a subtree with the single edge of length b , which is the minimum element of the submatrix in yellow box in (A). The same process is repeated and results in the whole tree in Panel (D).

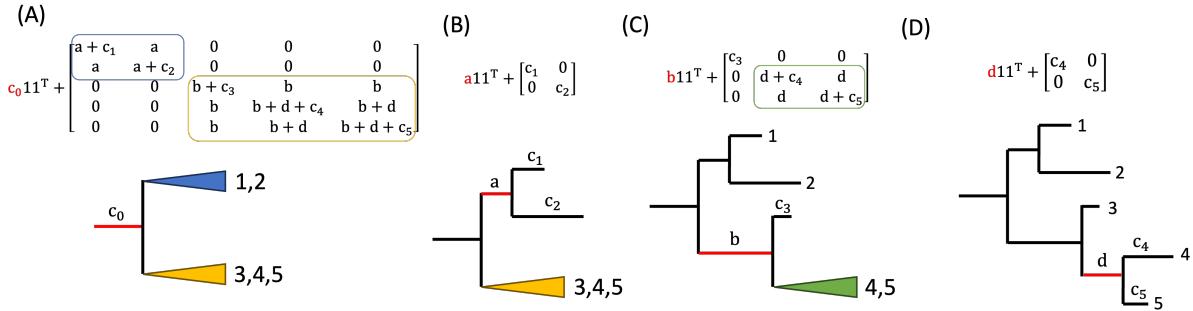


Figure 2: Panels (A)-(D) demonstrate how Φ maps $\Sigma^T \in \mathcal{U}_5$ to $T \in \mathcal{T}_6$: draw an edge of length $|e_{L \setminus \{0\}}|$ given by the minimum element of Σ^T and compute $\tilde{\Sigma}^T = \Sigma^T - |e_{L \setminus \{0\}}| \mathbf{1} \mathbf{1}^T$. Pattern of zeroes of this matrix determines subtrees. Repeat this process until all submatrices are of dimension at most two, which then determines the overall branching structure. Colored triangles in the trees represent subtrees corresponding to subsets of leaves.

Remark 2. The representation (5) provides a complementary formulation to that in equation (7) of Sturmfels et al. (2021), and matrices $\{E_A\}$ of binary matrices are linearly independent basis vectors that characterize the uncovered simplicial cone structure of the subset of ultrametric matrices for trees with a *fixed topology* within the set of $p \times p$ symmetric matrices. Each such collection \mathcal{E}_T of compatible edges/splits that identifies one of $(2p - 3)!!$ possible tree topologies determines a unique basis $\{E_A\}$ of rank-one binary matrices that are the extremal rays of the simplicial cone within the set of $p \times p$ symmetric matrices.

Remark 3. From Corollary 1, we note that the presence of a matrix of ones corresponding to the edge $e_{L \setminus \{0\}}$ attached to the root 0 ensures that Σ^T does not contain any zeroes. Since zero covariance between Gaussian random variables implies independence, the absence of zeroes is a desideratum when using a parametric Gaussian model $\{N_p(\mathbf{0}, \Sigma^T) : \Sigma^T \in \mathcal{U}_p\}$ for modelling hierarchically correlated components of a random vector. This further explains consideration of the root labeled by 0 as a leaf node. For covariance graph models (e.g., Dawid and Lauritzen, 1993; Letac and Massam, 2007), where the graph is a tree and zeroes of the covariance matrix determine the corresponding graph, it is more convenient to consider trees with p leaves and $(2p - 2)$ edges, where there is no single edge connecting the root; this results in the ultrametric matrix $\tilde{\Sigma}^T$ whose pattern of zeroes encodes the graph.

For every Σ^T , the image $\Phi(\Sigma^T)$ is a tree with $p + 1$ leaves (including the edge attached to the root 0), each with non-zero edge lengths. The tree consists of $(p - 2)$ internal edges, some of which may be of length zero. The preimage $\Psi(T)$ of every tree $T \in \mathcal{T}_{p+1}$ is a positive definite strictly ultrametric matrix in \mathcal{U}_p . The preimage of each of the $(2p - 3)!!$ orthants of dimension $(2p - 1)$ in \mathcal{T}_{p+1} is in \mathcal{U}_p and forms a simplicial cone within \mathcal{S}_p . The simplicial cone is spanned by rank-one basis matrices $\{E_A\}$ corresponding to a collection \mathcal{E}_T of compatible splits. Each of these simplicial cones is referred to as a *stratum* of \mathcal{U}_p . The subset of $(p - 2)$ internal splits identifies a stratum within \mathcal{U}_p and maps via Φ to the orthant in

the BHV space \mathcal{T}_{p+1}^I pertaining to the topology of T . This is achieved by noting that the set of $(p-2)$ internal splits is a subset of \mathcal{E}_T , determining $(p-2)$ binary matrices within the collection of $p \times p$ binary matrices $\{E_A : e_A \in \mathcal{E}_T\}$ corresponding to the internal edges. Of the remaining $p+1$ binary matrices within $\{E_A\}$, there is a single matrix of ones corresponding to the edge emanating from the root and the rest contains a single non-zero entry on the diagonal. Together, they represent the $(p+1)$ -axes in $\mathbb{R}_{>0}^{p+1}$ that identify splits associated with the leaf edges in \mathcal{T}_{p+1} .

Consequently, the ultrametric matrix $\Psi(T) = \sum_{e_A \in \mathcal{E}_T} |e_A| E_A$ is located within a stratum identified by $\{E_A : e_A \in \mathcal{E}_T\}$ coordinatized by the point $(|e_A|, \dots, |e_{2p-1}|)^T$ in \mathbb{R}^{2p-1} . Figure 3 illustrates the decomposition of an ultrametric matrix Σ^T in \mathcal{U}_4 with corresponding tree $T \in \mathcal{T}_5$ and a compatible edge set $\mathcal{E}_T = \{e_{123}, e_{23}, e_1, e_2, e_3, e_4\}$. An ultrametric matrix $\Psi(T)$ corresponding to a fully resolved tree T

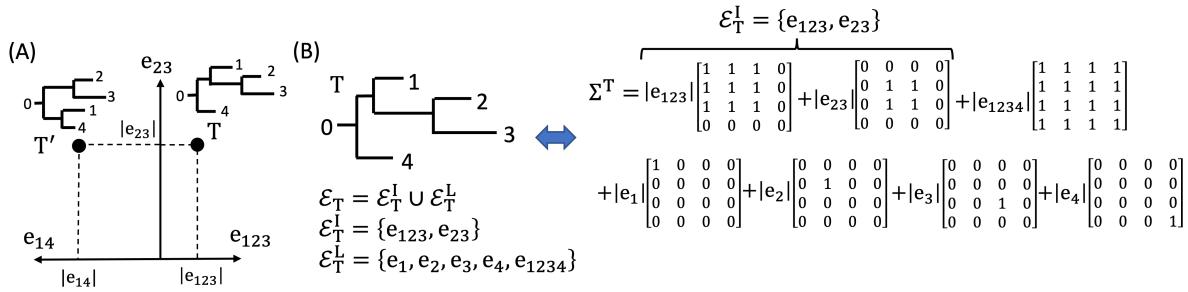


Figure 3: Bijection between the tree space \mathcal{T}_5 and the space \mathcal{U}_4 of ultrametric spaces in Theorem 1 and the corresponding decomposition in Corollary 1: Panel (A) shows the BHV tree space \mathcal{T}_5^I with axes identified with two compatible internal splits; Panel (B) shows how a rooted labeled tree in \mathcal{T}_5 uniquely identifies an ultrametric matrix Σ^T in \mathcal{U}_4 .

lies in the interior of one of the $(2p-3)!!$ strata of dimension $(2p-1)$ coordinatized by $\mathbb{R}_{>0}^{p-2} \times \mathbb{R}_{>0}^{p+1}$. If T is unresolved with a single node of degree four, $\Psi(T)$ lies within the $(2p-2)$ -dimensional boundary of a stratum reached by shrinking a coefficient $|e_A|$ corresponding to an internal edge $e_A \in \mathcal{E}_T^I$ to zero in the decomposition (5). Ultrametric matrices on the boundary of codimension $s \geq 1$, corresponding to unresolved trees with $p-2-s$ internal nodes, are obtained by shrinking s coefficients corresponding to the internal edges in (5) to zero.

There are clearly many paths from the interior of a stratum to a codimension $s \geq 1$ boundary of \mathcal{U}_p ; in other words, there are many copies of \mathcal{U}_q within \mathcal{U}_p for $q < p$. Moreover, an ultrametric matrix in \mathcal{U}_q on the boundary is still positive definite and full rank. It is then natural to query: How do ultrametric matrices on the boundaries ‘look’? Are they sparser in some appropriate sense to those in the interior?

To answer this in a relative sense, note that boundaries in the BHV space are identified (glued) isometrically along their common faces. Pulling this structure back onto \mathcal{U}_p engenders a similar identification of the boundaries of \mathcal{U}_p . With $\text{vec} : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^{p^2}$ as the vectorization map, consider the half-vectorisation map $\text{vech}(x) := \text{Avec}(x)$, where the matrix

$$A := \sum_{i \geq j} (u_{ij} \otimes e_j^\top \otimes e_i^\top) \in \mathbb{R}^{p(p+1)/2 \times p^2}$$

picks out the lower triangular part of the vectorization, $\{e_j\}$ is the standard basis in \mathbb{R}^{p^2} , and u_{ij} is a $p(p+1)/2$ -dimensional unit vector with 1 in position $(j-1)p+i-j(j-1)/2$ and 0 elsewhere. Consider

a partial order \preceq on \mathcal{U}_p defined as $A \preceq B$ if $\text{vech}(A) \leq \text{vech}(B)$, and $\mathbf{x} \leq \mathbf{y}$ on $\mathbb{R}^{p(p+1)/2}$ if every element of \mathbf{x} is utmost its corresponding element in \mathbf{y} .

Proposition 1. *For $1 < k \leq (p - 2)$, an ultrametric matrix Σ^T on a $(2p - 1 - k)$ -dimensional boundary can be reached along geodesics from distinct ultrametric matrices lying in the interior of $(2k + 1)!!$ strata such that Σ^T is smaller with respect to \preceq to each of the $(2k + 1)!!$ ultrametric matrices. Here $(2k + 1)!!$ is the number of binary trees with $k + 2$ leaves.*

From Proposition 1, we note that each codimension one boundary of \mathcal{U}_p is thus obtained by gluing, or identifying, three distinct codimension one boundaries of $\mathbb{R}_{\geq 0}^{p-2} \times \mathbb{R}_{>0}^{p+1}$ pertaining to the first factor, which coordinatize three different tree topologies. This structure plays an important role in development of sampling algorithms for inference in the sequel. For $I = \{1, \dots, (2p - 3)!!\}$, we introduce the notation

$$\mathcal{U}_p = \bigcup_{i \in I} \mathcal{U}_p^i$$

to represent \mathcal{U}_p as a disjoint union of $(2p - 3)!!$ spaces $\mathcal{U}_p^i = \mathbb{R}_{\geq 0}^{p-2} \times \mathbb{R}_{>0}^{p+1}$, $i \in I$ with common boundaries glued as described above. The origin common to each stratum \mathcal{U}_p^i consists of matrices of the form $\alpha(\mathbf{1}\mathbf{1}^\top - \mathbb{I}_p) + D$, where $\alpha > 0$ and D is a diagonal matrix with entries $d_{ii} > \alpha$, which as discussed in Remark 1 is positive definite and strictly ultrametric.

Recall that, by virtue of the definition, no matrix $\Sigma^T \in \mathcal{U}_p$ has a zero, but every matrix $\tilde{\Sigma}^T = \Sigma^T - |e_{L \setminus \{0\}}| \mathbf{1}\mathbf{1}^\top$ does, whose pattern determines topology of the corresponding tree T . The following result states that matrices $\tilde{\Sigma}^T$ get sparser, in terms of number of zeroes, with increasing codimension of boundaries of \mathcal{U}_p . Proof follows from that of Proposition 1.

Corollary 2. *An ultrametric matrix on a stratum of dimension j is smaller with respect to \preceq than any other on a stratum of dimension k , where $j < k$.*

From the decomposition of Σ^T in Corollary 1, we thus note that sparsity of $\tilde{\Sigma}^T$ is controlled by the number of non-zero coefficients in \mathcal{L}_T , and Corollary 2 implies that the (co)dimension of the boundary component $\tilde{\Sigma}^T$ lies in quantifies the level of sparsity of $\tilde{\Sigma}^T$.

Remark 4. Coefficients in the sum (5) are lengths of edges in the set \mathcal{E}_T of edges given by compatible splits of L . There are a total of $N = 2^p - 2$ possible nonempty splits of L of which $N - p$ correspond to the internal edges. Thus, using Φ , the set \mathcal{U}_p can be given *extrinsic* coordinates via an embedding of \mathcal{T}_{p+1} into \mathbb{R}^{N-p} where each set $\{|e_A|, e_A \in \mathcal{E}_T\}$ of edge lengths can be identified with a vector in \mathbb{R}^{N-p} , obtained upon choosing an ordering of the $N - p$ splits; the j th split is associated with the j th standard basis vector consisting of a one in the j th position and zero everywhere else. Such an embedding results in an alternative representation of Σ^T as very sparse vector in \mathbb{R}^{N-p} , since only a small subset of vectors correspond to compatible splits. The geometric picture of \mathcal{U}_p hence obtained with such an embedding is not transparent.

4 Prior distributions on ultrametric matrices

From Theorem 1 and Corollary 1, we note that an ultrametric matrix $\Sigma^T \in \mathcal{U}_p$ can be given coordinates $(\mathcal{E}_T, \mathcal{L}_T)$ corresponding to the tree T via Ψ . We can thus define a prior distribution on \mathcal{U}_p pushing forward

one on \mathcal{T}_{p+1} under Ψ . Consider the product measure $d\eta := dN dx$, where dN is the counting measure on $N := \{0\} \cup \{1, 2, \dots\}$ and dx is the Lebesgue measure on $\mathbb{R}_{>0}^{2p-1}$. Consider the probability distribution $\mu = \pi d\eta$ on \mathcal{T}_{p+1} with density $\pi := \pi_{\mathcal{E}} \pi_{\mathcal{L}|\mathcal{E}}$ that factors into a density $\pi_{\mathcal{E}}$ on sets of compatible splits and a density $\pi_{\mathcal{L}|\mathcal{E}}$ on edge lengths conditional on the splits; π can always be chosen to be a valid probability density since $\int_{\mathcal{T}_{p+1}} \pi d\eta < \infty$ by the CAT(0) property of \mathcal{T}_{p+1} . Define the distribution

$$\nu(A) = \mu \circ \Phi(A), \quad A \subset \mathcal{U}_p,$$

on \mathcal{U}_p . The distribution ν need not be absolutely continuous with respect to μ and may hence not have a density. Under this setup, we specify a prior distribution ν by first doing so on \mathcal{T}_{p+1} via the density π .

The density $\pi_{\mathcal{E}}$ can be specified to have mass on resolved trees in the interior and unresolved trees on the boundaries of \mathcal{T}_{p+1} . For the former, we can consider a binary fragmentation model based on a rule for recursively splitting \mathcal{E}_T into two blocks such that

$$\pi_{\mathcal{E}}(\mathcal{E}_T) = \prod_{e_A, e_B \in \mathcal{E}_T} \pi_{\text{split}}(e_A, e_B \mid e_{A \cup B}), \quad (6)$$

where $\pi_{\text{split}}(e_A, e_B \mid e_{A \cup B})$ is a symmetric function that captures the chance of a block $A \cup B$ splitting into two blocks of A and B . The model is rendered Markovian if π_{split} depends only on the sizes of blocks A and B , and is consistent when the distribution $\pi_{\mathcal{E}}$ when restricted to subtrees of T is the same as that of a tree that equals the subtree. Berestycki and Pitman (2007) considered a time-irreversible Markovian fragmentation process in order to define the splitting function π_{split} , while Pitman (2006) defined a Gibbs-type alternative which defines π_{split} as a product of weights that depend only on the size of sub-blocks. For such a Gibbs-type alternative, Theorem 2 of McCullagh et al. (2008) showed that the beta-splitting model of Aldous (1996) is the only one that results in a consistent Markovian binary fragmentation-based prior $\pi_{\mathcal{E}}$ on topology. The beta-splitting model prescribes the specification

$$\pi_{\text{split}}(e_A, e_B \mid e_{A \cup B}) \propto \frac{\Gamma(n_A + \beta + 1)\Gamma(n_B + \beta + 1)}{\Gamma(n_A + n_B + 2\beta + 2)}, \quad (7)$$

where n_A is the cardinality of the set A and $\beta \in (-2, \infty]$ is the hyper-parameter. The consistency here refers to when $\pi_{\mathcal{E}}$ remains the same when restricted to subtrees $T_A, A \subset L$. The Markov property of $\pi_{\mathcal{E}}$ implies that if for a fixed partition $\{A_1, \dots, A_k\}$ of L , the k subtrees restricted to A_1, \dots, A_k are independently distributed as $\pi_{\mathcal{E}_{T_{A_1}}}, \dots, \pi_{\mathcal{E}_{T_{A_k}}}$. For example, $\beta = -1.5$ leads to the uniform density $\frac{1}{(2p-3)!!}$ on topology, while $\beta = 0$ corresponds to the Yule model (Yule, 1925). For unresolved non-binary trees on the boundary of \mathcal{U}_p , a Poisson-Dirichlet model based on multifurcating Gibbs fragmentations can be used to define a consistent Markovian prior $\pi_{\mathcal{E}}$ (McCullagh et al., 2008, Theorem 8).

For simplicity, we decouple the density specifications for \mathcal{E}_T and \mathcal{L}_T and thus consider options for the choice of density $\pi_{\mathcal{L}}$. If the density is to have support only on fully resolved trees, several options are available for the prior, including products of $(2p-1)$ one-dimensional densities on $\mathbb{R}_{>0}$. For a specification with mass on the boundary of \mathcal{T}_{p+1} corresponding to unresolved trees with zero edge lengths, the Lebesgue component dx of the dominating measure $d\mu$ can be chosen to be a mixture of appropriate lower-dimensional Lebesgue measures with disjoint supports such that $\pi_{\mathcal{L}}$ will be mixture density consisting of individual components with disjoint supports.

Summarily, the decomposition in Corollary 1, made possible by the geometry inherited from \mathcal{T}_{p+1} , enables definition of prior distributions separately or jointly on the graph structure and the corresponding edge lengths encoded by Σ^T , as interpreted through the corresponding tree T . Moreover, the homeomorphic property of Φ ensures that we are able to define a prior on Σ^T that inherits the attractive Markov and consistency properties of the tractable binary fragmentation prior on the combinatorial tree T without edge lengths.

5 Posterior inference under a Gaussian latent tree model

We focus on the Gaussian latent tree model $\{N_p(\mathbf{0}, \Sigma^T) : \Sigma^T \in \mathcal{U}_p\}$ for multivariate data where correlations between components of a random vector is modeled through a tree T . The model is a special case of covariance graph models (Cox and Wermuth, 1993) in that a tree is a type of graph, and has received considerable attention (e.g., Choi et al., 2011; Shiers et al., 2016; Zwiernik, 2016; Leung and Drton, 2018), however, mostly restricted to estimation strategies. Our formulation of the Gaussian latent tree model differs from existing works in that the ultrametric covariance Σ^T we consider does not contain any zeros owing to the presence of an edge from the root.

Using the prior distribution from Section 4, we consider the following Bayesian model:

$$\begin{aligned} \mathbf{X}_1, \dots, \mathbf{X}_n | \Sigma^T &\stackrel{\text{i.i.d.}}{\sim} N_p(\mathbf{0}, \Sigma^T); \\ \Sigma^T &\sim \nu, \end{aligned}$$

where $\nu = \mu \circ \Phi = \pi d\eta \circ \Phi$. The density π is chosen with $\pi_{\mathcal{E}}$ corresponding to the density from a binary Gibbs fragmentation model with $\beta = -1.5$, and $\pi_{\mathcal{L}|\mathcal{E}} = \pi_{\mathcal{L}}$ such that the priors on topology and edge lengths are independent; $\pi_{\mathcal{L}}$ is taken to be a product of $(2p - 1)$ one-dimensional exponential densities $\text{Exp}(a)$ with a common mean a . Our choice implies that $\pi(\partial\mathcal{T}_{p+1}) = 0$ so that there is no mass on the boundary of \mathcal{T}_{p+1} . Extensions to non-Gaussian likelihoods (e.g., multivariate t -distribution and elliptical distributions) and presence of additional parameters (e.g., mean), can be carried out along similar lines once the advantages in the geometry-driven approach to sampling for the Gaussian case is well-understood.

The homeomorphism Φ in Theorem 1 enables us to consider the equivalent reparameterised model

$$\begin{aligned} \mathbf{X}_1, \dots, \mathbf{X}_n | \Psi(T) &\stackrel{\text{i.i.d.}}{\sim} N_p(\mathbf{0}, \Psi(T)); \\ T &\sim \mu, \end{aligned} \tag{8}$$

which is the model we use in what follows. The relatively simple choice for μ on \mathcal{T}_{p+1} is chosen to clearly demonstrate benefits of the stratified geometric structure of \mathcal{U}_p for developing an efficient sampler for the posterior distribution $\nu(\cdot | \mathbf{X}_1, \dots, \mathbf{X}_n)$ on \mathcal{U}_p corresponding to the consistent and Markovian prior ν . Sampling schemes that allow for moves between boundary components of varying codimensions within \mathcal{U}_p can get quite complicated, and will obfuscate the main benefits of incorporating geometric information into the scheme. However, the geometric framework does allow for exploration of the boundary of \mathcal{U}_p for certain applications, and we will take this up in future work.

5.1 Metropolis-Hastings algorithm

Denote by $\Sigma^{T(l)} = \Psi(T_{(l)}) \in \mathcal{U}_p$ the l th iterate of an algorithm to sample from the posterior distribution $\nu(\Sigma^T | \mathbf{X}_1, \dots, \mathbf{X}_n)$. More generally, an (l) in either the subscript or superscript of any quantity is used to denote its value at the l th iteration. An algebraic approach to explore \mathcal{U}_p involves exploiting the semigroup structure of \mathcal{U}_p under the operation of taking element-wise minimum (McCullagh, 2006): the matrix $\Sigma^{T_1} \wedge \Sigma^{T_2}$ obtained by taking element-wise minimum of corresponding elements of two ultrametric matrices Σ^{T_1} and Σ^{T_2} is again ultrametric, and this can be used to propose a new ultrametric matrix $\Sigma^{T(l+1)}$ by computing $\Sigma^{T(l)} \wedge \Sigma^T$, where Σ^T is a sample from the prior ν . However, it may happen that the resulting matrix lies on the boundary of \mathcal{U}_p , where the chosen prior π has no mass.

An alternative is to exploit the fact that \mathcal{U}_p is closed under the action $P\Sigma^TP^\top$ of any $p \times p$ permutation matrix P (McCullagh, 2006), and one can propose $\Sigma^{T(l+1)} = P\Sigma^{T(l)}P^\top$ for a random permutation P . While such updates ensure that one stays within \mathcal{U}_p , there is no control over the topological component, and moves may result in large jumps across strata in \mathcal{U}_p .

Specification of the density π on \mathcal{T}_{p+1} was facilitated by the decomposition (5) of an ultrametric metric into its constituent topological component $\{E_A\}$ of binary matrices which encode information of the set of compatible splits \mathcal{E}_T , and geometric component \mathcal{L}_T consisting of edge lengths. From Theorem 1, since \mathcal{U}_p is a geodesic metric space with respect to the pullback geometry under Φ , it is possible to design a sampling scheme that makes moves along geodesics in \mathcal{U}_p . However, since \mathcal{U}_p contains $(2p - 3)!!$ strata glued along their common boundaries, speed of mixing of the algorithm will highly depend on the probability of $\Sigma^{T(l+1)} = \Psi(T_{(l+1)})$ possessing a topology different from that of $\Sigma^{T(l)}$.

Our approach is based on an algorithm that makes a *local* move from $\Sigma^{T(l)}$ along a geodesic in \mathcal{U}_p that is informed by the map Φ . Such a move might involve change in topology and/or edge lengths, where ‘locality’ is determined the distance from $\Sigma^{T(l)}$ along a geodesic. More precisely, suppose that the current iterate $\Sigma^{T(l)}$ lies in stratum \mathcal{U}_p^i for some i :

- (i) Compute $\Phi(\Sigma^{T(l)}) = (\mathcal{E}_{T(l)}, \mathcal{L}_{T(l)})$, and determine the orthant of \mathcal{T}_{p+1} the tree $T_{(l)}$ lies in;
- (ii) Randomly choose an internal edge/split $e_A \in \mathcal{E}_{T(l)}^I$ and set $|e_A| = 0$ so that $T_{(l)}$ now lies on a codimension one boundary of \mathcal{T}_{p+1} ; identify, via the corresponding changes to $\mathcal{E}_{T(l)}$, the two nearest neighbours, distinct orthants j and k of \mathcal{T}_{p+1} with $j \neq i$ and $k \neq i$, of $T_{(l)}$ (there exist two and only two neighboring orthants distinct from the one where $T_{(l)}$ lies; see Proposition 1 with $k = 1$);
- (iii) Randomly choose between orthants j or k of \mathcal{T}_{p+1} to decide on the updated topology $\mathcal{E}_{T(l+1)}$ of $T_{(l+1)}$;
- (iv) Update $\mathcal{L}_{T(l)}$ to $\mathcal{L}_{T(l+1)}$ to obtain $T_{(l+1)} = (\mathcal{E}_{T(l+1)}, \mathcal{L}_{T(l+1)})$;
- (v) Compute $\Psi(T_{(l+1)}) = \Sigma^{T(l+1)}$.

Summarily, the update $\Sigma^{T(l+1)}$ is obtained from $\Sigma^{T(l)}$ by implementing a *nearest neighbour interchange* along a geodesic that starts in stratum \mathcal{U}_p^i and ends in stratum \mathcal{U}_p^j with $i \neq j$ by passing through the codimension one boundary shared by the two strata. If the topology update in steps (ii) and (iii) is rejected, then a local move constitutes moving along a geodesic within the same orthant by merely updating

edge lengths. This is the benefit of endowing \mathcal{U}_p with a continuous BHV geometry of combinatorial trees with edge lengths. The motivation behind making a mandatory change in topology (which indeed can be rejected) is to make many computationally cheap local moves over a large number of iterations rather than a few computationally expensive moves, resulting in better exploration of the space of rooted trees.

To illustrate the local moves, we detail the update step in the space \mathcal{T}_5^I of four-leaf trees in Figure 4. Given a tree $T_{(l)}$ with four leaves at the l th iteration, we propose a candidate via a geodesic move from $T_{(l)}$ to $T_{\text{cand},(l+1)}^1$ with every matrix on the path being ultrametric; this is highlighted by a path (red dashed line) in panel (F). Five matrices in \mathcal{U}_4 on the geodesic path from $T_{(l)}$ to $T_{\text{cand},(l+1)}^1$ are shown in panels (A)-(E). Our algorithm is an adaptation of the random walk algorithm on the BHV space \mathcal{T}_{p+1}^I by

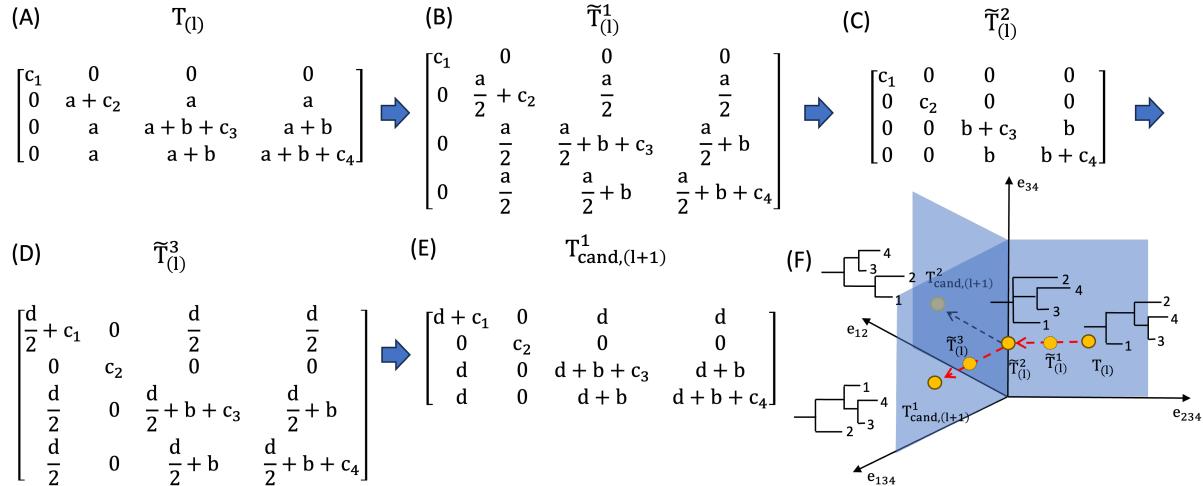


Figure 4: Updating topology in \mathcal{T}_5^I using Algorithm 1: Given tree $T_{(l)}$ at the l th iteration in orthant i , the proposal function randomly shrinks an internal edge and moves to an intermediate unresolved tree $\tilde{T}_{(l)}^2$ on a codimension one boundary common to two other distinct orthants j and k with two candidate trees $T_{\text{cand},(l+1)}^1$ and $T_{\text{cand},(l+1)}^2$, one of which is chosen with equal probability, and reached by following geodesics (red and black dashed lines) from $T_{(l)}$. Panels (A)-(E) show three ultrametric matrices along the geodesic connecting $\Psi(T_{(l)})$ to candidate $\Psi(T_{\text{cand},(l+1)}^1)$ (red dashed line).

Nye (2020), where there is an equal chance to change or remain within an orthant. The objective in their work was to construct a Brownian motion on \mathcal{T}_{p+1}^I as the limit of a geodesic random walk, and as such may lead to a slower mixing of a Markov chain to sample from the posterior owing to the rather large probability of not changing topology within an update; see Section S1.2 of the Supplementary Material for evidence of this.

The ability to make local geometry-driven moves on \mathcal{U}_p leads to a geometric Metropolis-Hastings algorithm that explores \mathcal{U}_p via the image of an exploration of \mathcal{T}_{p+1} under Ψ , where *two* acceptance probabilities, both under the reparameterised Bayesian model (8), corresponding to the decoupled topology and edge length updates are computed to first update the topology and then the edge lengths. The acceptance probability is

$$\max \left\{ 1, \frac{\pi(T') N_p(\mathbf{0}, \Psi(T')) q(T | T')}{\pi(T) N_p(\mathbf{0}, \Psi(T)) q(T' | T)} \right\}, \quad (9)$$

where $q(\cdot | \cdot)$ is a transition kernel on \mathcal{T}_{p+1} , applies to updates of both \mathcal{E}_T and \mathcal{L}_T with suitable changes to π and q . When updating \mathcal{E}_T for a fixed \mathcal{L}_T , the transition kernel q is symmetric since because of the uniform probabilities assigned to first choosing an internal edge to shrink to zero, and second in choosing a new orthant to move to from the codimension one boundary. This implies that (9) reduces to

$$\alpha_{\mathcal{E}} := \max \left\{ 1, \frac{\pi_{\mathcal{E}}(\mathcal{E}'_T) N_p(0, \Psi(T'))}{\pi_{\mathcal{E}}(\mathcal{E}_T) N_p(0, \Psi(T))} \right\}. \quad (10)$$

where $\pi_{\mathcal{E}}(\mathcal{E}_T)$ is the density is defined in (6). For updating an edge length $|e_A| \in \mathcal{L}_T$, for a fixed \mathcal{E}_T , evidently,

$$\alpha_{\mathcal{L}} := \max \left\{ 1, \frac{\text{Exp}(|e'_A|; a) N_p(\mathbf{0}, \Psi(T')) \text{TN}_{(0, \infty)}(|e_A|; |e'_A|, \sigma_{\mathcal{L}})}{\text{Exp}(|e_A|; a) N_p(\mathbf{0}, \Psi(T)) \text{TN}_{(0, \infty)}(|e'_A|; |e_A|, \sigma_{\mathcal{L}})} \right\}, \quad (11)$$

where $\text{Exp}(\cdot; a)$ is the exponential density with mean parameter a , and the transition kernel q is taken to be $\text{TN}_{(0, \infty)}(\cdot; \mu, \sigma_{\mathcal{L}})$, the truncated normal density on $(0, \infty)$ with mean μ and standard deviation $\sigma_{\mathcal{L}}$. The Metropolis-Hastings algorithm is summarized in Algorithm 1.

5.2 Posterior summaries

Algorithm 1 outputs posterior samples of $\Sigma^{T(l)}$, $l = 1, \dots, M$, from the posterior distribution $\nu(\Sigma^T | \mathbf{X}_1, \dots, \mathbf{X}_n)$. We consider estimates of two functionals of $\nu(\Sigma^T | \mathbf{X}_1, \dots, \mathbf{X}_n)$: (i) the maximum a posteriori (MAP) ultrametric matrix; (ii) Fréchet mean of $\nu(\Sigma^T | \mathbf{X}_1, \dots, \mathbf{X}_n)$ defined as a minimizer of

$$\Psi(\tilde{T}) \mapsto \int d^2(\Psi(T), \Psi(\tilde{T})) d\mu(\Psi(T) | \mathbf{X}_1, \dots, \mathbf{X}_n) \quad (12)$$

under the reparameterised model (8). Existence of the Fréchet mean minimising the above functional depends on the support of the posterior distribution; when it exists, it will be unique owing to the inherited CAT(0) geometry of \mathcal{U}_p . On \mathcal{T}_{p+1} , the sample Fréchet mean \tilde{T} , obtained by minimising

$$\mathcal{T}_{p+1} \ni \tilde{T} \mapsto \sum_{l=1}^M d_{\text{tree}}^2(\tilde{T}, T_{(l)})$$

with respect to the empirical measure on trees $\{T_{(l)}\}$, exists and is unique owing to the CAT(0) geometry of \mathcal{T}_{p+1} (Bridson and Haefliger, 1999). From Theorem 1, we thus note that $\Psi(\tilde{T})$ estimates the Fréchet mean of the posterior $\nu(\Sigma^T | \mathbf{X}_1, \dots, \mathbf{X}_n)$, when its exists. On \mathcal{T}_{p+1} , \tilde{T} is a consistent estimator of the Fréchet mean of $\mu(T | \mathbf{X}_1, \dots, \mathbf{X}_n)$ (Barden et al., 2018), and by continuity of Ψ , $\Psi(\tilde{T})$ consistently estimates the Fréchet mean of $\nu(\Sigma^T | \mathbf{X}_1, \dots, \mathbf{X}_n)$.

The estimate $\Psi(\tilde{T})$ is computed using the algorithm by Miller et al. (2015), and accuracy of approximations is ascertained chiefly with respect to the intrinsic distance d using the polynomial-time algorithm of Owen and Provan (2011) to compute d_{tree} . Using these tools, we summarise the posterior by quantifying the proportion of subtrees present in $\Sigma^{T(l)}$, $l = 1, \dots, M$, and construct 95% credible intervals for each element of the true data generating ultrametric matrix.

Algorithm 1. Geometry-driven Metropolis-Hastings algorithm for posterior sampling

```

Input :
(a) Initialization: Current iterate  $\Sigma^{T(l)}$ ;
(b) Probability densities  $\pi_{\mathcal{E}}$  and  $\pi_{\mathcal{L}}$ ;
(c) Number of iterations  $M$  and standard deviation  $\sigma_{\mathcal{L}}$ .

Output: Samples from posterior distribution  $\nu(\Sigma^T | \mathbf{X}_1, \dots, \mathbf{X}_n)$ .
for  $l = 1$  to  $M$  do
    Update topology
    1   Compute  $\Phi(\Sigma^{T(l)}) = (\mathcal{E}_{T(l)}, \mathcal{L}_{T(l)})$ ;
    2   Randomly remove a split  $e_A \in \mathcal{E}_{T(l)}^I$ ;
    3   Randomly choose an internal split, say  $e_B$ , between two splits  $\{e_{A'}, e_{A''}\}$  different
        from  $e_A$  that are compatible with  $\mathcal{E}_{T(l)}^I \setminus \{e_A\}$  (Proposition 1 ensures that there
        exist two and only two such splits by setting  $k = 1$  with one internal edge  $e_A$ 
        shrunk to zero);
    4   Set edge length  $|e_B| = |e_A|$ ;
    5   Compute acceptance rate  $\alpha_{\mathcal{E}}$  from (10) and generate  $u \sim \text{Unif}(0, 1)$ ;
        if  $u \leq \alpha_{\mathcal{E}}$  then
            | Return  $\mathcal{E}_{T(l+1)} = \{e_B\} \cup \mathcal{E}_{T(l)} \setminus \{e_A\}$ ;
        else
            | Return the edge set  $\mathcal{E}_{T(l+1)} = \mathcal{E}_{T(l)}$ ;
    Update edge lengths
    for  $e_A \in \mathcal{E}_{T(l+1)}$  do
        6   Generate the new edge length  $|e'_A|$  from  $TN_{(0, \infty)}(e_A, \sigma_{\mathcal{L}})$ ;
        7   Compute acceptance rate  $\alpha_{\mathcal{L}}$  from (11) and generate  $u \sim \text{Unif}(0, 1)$ ;
        if  $u \leq \alpha_{\mathcal{L}}$  then
            | Update  $|e_A| = |e'_A|$  and return  $\mathcal{L}_{T(l+1)} = \{|e'_A|\} \cup \mathcal{L}_{T(l)} \setminus \{|e_A|\}$ ;
        else
            | Return  $\mathcal{L}_{T(l+1)} = \mathcal{L}_{T(l)}$ ;
    Return  $\Sigma^{T(l+1)} = \Psi(T(l+1))$ .

```

6 Simulation

We empirically demonstrate the utility of the proposed method through a series of simulation studies and show that the proposed method can restore the underlying ultrametric matrix under different true data generating mechanisms. By incorporating the geometry of ultrametric matrices, our method enables inference based on the posterior samples from the MCMC algorithm that makes local geometry-driven moves across on \mathcal{U}_p and is robust to mis-specified data generating mechanisms with a heavier tail.

Data generating mechanism. We generate a tree structure of p leaves as the true underlying tree T^0 and map the true tree structure to an ultrametric matrix of dimension p . Given the true ultrametric matrix Σ^{T^0} , we consider three data generating mechanisms of (i) correct specified normal distribution $X_i \stackrel{i.i.d.}{\sim} N(0, \Sigma^{T^0})$ and (ii) mis-specified t distribution $X_i \stackrel{i.i.d.}{\sim} t_{\nu}(0, \Sigma^{T^0})$ with degrees of freedom four and

three ($\nu = 3$ and 4). We generate the data with five different sample sizes of $n \in \{3p, 5p, 10p, 25p, 50p\}$ and 50 independent independent replicates. In this simulation, we obtain the true tree by the function `rtree` from the R package of `ape` (Emmanuel Paradis, 2012) and map the tree to the true ultrametric matrix shown in Panel (F) in Figure 5. We consider the dimension of $p = 10$ due to the intensive computation requirement for the competing method. (See Supplementary Material Section S1.6.)

We summarize the posterior samples by using the statistics in Section 5.2 in two aspects: (i) uncertainty quantification via the frequency of splits and element-wise 95% credible interval and (ii) point estimation with the representative matrices. For the uncertainty quantification, we first focus on the topology of the matrix of $\mathcal{E}_T = \Phi_V(V)$ and measure the topology recovery in terms of the splits frequency. For each split in posterior matrices, we compute the frequency of the tree topologies in the posterior samples that contain the true splits. We also investigate the coverage for each element in the matrix for the element-wise 95% credible interval. To our best knowledge, no existing method can directly quantify the uncertainty and be considered as the competing method. For the point estimation, we calculate the MAP tree and the mean tree (Miller et al., 2015) as representative trees, map the representative trees via $\Psi(\bar{T})$, and measure the matrix norm and the distance d of (3) between the true underlying matrix and the estimated matrices from representative trees. We compare the estimated matrices from our method to Bravo et al. (2009), which formulates the matrix estimation as a mixed-integer programming (MIP) problem. Under the matrix norm, we also consider the sample covariance which does not preserve ultrametric property. We assign priors of $\beta = -1.5$ as the uniform prior on all topology and independent $\exp(1)$ on the edge lengths. We run the MCMC in Algorithm 1 for 10,000 iterations and discard the first 9,000 iterations.

Uncertainty quantification. We quantify the uncertainty for ultrametric matrices by examining split-wise recovery in Table 1 and element-wise coverage in Figure 5. In Table 1, true splits with the corresponding edge lengths from the underlying true matrix are listed in the first row of Table 1. For each true split, we calculate the proportion of the posterior samples that contain the true split shown by each column. The split-wise recovery performs better when the sample size increases for all data generating mechanisms. For different data generating mechanisms, the correct specified model performs the best with sample size of $n = 50$ to ensure around 90% of the posterior samples having correct splits. On the other hand, to achieve a similar level of 90% split recovery, the mis-specified t-distribution requires sample size over 100 and 250 for t_4 and t_3 , respectively. Among all splits, we also observe that the split with a smaller length ($|e_{1,2,4}| = 0.231$) has the worst recovery (mean recovery of $|e_{1,2,4}|$ is 67.8% while the recovery for other split are around 90% for correct specified model with $n = 50$). We present the results of element-wise coverage of the 95% credible interval for the normal distribution in Figure 5. The results for t-distribution are available in Supplementary Material Section S1. Overall, the estimated coverage are high but slightly lower than the nominal coverage (around 0.75 to 0.94) and is higher when the sample size increase (medians of the coverage rates for sample size of (30, 50, 100, 250, 500) are (0.84, 0.78, 0.88, 0.82, 0.90)).

Point estimation. We show the distance from the estimated matrices to the true matrix in Figure 6. We observe that the matrices mapped from the mean and MAP trees from our method are comparable to

Table 1: Split-wise recovery for the proposed algorithm with $\exp(1)$ prior on the edge lengths under different sample size, data generating distribution. Each column shows proportion of posterior splits that contains specific split in the true underlying matrix. The average and the standard deviation of the proportion are obtained from 10,000 iterations with first 9,000 iterations discarded over 50 independent replicates.

Sample Size	Generating Distribution	1,2,3,4,5, 6,7,8,9	3,5,6,7,8,9	3,5,6,8,9	3,9	5,6,8	5,6	1,2,4	2,4
True Edge Lengths		0.701	0.872	0.712	0.88	0.878	0.854	0.231	0.869
30	Normal	79.7(23.7)	67.8(28.1)	82.3(20.3)	80.3(22.6)	83.2(22.7)	81.5(20.5)	43.8(27.6)	78.7(26.9)
	t_4	60.4(35)	72.2(32.6)	70.8(29.6)	55.5(37.8)	65.1(39.4)	68.4(33.1)	37(26.9)	74.2(31.7)
	t_3	59(38.8)	60.7(36.2)	72.2(30.2)	66.6(38.1)	68.5(36.7)	68(33.8)	42.9(33.7)	67.1(37.7)
50	Normal	86.1(18.8)	90.6(16.6)	94.8(10.1)	87.6(19)	90.5(18.4)	88.5(22)	67.8(25.2)	95(12.8)
	t_4	78(26.4)	82.8(27.5)	78.7(28)	76.8(30.4)	76.1(31.8)	85.9(23.2)	69.2(30.2)	90.2(22.9)
	t_3	72.9(31.5)	78.3(32.7)	85.7(23.3)	72.3(36.9)	75.4(37.6)	79.5(32.5)	46.8(36.4)	77.6(32)
100	Normal	95.2(9)	99.6(1)	98.6(4.4)	95.1(11.1)	98.5(4.3)	98.7(4.5)	87.2(17.9)	99.9(0.3)
	t_4	88.5(20)	93.2(18.3)	96.1(13.2)	87.5(23.8)	90(25.3)	95.2(13.9)	74.7(34.5)	93.2(23.9)
	t_3	71.4(37.3)	80.4(37.4)	80.9(37)	75.9(38.9)	76.1(39.5)	87.3(30.4)	62.8(40.8)	83.2(35)
250	Normal	99.6(3)	100(0.1)	100(0)	99.9(0.3)	100(0)	100(0)	98.3(6.4)	100(0)
	t_4	95.7(17.4)	100(0.1)	99.5(3.1)	98.9(6.7)	100(0.1)	97.9(14.1)	89.7(20.3)	100(0)
	t_3	79.3(38)	90.4(26.1)	89.6(29.6)	90.7(28.5)	87.8(31.3)	93.2(23.5)	86.8(30.3)	93(24)
500	Normal	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)	100(0)
	t_4	100(0)	98.1(13.4)	100(0)	100(0)	100(0)	100(0)	95.8(16)	100(0)
	t_3	92.9(24.8)	91.1(26.3)	94.1(23.4)	92.3(26.4)	95.5(19.8)	98(14.1)	88.2(30.6)	98(14.1)

the estimated matrix from Bravo et al. (2009) and sample covariance in terms of matrix norm (Panel (A) to (C)) and distance d from (3) (Panel (D) to (F)) across different data generating mechanisms and sample sizes. When the model is correctly specified (Panel (A) and (D)), all methods benefit from the increase of the sample size with a smaller distance to the true matrix (medians of matrix norm and distance d for sample sizes of (30, 50, 100, 250, 500) are (6.75, 5.89, 3.91, 2.50, 1.56) and (2.01, 1.53, 0.987, 0.627, 0.435), respectively). For the mis-specified scenario (Panel (B) and (E) for t_4 , and (C) and (F) for t_3), the advantage from the larger sample size is moderate (medians of matrix norm and distance d for sample sizes of (30, 50, 100, 250, 500) are (21.9, 24.6, 25.2, 23.7, 24.8) and (4.97, 4.46, 3.97, 3.25, 3.28) for t_4 , and (31.6, 39.5, 42.5, 44.0, 47.5) and (7.00, 6.70, 7.39, 6.10, 6.28) for t_3).

We also provide additional results for simulation including (i) convergence diagnostics, (ii) element-wise coverage for mis-specified t-distribution, (iii) the topology trajectory for the proposed method, and (iv) the simulation results for the data generated from a underlying tree with the same sum of the edge lengths from the root to all leaves. Comparing to other methods (e.g. MIP), we also observe that our proposed MCMC is relative scalable in terms of the number of leaves. Specifically, when the number of leaves is doubled from $p = 10$ to $p = 20$, the time required for MIP is 21.5 times, while it is only 3.5 times for the proposed method. All additional simulation studies are provided in Supplementary Material Section S1.

In summary, the geometry of the set \mathcal{U}_p provides three main inferential and computational advantages

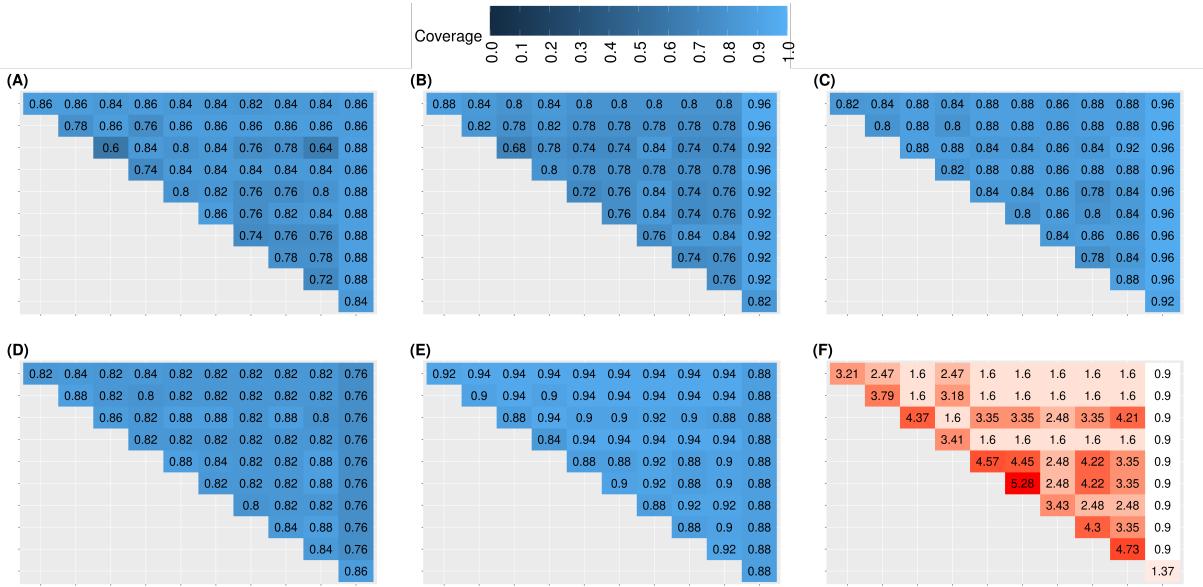


Figure 5: Element-wise coverage rates of the nominal 95% credible intervals for the correctly specified normal distribution with five different sample sizes. Panel (A) to (E) show the empirical coverage of 95% credible interval for sample size of 30, 50, 100, 250, and 500. The true underlying covariance is shown in Panel (F).

comparing to existing models: (i) enables the uncertainty quantification on ultrametric matrices along with the point estimator, (ii) is robust to mis-specified model, and (iii) is computational efficient in terms of convergence and scalability of number of leaves.

7 Analysis of Treatment Tree in Cancer

We demonstrate the proposed method on a pre-clinical patient-derived xenograft (PDX) data to discover promising cancer treatments. Due to the impracticality of testing multiple treatments on the same patient, PDX is a experiment design that evaluates multiple treatments administered to samples from the same human tumor implanted into genetically identical mice. The mice are then treated as the “avatars” to mimic potential responses to different treatments. In this analysis, we leverage a PDX dataset of Novartis Institutes for BioMedical Research - PDX Encyclopedia [NIBR-PDXE, (Gao et al., 2015)] that has collected over 1,000 PDX lines across multiple cancers with a $1 \times 1 \times 1$ design (one animal per PDX model per treatment).

For our analysis, we focus on cutaneous melanoma, which consists of 14 treatments and 32 PDX lines. The primary response is the tumor size difference before and after treatment administration, following the approach by Rashid et al. (2020), with the untreated group as the reference group. Positive responses indicate that the treatment shrunk the tumor more than the untreated group with a higher value representing a better efficacy. We assume that treatments with similar mechanism should induce similar levels of responses, and we aim to construct a tree structure to reveal the mechanism similarity based on the main responses. We ran our method with 10,000 iterations and discarded the first 9,000

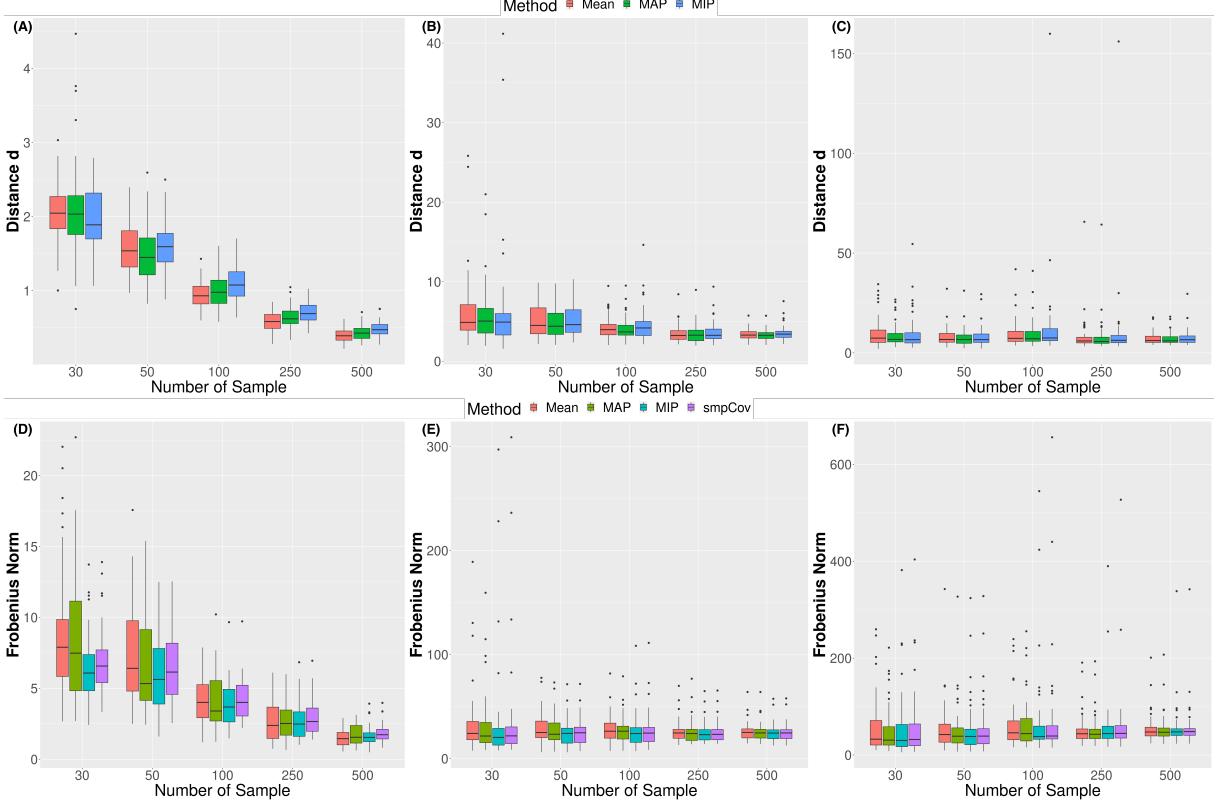


Figure 6: Distances between the estimated matrix and the true matrix under different data generating mechanism and sample sizes. The mean (red) and MAP (green) from our method is comparable to competing methods (blue for MIP and purple for sample covariance) in terms of the distance d from (3) ((A) to (C)) and matrix norm ((D) to (F)).

iterations. The convergence diagnostics of the algorithm are shown in Supplementary Material Section S2. We summarize the results with the MAP and mean trees and highlight subtrees with the frequency over 90%.

Figure 7 shows the mean (Panel (A)) and MAP trees (Panel (B)) with subtrees that consistently appear in the posterior samples. Two subtrees with frequencies higher than 90% are emphasized by boxes: blue (91%), and yellow (98%). We observe that the MAP and mean trees share many subtrees with the same topology. For example, the subtrees in the boxes are identical in both the mean and MAP trees. Additionally, two combination treatments highlighted by blue box form a tight sub-tree that appears over 90% of posterior samples, indicating a high level of mechanism similarity of two combination therapies. Two combination therapies consist of two agents, with one agent being encorafenib and the other targeting one of the following pathways: phosphoinositide 3-kinases (BKM120), and cyclin-dependent kinases (LEE011). As these pathways are closely related and share common downstream mechanisms (e.g., Repetto et al., 2018; Kurtzeborn et al., 2019), it is not surprising to see that all combination therapies form a tight subtree in the tree structure.

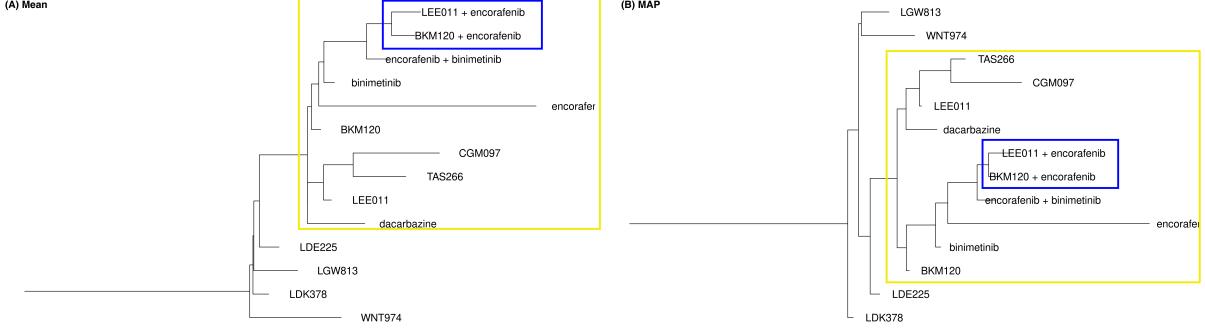


Figure 7: The mean (Panel (A)) and MAP trees (Panel (B)) for the melanoma. Two boxes emphasize the subtrees with high frequencies ($> 90\%$) in the posterior samples: blue: 91%, and yellow: 98%.

8 Discussion

In this paper, we develop a novel Bayesian framework that allows the inference on ultrametric matrices that makes local geometry-driven moves along geodesic within the set of ultrametric matrices \mathcal{U}_p . Leveraging the bijection map of the ultrametric matrix and the tree structure, we characterize the geometry of \mathcal{U}_p by pulling back the intrinsic stratified geometry of the BHV space onto \mathcal{U}_p . The homeomorphism further constructs the decomposition of (5) that represents the ultrametric matrices by coordinates of BHV space. The ‘‘pulling back’’ also allows Markovian and consistent priors on ultrametric matrices. By utilizing the geometry of the BHV, we design an algorithm that draws posterior ultrametric matrices without relaxation or projection, which allows us to summarize the posterior samples through existing tree modeling tools. In simulation studies, our proposed algorithm produces element-wise credible intervals with nominal coverage rates, while having point estimates comparable with existing projection-based method in terms of matrix recovery errors. As a byproduct, the posterior samples of the tree topology also affords direct calculation of the posterior probability of any split as another way of assessing recovery. We demonstrate our method in a preclinical dataset and discover that subsets of treatments sharing high mechanism similarities which aligns with existing literature.

The present work considered trees with strictly positive leaf lengths. Choice of the Euclidean distance on $\mathbb{R}_{>0}^{p+1}$ resulted in using the Lebesgue measure $d\mathbf{x}$ when defining the prior distribution on \mathcal{U}_p , via one on \mathcal{T}_{p+1} . If it is desired for the prior distribution to place negligible mass on trees with infinitesimally small leaf lengths, one can use a Riemannian volume measure related to an alternative distance arising from a Riemannian metric on the positive orthant $\mathbb{R}_{>0}^{p+1}$ under which trees with leaf lengths zero, or close to zero, are infinitely far away. An example is the induced Riemannian metric on $\mathbb{R}_{>0}^{p+1}$ obtained by pulling back the standard metric on \mathbb{R}^{p+1} under the map $F(\mathbf{x}) = (\log x_1, \dots, \log x_{p+1})^\top$:

$$\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{x}} = \sum_{i=1}^{p+1} \frac{a_i b_i}{x_i^2},$$

for two vectors \mathbf{a}, \mathbf{b} in the tangent space $T_{\mathbf{x}}(\mathbb{R}_{>0}^{p+1})$. Such a specification would result in a new geometry on \mathcal{U}_p , and the proposed inferential framework would need to be suitably modified.

We list five important future directions that may further improve the utility of the proposed method.

First, in certain contexts, latent tree models for phylogenetic studies assume that trees have edge lengths proportional to the length of time between speciation events represented by each node. This results in a sum constraint on the edge lengths from the root to leaves. Geometry of the BHV treespace is unsuitable for such trees. Suitable modification (e.g., Gavryushkin and Drummond, 2016) that results in a similar stratified space, to which the current geometric framework can be extended, can be used.

Second, the bijection $\Phi : \mathcal{U}_p \rightarrow \mathcal{T}_{p+1}$ may lead to a useful set of coordinates in terms of the matrices (V, D) for further study of $(\mathcal{T}_{p+1}, d_{\text{tree}})$, especially in simplifying computations presently used on the popular treespace.

Third, from a modelling perspective, for trees containing a large number of leaves (large p), the prior construction in Section 4 makes possible development of Bayesian models based on principled variable selection priors that decouples topology and edge length information on latent trees. For example, spike-and-slab priors on the coefficients in the representation (5) will allow modelling of sparse representations of Σ^T corresponding to trees T of a fixed topology.

Fourth, it is possible to develop variants of Algorithm 1 that explicitly makes moves to or between boundary components of positive codimensions. This will extend the utility of the proposed algorithm for use on latent tree models with non-binary trees. Such algorithms require specifying a prior with full support on \mathcal{U}_p , including the boundaries. For example, the two-parameter Poisson-Dirichlet distribution from a Gibbs fragmentation model (McCullagh et al., 2008) is appropriate in this context.

Finally, a natural extension to the latent tree model is to introduce covariates \mathbf{z} through the decomposition (5) that leads to case-specific ultrametric covariance matrices $\Sigma^{T(\mathbf{z})}$. A prior distribution using the construction in Section 4 will assign weights on different tree topologies and edge lengths depending on \mathbf{z} , while its fragmentation process-based definition will allow borrowing of information as \mathbf{z} varies. We leave these topics for future research.

Acknowledgements

KB acknowledges support from grants NSF 2015374 and NIH R37-CA21495. The work is partly supported by a Michigan Institute for Data Science (MIDAS) seed grant to ZW. VB acknowledges support from grants NIH P30 CA046592 and NIH R01 CA244845-01A1.

Appendix

We need the following Lemma on the ultrametric matrix corresponding to any tree $T \in \mathcal{T}_{p+1}$, including those on the boundary.

Lemma 1. Σ^T corresponding to any tree $T \in \mathcal{T}_{p+1}$ is a positive definite strictly ultrametric matrix.

Proof. The origin $\{\mathbf{0}\} \times \mathbb{R}_{>0}^{p+1}$ of \mathcal{T}_{p+1} is the set of trees with a single edge of positive length emanating from the root 0 and p edges of positive lengths connecting the same parent to p leaves. From decomposition (2) the ultrametric matrices for such tree are of the form

$$d_1 \mathbf{1} \mathbf{1}^\top + \sum_{i=1}^p d_{j_i} \mathbf{u}_i \mathbf{u}_i^\top,$$

where $\{\mathbf{u}_i, i = 1, \dots, p\}$ is the standard basis vectors in \mathbb{R}^p with one in the i th position and zero elsewhere, and d_{j_1}, \dots, d_{j_p} is a subsequence from $\{d_i\}$ in (2) corresponding to leaf edges. Clearly, such matrices satisfy property (ii) in Definition 1, and are strictly ultrametric. They are also positive definite. To see this, note that the second term in the sum gives rise to positive definite diagonal matrices with positive entries. For such a diagonal matrix W , note that $W + d_1 \mathbf{1}\mathbf{1}^\top$ is its rank-one perturbation with scale $d_1 > 0$, and symmetric. From the Sherman-Morrison formula, $|W + d_1 \mathbf{1}\mathbf{1}^\top| > 0$ since $1 + d_1 \mathbf{1}W^{-1}\mathbf{1} > 0$ when $d_1 > 0$.

Growing an internal edge to a tree at the origin to move to the boundary component of codimension $(p - 1)$ translates to the operation

$$d_1 \mathbf{1}\mathbf{1}^\top + \sum_{i=1}^p d_{j_i} \mathbf{u}_i \mathbf{u}_i^\top + d_{j_{p+1}} \mathbf{v} \mathbf{v}^\top,$$

where \mathbf{v} is a binary vector and $\mathbf{v}\mathbf{v}^\top$ is of rank 1 (Nabben and Varga, 1994) with $d_{j_{p+1}} > 0$, on ultrametric matrices. Using the argument from above, such matrices are positive definite and strictly ultrametric. Adding additional internal edges amounts to performing repeated rank-one perturbations of a positive definite matrix with positive d_{j_k} , $k = p + 2, \dots, 2p - 1$, which preserves strict ultrametricity and positive definiteness. \square

Proof of Theorem 1

Proof. We prove the result for fully resolved binary trees T and the corresponding strictly ultrametric matrices Σ^T ; the proof for unresolved trees on the boundary of \mathcal{T}_{p+1} follows along similar lines once identifications, or gluing, of the common boundaries are accounted for.

We first establish that Φ is a bijection by using the representation $\Sigma^T = VDV^\top$ in (2), where $V \in \{0, 1\}^{p \times 2p-1}$ possesses the partition property and $D \in \mathbb{R}^{(2p-1) \times (2p-1)}$ is a diagonal matrix with positive entries. We then show that Φ is continuous with a continuous inverse with respect to the pullback or quotient topology from \mathcal{T}_{p+1} , which will imply that Φ is indeed a homeomorphism.

Consider two maps $V \mapsto \Phi_V(V) = \mathcal{E}_T$ and $D \mapsto \Phi_D(D) = \mathcal{L}_T$ that, respectively, map a V with the partition property to a compatible edge set \mathcal{E}_T and D to a vector in $\mathcal{L}_T \cong \mathbb{R}_{>0}^{2p-1}$. We note that $\Phi_D(D) = \sum_{i=1}^{2p-1} \mathbf{e}_i^\top D \mathbf{e}_i \mathbf{e}_i$, where $\{\mathbf{e}_i\}$ is the standard basis in \mathbb{R}^{2p-1} .

The map Φ_V will be defined implicitly via a construction described below. Note first that the representation $\Sigma^T = VDV^\top$ decouples V and D , and thus

$$\Phi(\Sigma^T) = (\Phi_V(V), \Phi_D(D)), \quad \Sigma^T \in \mathcal{U}_p.$$

For each edge e_A associated with a split A define the unique p -dimensional binary vector b_A with ones at indices that are in A and zero for indices in A^c . Arrange the vectors into a $p \times (2p - 1)$ binary matrix $B = (b_{A_1}, \dots, b_{A_{2p-1}})^\top$ corresponding to the $(2p - 1)$ edges in a fully resolved T . In order to relate the columns of B to those of V possessing the partition property, we use the logical **and** operator \wedge on columns of B . In other words, the compatibility criterion that one of $A_1 \cap A_2, A_1 \cap A_2^c, A_1^c \cap A_2$ associated with two splits e_{A_1} and e_{A_2} be empty translates to one of $b_{A_1} \wedge b_{A_2}, \bar{b}_{A_1} \wedge b_{A_2}, b_{A_1} \wedge \bar{b}_{A_2}$ equalling the zero vector $\mathbf{0}$, where \bar{b} denotes the negation of b . The construction defines the map Φ_V .

Every column vector v of V maps to an edge of a tree T the matrix Σ^T uniquely determines (Dellacherie et al., 2014, Proposition 3.14), and thus to an edge e_A associated with a split $L = A \cap A^c$ on the leaves indexed by $L = \{0, 1, \dots, p\}$ of T . Clearly, every point in \mathcal{L}_T is the image of some D . Thus to every tree $T \in \mathcal{T}_{p+1}$ in orthant j there exists an ultrametric matrix $\Sigma^T = VDV^\top$ such that V uniquely identifies stratum j and D the point within it. The maps Φ_V and Φ_D are surjective, and when combined with Lemma 1 the map Φ is surjective.

The map Φ_D is clearly injective. Injectivity of Φ_V is established once it is shown that V , which encodes topology of a tree T , uniquely determines an edge set \mathcal{E}_T of compatible splits that identifies an orthant in \mathcal{T}_{p+1} . From the partition property of V we thus need to prove that any triplet (v_i, v_j, v_k) of columns vectors in V satisfy $v_i = v_j + v_k$ if and only if there exists a unique triplet $(e_{A_i}, e_{A_j}, e_{A_k})$ of edges/splits in T that are mutually pairwise compatible.

Each column v_i of V has ones at component indices corresponding to the set of leaves that are descendants of edge i , and maps to a column in B , since they are of the same dimension. Owing to the compatibility condition amongst the edges, such a column in B is then necessarily unique. Every triplet (v_i, v_j, v_k) of columns in V thus map to a unique triplet $(b_{A_i}, b_{A_j}, b_{A_k})$ of columns in B . Then, with \vee is the logical or operator, we note that $v_i = v_j + v_k$ if and only if $b_{A_i} = b_{A_j} \vee b_{A_k}$ and $b_{A_j} \wedge b_{A_k} = \mathbf{0}$ while each of $\bar{b}_{A_j} \wedge b_{A_k}, b_{A_j} \wedge \bar{b}_{A_k}$ does not equal $\mathbf{0}$, rendering the splits corresponding to the pair (b_{A_j}, b_{A_k}) compatible. Similarly, $b_{A_i} \wedge b_{A_j} \neq \mathbf{0}$, and only one of $\bar{b}_{A_i} \wedge b_{A_j}, b_{A_i} \wedge \bar{b}_{A_j}$ equals $\mathbf{0}$, since otherwise $b_{A_i} \neq b_{A_j} \vee b_{A_k}$. The splits corresponding to pair (b_{A_i}, b_{A_j}) are hence compatible; a similar argument is used to show compatibility of (b_{A_i}, b_{A_k}) . This shows that Φ_V is injective. Since both Φ_D and Φ_V are injective, the map Φ is injective. We have thus established that Φ is bijective with inverse Ψ .

Consider the metric topology on \mathcal{T}_{p+1} consisting of open sets defined by open balls with respect to the metric d_{tree} . Define the quotient topology

$$\tau_{\mathcal{U}_p} = \{\Psi(A) : A \text{ is open in } \mathcal{T}_{p+1}\}.$$

By definition, $\tau_{\mathcal{U}_p}$ is a topology of open sets on \mathcal{U}_p . Φ is bijective and continuous with respect to $\tau_{\mathcal{U}_p}$ since the preimage of an open set is open. It is an open map, since by definition, $\Phi(U)$ is open in \mathcal{T}_{p+1} for every $U \in \tau_{\mathcal{U}_p}$. This implies that Φ is a homeomorphism.

To turn $(\mathcal{T}_{p+1}, d_{\text{BVH}})$ into a geodesic metric space we need to ensure that geodesics in \mathcal{T}_{p+1} map to geodesics in \mathcal{U}_p with respect to the pullback metric d . This is true if it is shown that \mathcal{U}_p is a *length space* with the induced metric d . Define

$$\tilde{d}(\Sigma_1^T, \Sigma_2^T) := \inf \left\{ \text{Len}(\Phi \circ \gamma) : \text{a continuous curve } \gamma : [0, 1] \rightarrow \mathcal{U}_p, \gamma(0) = \Sigma_1^T, \gamma(1) = \Sigma_2^T \right\};$$

here Len measures the length of a curve $\sigma : [0, 1] \rightarrow \mathcal{T}_{p+1}$ defined as

$$\text{Len}(\sigma) := \sup \sum_{i=1}^k d_{\text{tree}}(\sigma(t_{i-1}), \sigma(t_i)),$$

where the supremum is taken over all k and all sequences $t_0 \leq t_1 \leq \dots \leq t_k$ in $[0, 1]$. Note the \tilde{d} is indeed a metric since \mathcal{T}_{p+1} is Hausdorff. The homeomorphic property of Φ ensures that it is a covering map of \mathcal{T}_{p+1} and turns $(\mathcal{U}_p, \tilde{d})$ into a length space with $\tilde{d} = d$ (Bridson and Haefliger, 1999, Chapter I.3).

The BHV space $(\mathcal{T}_{p+1}^I, d_{\text{BHV}})$ is a CAT(0) space (Billera et al., 2001, Lemma 4.1). The space $(\mathbb{R}_{>0}^{p+1}, \|\cdot\|_2)$ is Euclidean and hence CAT(0), and $(\mathcal{T}_{p+1}, d_{\text{tree}})$ as a product of two CAT(0) spaces is thus CAT(0) (Bridson and Haefliger, 1999). In addition to being a homeomorphism the map Φ is an (local) isometry (Bridson and Haefliger, 1999, Proposition 3.25), and we can hence pull back the nonpositive curvature of \mathcal{T}_{p+1} onto \mathcal{U}_p making it globally CAT(0).

Inclusion of the leaf edge lengths ensures that each stratum of \mathcal{T}_{p+1} can now be identified with the product $\mathbb{R}_{\geq 0}^{p-2} \times \mathbb{R}_{>0}^{p+1}$, a stratified space with the strata glued isometrically along their shared boundaries of only the first factor in the product. Then \mathcal{U}_p is prescribed a stratification under Ψ with the same $(2p - 3)!!$ number of strata each of the same dimension as that of \mathcal{T}_p . The proof is now complete. \square

Proof of Proposition 1

Proof. On a codimension one boundary, from the decomposition in (5), we note that the unresolved tree corresponding to $\tilde{\Sigma}^T$ is obtained by creating a degree 4 vertex. Starting from such an unresolved tree there are exactly three possible ways to grow an edge and create a fully resolved tree. This procedure identifies the three unique ultrametric matrices $\Sigma_p^T, \Sigma_q^T, \Sigma_r^T$ on three distinct strata p, q, r . The claim on k -codimensional boundaries follows from the corresponding picture in the BHV space (Billera et al., 2001). The order inequality on $\tilde{\Sigma}^T$ follows upon noting that an element Σ_{ij}^T of Σ^T is the sum of edge lengths from the root to the most recent common ancestor of i and j . \square

References

- Agrawal, R., Roy, U., and Uhler, C. (2020). Covariance Matrix Estimation under Total Positivity for Portfolio Selection*. *Journal of Financial Econometrics*, 20(2):367–389.
- Aldous, D. (1996). Probability distributions on cladograms. In Aldous, D. and Pemantle, R., editors, *Random Discrete Structures*, pages 1–18, New York, NY. Springer New York.
- Anderson, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. *Essays in Probability and Statistics*, pages 1–24.
- Barden, D., Le, H., and Owen, M. (2018). Limiting behaviour of Fréchet means in the space of phylogenetic trees . *Annals of the Institute of Statistical Mathematics*, pages 99–129.
- Berestycki, N. and Pitman, J. (2007). Gibbs distributions for random partitions generated by a fragmentation process. *Journal of Statistical Physics*, 127(2):381–418.
- Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767.
- Bravo, H. C., Wright, S., Eng, K. H., Keles, S., and Wahba, G. (2009). Estimating tree-structured covariance matrices via mixed-integer programming. *J Mach Learn Res*, 5:41–48.

- Bridson, M. R. and Haefliger, A. (1999). *Metric Spaces of Non-Positive Curvature*. Springer Berlin, Heidelberg.
- Brown, D. G. and Owen, M. (2020). Mean and variance of phylogenetic trees. *Sytematic Biology*, 69:139–154.
- Buneman, P. (1974). A note on the metric properties of trees. *Journal of Combinatorial Theory, Series B*, 17:48–50.
- Choi, M. J., Tan, V. Y. F., and Anandkumar, A. and Willsky, A. S. (2011). Learning latent tree graphical models. *Journal of Machine Learning Research*, 12:1771–1812.
- Cox, D. R. and Wermuth, M. (1993). Linear dependencies represented by chain graphs (with discussion). *Statistical Science*, 8:204–218.
- Dawid, A. P. and Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable models. *Annals of Statistics*, 21:1272–1317.
- Dellacherie, C., Martinez, S., and San Martin, J. (2014). *Inverse M-Matrices and Ultrametric Matrices*. Springer Cham.
- Emmanuel Paradis (2012). *Use R ! Analysis of Phylogenetics and Evolution with R*.
- Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, pages 471–492.
- Felsenstein, J. (2003). *Inferring Phylogenies*. Oxford University Press.
- Gao, H., Korn, J. M., Ferretti, S., Monahan, J. E., Wang, Y., Singh, M., Zhang, C., Schnell, C., Yang, G., Zhang, Y., Balbin, O. A., Barbe, S., Cai, H., Casey, F., Chatterjee, S., Chiang, D. Y., Chuai, S., Cogan, S. M., Collins, S. D., Dammassa, E., Ebel, N., Embry, M., Green, J., Kauffmann, A., Kowal, C., Leary, R. J., Lehar, J., Liang, Y., Loo, A., Lorenzana, E., Robert McDonald, E., McLaughlin, M. E., Merkin, J., Meyer, R., Naylor, T. L., Patawaran, M., Reddy, A., Röelli, C., Ruddy, D. A., Salangsang, F., Santacroce, F., Singh, A. P., Tang, Y., Tinnetto, W., Tobler, S., Velazquez, R., Venkatesan, K., Von Arx, F., Wang, H. Q., Wang, Z., Wiesmann, M., Wyss, D., Xu, F., Bitter, H., Atadja, P., Lees, E., Hofmann, F., Li, E., Keen, N., Cozens, R., Jensen, M. R., Pryer, N. K., Williams, J. A., and Sellers, W. R. (2015). High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.*, 21(11):1318–1325.
- Garba, M. K., Nye, T. M. W., Lueg, J., and Huckemann, S. F. (2021). Information geometry for phylogenetic trees. *Journal of Mathematical Biology*, 82:19.
- Gavryushkin, A. and Drummond, A. J. (2016). The space of ultrametric phylogenetic trees. *Journal of Theoretical Biology*, 403:197–208.
- Karlin, S. and Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. i. multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4):467–498.

- Karlin, S. and Rinott, Y. (1983). M-matrices as covariance matrices of multinormal distributions. *Linear Algebra and its Applications*, 52-53:419–438.
- Khare, K. and Rajaratnam, B. (2011). Wishart distributions for decomposable covariance graph models. *Annals of Statistics*, 39:514–555.
- Kim, J. (2000). Slicing hyperdimensional oranges: The geometry of phylogenetic estimation. *Molecular Phylogenetics and Evolution*, 17:58–75.
- Kurtzeborn, K., Kwon, H. N., and Kuure, S. (2019). MAPK/ERK Signaling in regulation of renal differentiation. *Int J Mol Sci*, 20(7).
- Lauritzen, S., Uhler, C., and Zwiernik, P. (2019). Maximum likelihood estimation in Gaussian models under total positivity. *Annals of Statistics*, 47(4):1835–1863.
- Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.
- Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs. *Annals of Statistics*, 35:1278–1323.
- Leung, L. and Drton, M. (2018). Algebraic tests of general gaussian latent tree models. In *Advances in Neural Information Processing Systems*.
- Lueg, J., Garba, M. K., Nye, T. M. W., and Huckemann, S. F. (2022). Foundations of the Wald Space for Phylogenetic Trees. [arxiv.org/pdf/2209.05332](https://arxiv.org/pdf/2209.05332.pdf).
- McCullagh, P. (2006). Structured covariance matrices in multivariate regression models. Technical report, Department of Statistics, University of Chicago.
- McCullagh, P., Pitman, J., and Winkel, M. (2008). Gibbs fragmentation trees. *Bernoulli*, 14(4):988 – 1002.
- Miller, E., Owen, M., and Provan, J. S. (2015). Polyhedral computational geometry for averaging metric phylogenetic trees. *Advances in Applied Mathematics*, 68:51–91.
- Moulton, V. and Steel, M. (2004). Peeling phylogenetic oranges. *Advances in Applied Mathematics*, 33:710–727.
- Nabben, R. and Varga, R. S. (1994). A Linear Algebra Proof that the Inverse of a Strictly Ultrametric Matrix is a Strictly Diagonally Dominant Stieltjes Matrix. *SIAM Journal on Matrix Analysis and Applications*, 15(1):107–113.
- Neal, R. (2003). Density Modeling and Clustering Using Dirichlet Diffusion Trees. *Bayesian Statistics*, 7:619–629.
- Nye, T. M. (2011). Principal components analysis in the space of phylogenetic trees. *Annals of Statistics*, 39:2716–2739.

- Nye, T. M. (2020). Random walks and Brownian motion on cubical complexes. *Stochastic Processes and their Applications*, 130(4):2185–2199.
- Ostrowski, A. (1937). Über die determinanten mit überwiegender Hauptdiagonale. *Commentarii Mathematici Helvetici*, 10:69–96.
- Owen, M. and Provan, J. S. (2011). A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(1):2–13.
- Pitman, J. (2006). *Combinatorial stochastic processes*, volume 1875.
- Rashid, N. U., Luckett, D. J., Chen, J., Lawson, M. T., Wang, L., Zhang, Y., Laber, E. B., Liu, Y., Yeh, J. J., Zeng, D., and Kosorok, M. R. (2020). High-dimensional precision medicine from patient-derived xenografts. *Journal of the American Statistical Association*, 0(0):1–15.
- Repetto, M. V., Winters, M. J., Bush, A., Reiter, W., Hollenstein, D. M., Ammerer, G., Pryciak, P. M., and Colman-Lerner, A. (2018). CDK and MAPK synergistically regulate signaling dynamics via a shared multi-site phosphorylation region on the scaffold protein Ste5. *Mol Cell*, 69(6):938–952.
- Semple, C., Steel, M., et al. (2003). *Phylogenetics*, volume 24. Oxford University Press.
- Shiers, N., Zwiernik, P., Aston, J. A. D., and Smith, J. Q. (2016). The correlation space of Gaussian latent tree models and model selection without fitting. *Biometrika*, 103:531–545.
- Slawski, M. and Hein, M. r. (2015). Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields. *Linear Algebra and its Applications*, 473:145–179.
- Sturmefels, B., Uhler, C., and Zwiernik, P. (2021). Brownian Motion Tree Models Are Toric. *Kybernetika*, 56(6):1154–1175.
- Wang, Y., Zhang, L., N., and Chen, T. (2008). Latent tree models and approximate inference in Bayesian networks. *Journal of Artificial Intelligence Research*, 32:879–900.
- Yao, T.-H., Wu, Z., Bharath, K., Li, J., and Baladandayuthapan, V. (2023). Probabilistic learning of treatment trees in cancer. *Annals of Applied Statistics*, page Forthcoming.
- Yule, G. U. (1925). Ii.—a mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f. r. s. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213(402-410):21–87.
- Zwiernik, P. (2016). *Semialgebraic Statistics and Latent Tree Models*. CRC Press.
- Zwiernik, P., Uhler, C., and Richards, D. (2017). Maximum likelihood estimation for linear gaussian covariance models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(4):1269–1292.