Advances in Applied Mathematics 68 (2015) 51–91

Contents lists available at ScienceDirect

Advances in Applied Mathematics

www.elsevier.com/locate/yaama

Polyhedral computational geometry for averaging metric phylogenetic trees

Ezra Miller [a], Megan Owen [b,*], J. Scott Provan [c]

[a] *Department of Mathematics, Duke University, Durham, NC 27708, USA*
[b] *Department of Mathematics and Computer Science, Lehman College, CUNY, Bronx, NY 10468, USA*
[c] *Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-3180, USA*

ARTICLE INFO

Article history:
Received 16 February 2014
Received in revised form 23 April 2015
Accepted 23 April 2015

MSC:
68U05
05C05
52B99
62P10
92D15
60B05
52-04
62H99
52A41
90C57
92B10
92-08
53C23

Keywords:
Tree space
Fréchet mean
Polyhedral subdivision
Descent method

ABSTRACT

This paper investigates the computational geometry relevant to calculations of the Fréchet mean and variance for probability distributions on the phylogenetic tree space of Billera, Holmes and Vogtmann, using the theory of probability measures on spaces of nonpositive curvature developed by Sturm. We show that the combinatorics of geodesics with a specified fixed endpoint in tree space are determined by the location of the varying endpoint in a certain polyhedral subdivision of tree space. The variance function associated to a finite subset of tree space has a fixed C^∞ algebraic formula within each cell of the corresponding subdivision, and is continuously differentiable in the interior of each orthant of tree space. We use this subdivision to establish two iterative methods for producing sequences that converge to the Fréchet mean: one based on Sturm's Law of Large Numbers, and another based on descent algorithms for finding optima of smooth functions on convex polyhedra. We present properties and biological applications of Fréchet means and extend our main results to more general globally nonpositively curved spaces composed of Euclidean orthants.

* Corresponding author.
E-mail addresses: ezra@math.duke.edu (E. Miller), megan.owen@lehman.cuny.edu (M. Owen), scott_provan@unc.edu (J.S. Provan).

http://dx.doi.org/10.1016/j.aam.2015.04.002
0196-8858/

Phylogenetics
Nonpositively curved space

0. Introduction

The development of statistical methods for studying phylogenetic trees, and in particular the search for meaningful notions of consensus tree for phylogenetic data, has been of considerable importance in biology for four decades. Starting with the problem as posed by Adams [1], a great deal of research has been done, and a myriad of definitions proposed, relating to consensus trees in phylogenetics; see [9] for an excellent overview. The problem has been confounded by the combinatorial nature of the trees themselves. According to Cranston and Rannala [11], "Phylogenetic inference has long been troubled by the difficulty of performing statistical analysis on tree topologies. The topologies are discrete, categorical, and non-nested hypotheses about the species relationships. They are not amenable to standard summary analyses such as the calculation of means and variances and cause difficulties for many traditional forms of hypothesis testing." Other papers share concerns about issues such as these [5,16].

The introduction by Billera, Holmes, and Vogtmann of phylogenetic tree space [7] opened statistical analysis of tree-like data to a wide and computationally tractable variety of techniques [17]. Tree space, with its geodesic distance, is a *globally nonpositively curved* (abbreviated to *global NPC*) space, and as a result it has convexity properties that imply uniqueness of means as well as other important statistical and geometric objects, while also giving a framework for effective computational methods to calculate these objects. One of the major uses of the convexity properties was the discovery by Owen and Provan [27] of a fast algorithm for computing geodesics in this space (see Section 1 for this algorithm as well as the background tree space geometry necessary to state it). Chakerian and Holmes [10] subsequently showed that phylogenetic tree space provides an excellent platform for implementing several distance-based statistical techniques, and Nye [24] has shown how this space can be used to perform principal component analysis on tree data.

Perhaps the two most fundamental concepts of interest in statistical analysis of data are that of *sample mean* (or *average*) and its associated *variance*. The basic goal of this paper is to demonstrate the computational effectiveness of certain notions of statistical mean and variance for probability distributions on tree space. The average that we use is the *Fréchet mean*, or *barycenter*: the point in tree space that minimizes its sum of squared geodesic distances to the sample points (Section 2). Our decision to use this definition is motivated by work of Sturm [32], who identified the Fréchet mean as a theoretically rich statistical object associated with sampling from a specified distribution on a global NPC space (see Theorem 2.4). Fréchet means in tree space and the algorithm for computing them that arises from Sturm's work (Algorithm 2.5) have been independently developed by Bačák [4].