Tropical pricipal component analysis on the space of ultrametrics

Robert Page Ruriko Yoshida Leon Zhang

1 Introduction

This document is a supplement for the paper "Tropical principal component analysis on the space of phylogenetic trees".

2 Mixture of coalescent models

In order to compare our results, we also applied the same simulated data sets to `geophytter` which approximate the BHV PCA on the BHV tree space [3].

3 Sensitivity analysis

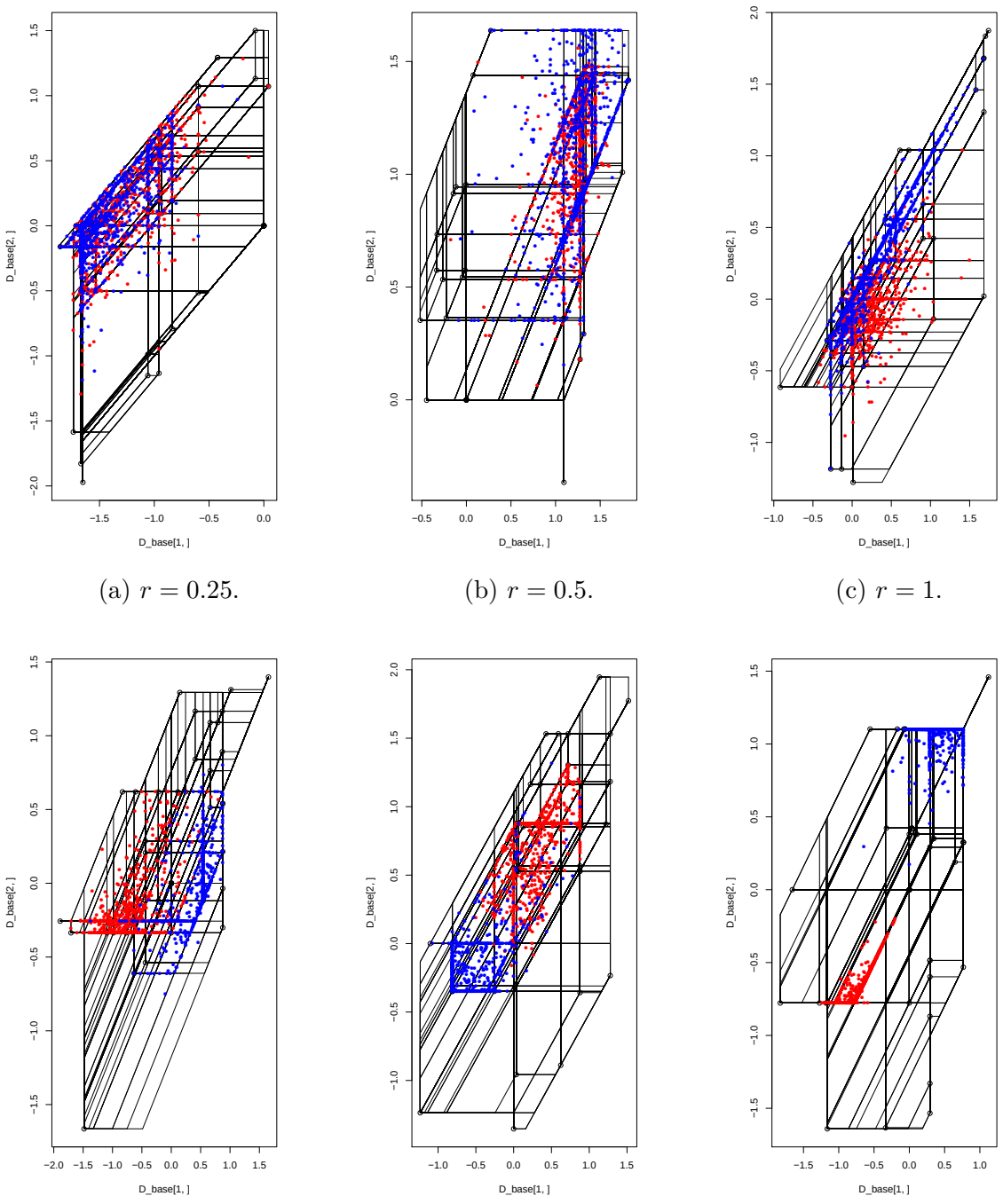
We examine the sensitivity of our MCMC method for estimating the second-order tropical principal polytope from the given data. We conducted this sensitivity analysis by running our MCMC approach 10 times on the same set of data and studying how the corresponding value of R changed. We applied this approach on two types of dataset: one with a fixed underlying tree topology and one in which the datapoints were essentially random.

4 Empirical data

We applied our method to three empirical data sets: Apicomplexa gene trees [1], the African coelacanth genome [2], and flu virus data [4].

4.1 African coelacanth genome data

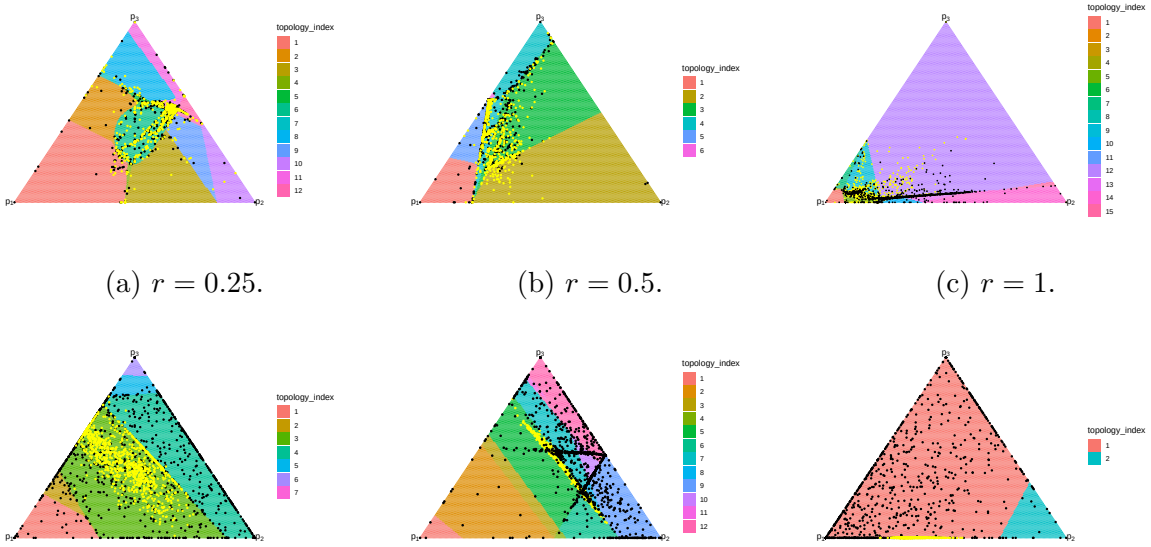
We applied our MCMC technique to estimate the tropical principal polytope of the dataset consisting of 1,290 genes on 690,838 amino acid residues obtained from genome and transcriptome data [2]. The result is shown in Figure 5. In Figure 5 each tree topology of a projection onto the second order tropical principal polytope has a color and black color represents tree topologies in the lower five percentile.



(a) $r = 0.25$. (b) $r = 0.5$. (c) $r = 1$.

(d) $r = 2$. (e) $r = 5$. (f) $r = 10$.

Figure 1: We applied tropical PCAs on the mixture of two coalescent distributions using Algorithm 5.1. We colored blue for projected trees whose gene trees are generated from one coalescent distribution and red for the other distribution. We varied the ratio r.



(a) $r = 0.25$. (b) $r = 0.5$. (c) $r = 1$.

(d) $r = 2$. (e) $r = 5$. (f) $r = 10$.

Figure 2: We applied BHV PCAs on the mixture of two coalescent distributions. We colored blue for projected trees whose gene trees are generated from one coalescent distribution and red for the other distribution. We varied the ratio r.

4.2 Apicomplexa

The second empirical dataset we have applied is from 268 orthologous sequences with eight species of protozoa presented in [1]. This data set has gene trees reconstructed from the following sequences: *Babesia bovis* (Bb), *Cryptosporidium parvum* (Cp), *Eimeria tenella* (Et) [15], *Plasmodium falciparum* (Pf) [11], *Plasmodium vivax* (Pv), *Theileria annulata* (Ta), and *Toxoplasma gondii* (Tg). An outgroup is a free-living ciliate, *Tetrahymena thermophila* (Tt).

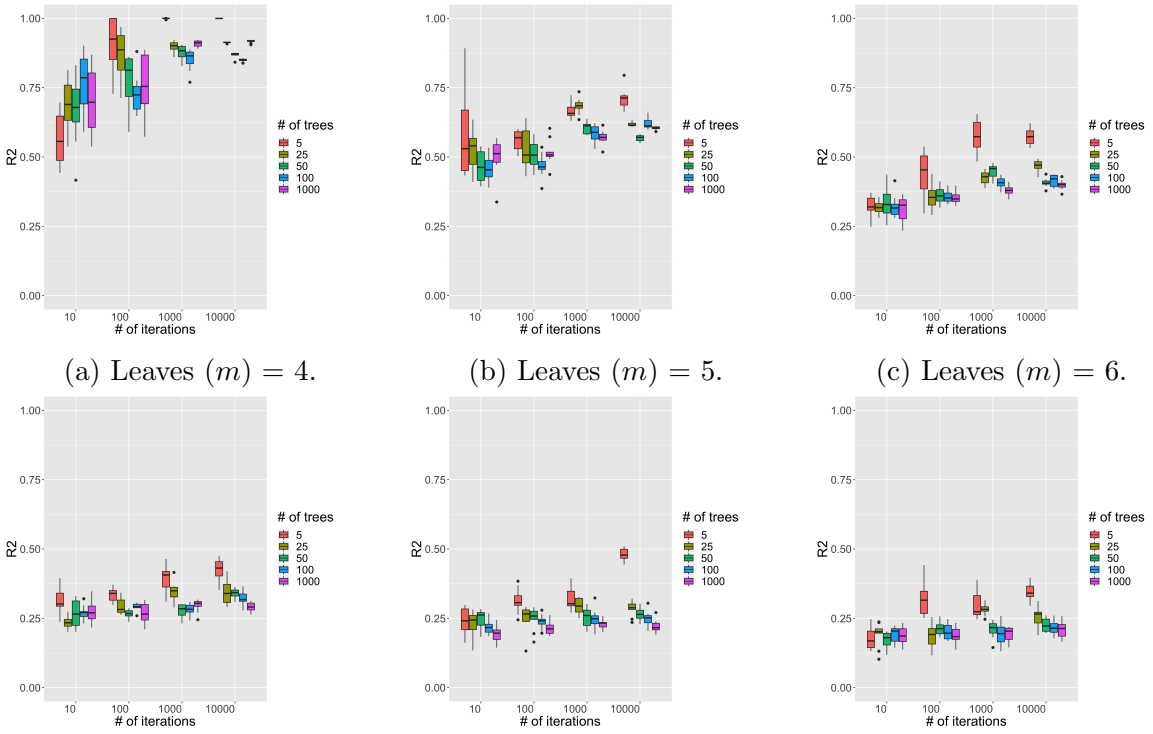
The result is shown in Figure 6. In Figure 6 each tree topology of a projection onto the second order tropical principal polytope has a color and black color represents tree topologies with their frequencies in the lower five percentile.

References

[1] C. Kuo, J. P. Wares, and J. C. Kissinger. The apicomplexan whole-genome phylogeny: An analysis of incongruence among gene trees. *Mol Biol Evol*, 25(12):2689–2698, 2008.

[2] D. Liang, X. X. Shen, and P. Zhang. One thousand two hundred ninety nuclear genes from a genome-wide survey support lungfishes as the sister group of tetrapods. *Mol. Biol. Evol.*, 30(8):1803–1807, 2013.

[3] T. M. W. Nye, X. Tang, G. Weyenberg, and R. Yoshid. Principal component analysis and



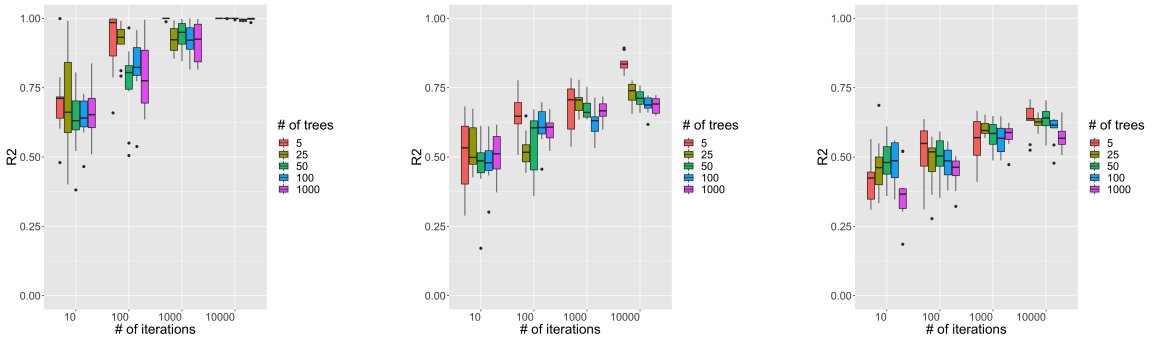
(a) Leaves $(m) = 4$. (b) Leaves $(m) = 5$. (c) Leaves $(m) = 6$.

(d) Leaves $(m) = 7$. (e) Leaves $(m) = 8$. (f) Leaves $(m) = 9$.

Figure 3: We applied the MCMC approach to compute a second order tropical principal polytope on the datasets of random trees. The y-axis represents R and the x-axis represents the number of iterations for a MCMC.

the locus of the fréchet mean in the space of phylogenetic trees. *Biometrika*, 104:901–922, 2017.

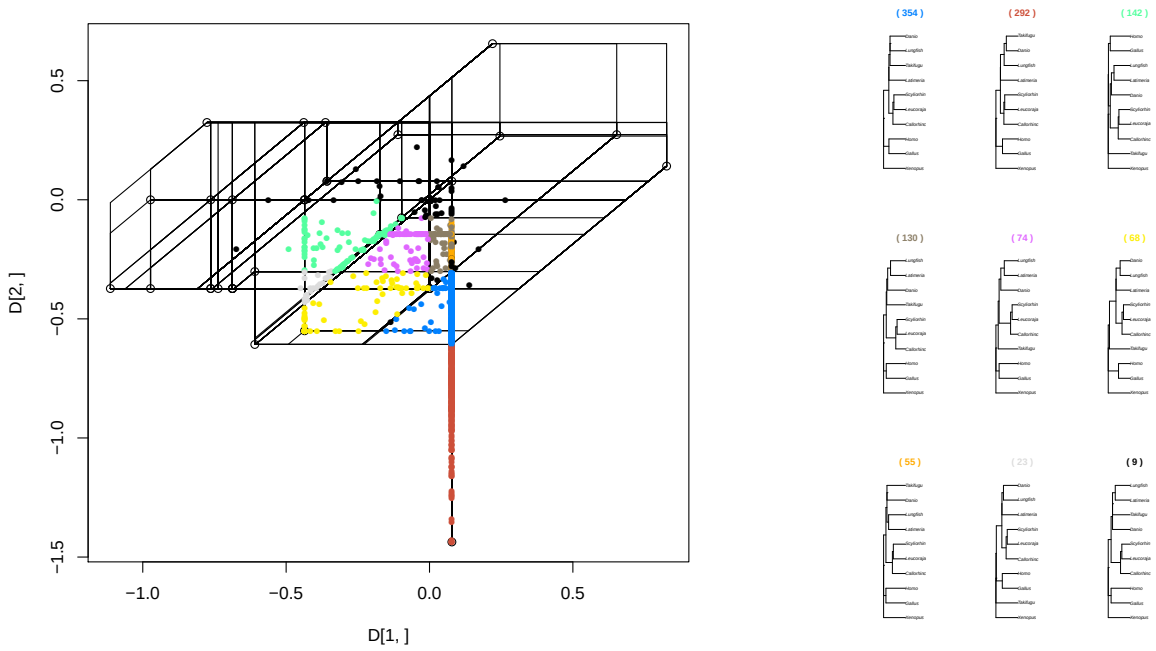
[4] S. Zairis, H. Khiabanian, A.J. Blumberg, , and R. Rabadán. Tropical principal component analysis and its application to phylogenetics. *Bulletin of Mathematical Biology*, 81:568–597, 2019.



(a) Leaves $(m) = 4$. (b) Leaves $(m) = 5$. (c) Leaves $(m) = 6$.

(d) Leaves $(m) = 7$. (e) Leaves $(m) = 8$. (f) Leaves $(m) = 9$.

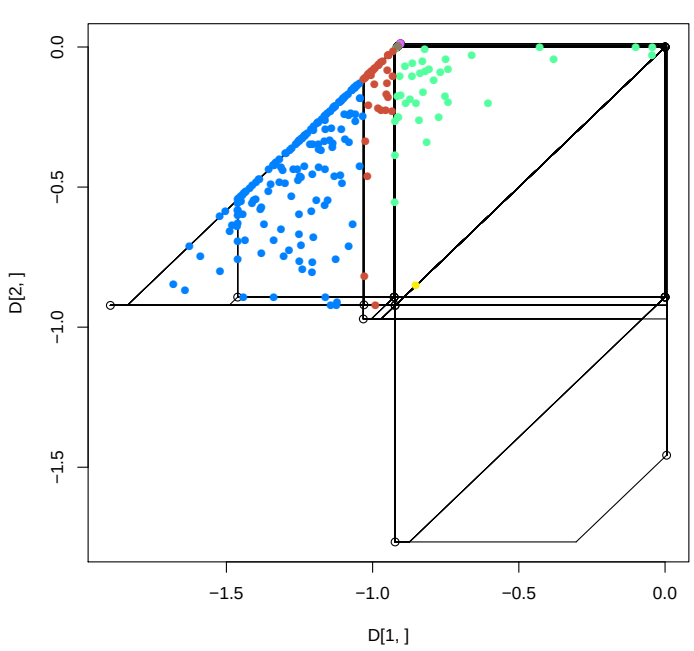
Figure 4: We applied the MCMC approach to compute a second order tropical principal polytope on datasets of trees with the same underlying topology. We ran 10 Markov chains and we computed a box plot for each case. The y-axis represents R and the x-axis represents the number of iterations for a MCMC.



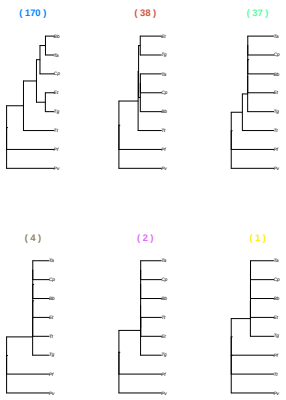
(a) Second order tropical principal polytope for African coelacanth genome data. Black colored dots are trees with the tree topologies with frequencies in the lower 5 percentile.

(b) Tree topologies projected on the tropical principal polytope.

Figure 5: Estimated tropical principal polytope of African coelacanth genome data via our MCMC method.



(a) Second order tropical principal polytope for Apicomplexa gene data. Black colored dots are trees with the tree topologies with frequencies in the lower 5 percentile.



(b) Tree topologies projected on the tropical principal polytope.

Figure 6: Estimated tropical principal polytope of gene trees on Apicomplexa data set via our MCMC method.