A litte bit of algebraic geometry for phylogenetic invariants

M. Casanellas

April 27, 2017

When one wants to use phylogenetic invariants on a tree T, one has to

- first select a set of phylogenetic invariants on T $I = \{f_1, \dots, f_r\}$ ($f_1(p) = 0$ for any distribution p arising on T)[1]
- and when one uses this set, one is considering the whole set of solutions to these polynomials, lets call it $V(I)$:
$$V(I) = \{p | f_1(p) = \dots f_r(p) = 0\}$$
(sets like this are called *algebraic varieties*).

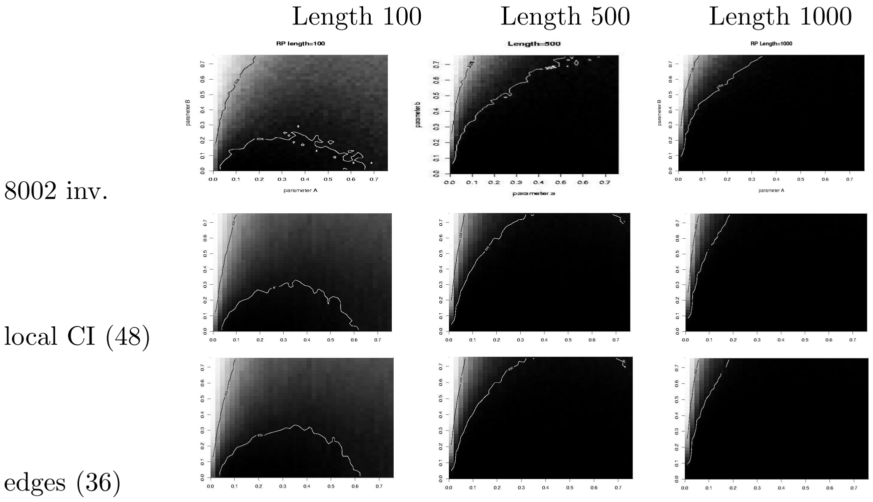
It may happen that this set $V(I)$ does not only contain the distributions in your tree, but contains also many other points: $V(I) = D_T \cup Z$, where D_T is the set of distributions on T. This is precisely what happens when you use the "full" set of phylogenetic invariants.

If we want to get rid of Z and detect the part that is relevant to us, D_T, one can use a smallest set of equations. Namely, taking into account that our $p \in D_T$ arise from a Markov processes in the tree where the substitution matrices shouldn't be too far from the identity matrix, it is possible to find a smallest set of invariants that "defines" the set D_T around these kind of points. Finding this smallest set (which is called a *local complete intersection*) is a hard task that we've done in [**?**], [**?**], but it is very useful: e.g. for a quartet tree on K81 model, it reduces the 8002 phylogenetic invariants that generate the full set of phylogenetic invariants down to 48 phylogenetic invariants!

Finally, if we only want to use invariants for topology reconstruction purposes, we are only interested in *topology invariants*. Also, we have to bear in mind that when one wants to infer the tree topology, one is actually assuming that the distribution comes from some tree, that is the points p we are interested in satisfy $p \in V_{T_1} \cup \dots V_{T_m}$ (it $T_1, \dots, T_m$ is the set of all tree topologies). This is an assumption we have not yet used. Thus, we actually want those invariants that define V_{T_i} inside this union (that is, if we already knew the invariants that vanish on this union, we only need to care about the extra invariants that vanish on V_{T_i} and not on the whole union). What we prove in [1] is that the set of *edge invariants* can play this role. For the K81 example above, this reduces the set to 36 polynomials.

So I don't think it would be more powerful to consider the full set of invariants. I attach a figure where you can see the performance of these sets of equations on quartet trees for the K81 model, which actually shows that using the full set of invariants is not a good idea and using the edge invariants is the same as using the local complete intersection.

[1] The set of phylogenetic invariants is infinite but there are finite sets $\{f_1, \dots, f_r\}$ that generate all invariants, that is, any other invariant f is an algebraic combination of them (this is a Theorem of Noether form 100 years ago). This is what you called a "full" set of invariants I guess.



References

[1] M. Casanellas and J. Fernandez-Sanchez. Geometry of the Kimura 3-parameter model. *Advances in Applied Mathematics*, 41:265–292, 2008.

[2] M. Casanellas and J. Fernández-Sánchez. Relevant phylogenetic invariants of evolutionary models. *J. Math. Pure. Appl.*, 96:207–229, 2010.

[3] M. Casanellas, J. Fernańdez-Sánchez, and M. Michalek. Complete intersection for equivariant models. arxiv:1512.07174.