

Identifiability of Phylogenetic Models

Elizabeth S. Allman¹, Cécile Ané², John A. Rhodes¹

1. Dept. of Mathematics and Statistics

University of Alaska Fairbanks

2. Depts. of Statistics and Botany

Univeristy of Wisconsin Madison

Evolution 2007 – Christchurch, NZ

June 16–20, 2007



Identifiability:

Defn: A statistical model is *identifiable* if the values of all model parameters can be determined from a predicted distribution of data.

Ex: For a distribution from the Jukes-Cantor model,
by first computing JC distances it is possible to determine

- 1) tree topology, and
- 2) edge lengths.

Hence the JC model is identifiable.

Identifiability of models is necessary to have *consistency* of statistical inference, whether using ML or Bayesian methods.

Defn: An inference method is *consistent* if the correct parameter values will be inferred from ‘perfect data’.

Parameters = tree topology, edge lengths, Γ rate-distribution parameter, proportion of invariable sites, etc.

Perfect data = infinite amounts of data in perfect accord with the assumed model.

Consistency is a *minimal* requirement for an inference method.

It says nothing about behavior with

- 1) finite amounts of data (*efficiency*), or
- 2) misspecified model (*robustness*).

Q: Is ML, Bayesian inference with GTR+ Γ +I model consistent?

This is fundamental to statistical underpinnings of much current work in phylogenetics.

Rogers (Sys. Biol., 2001) — claimed a proof, widely cited, but

Argument has several major gaps in showing identifiability:

- 1) crucial use of an unjustified graphical claim
- 2) generic vs. non-generic parameters

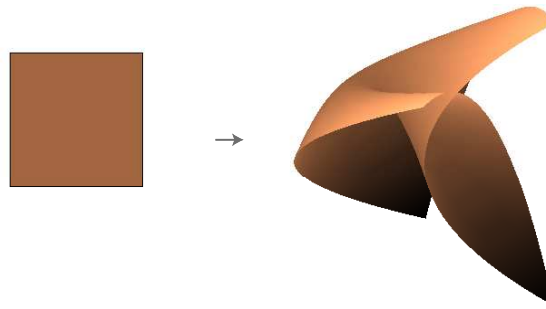
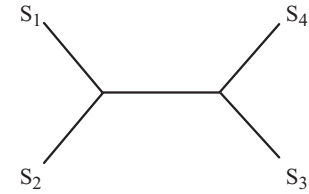
It is not clear whether Rogers' argument can be fixed. For full discussion of technical details of gaps, see

<http://www.dms.uaf.edu/~jrhodes/rogers.pdf>

↪ There is **no valid, published proof** that ML or Bayesian inference using the GTR+ Γ +I model is consistent.

Generic vs. non-generic identifiability

Consider a fixed tree topology,
all possible numerical parameters



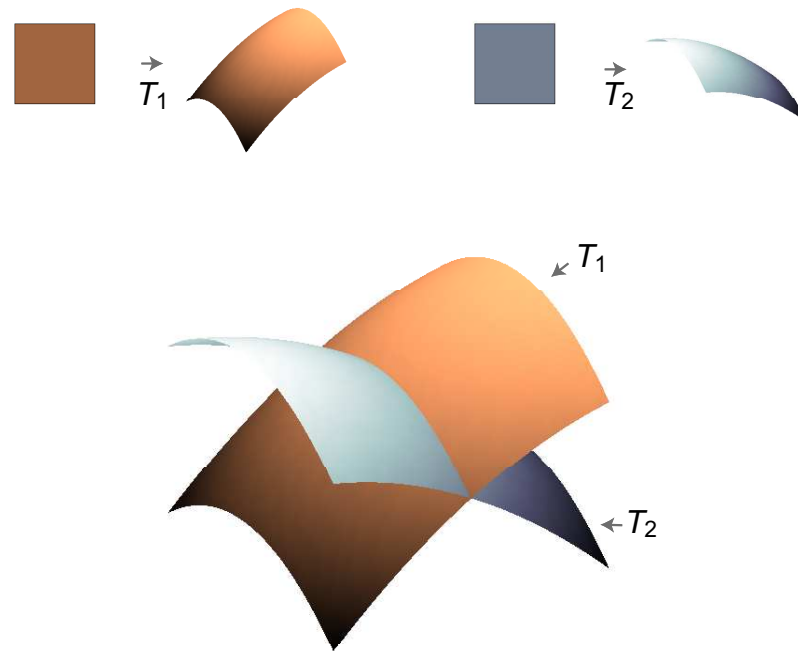
parameter space \rightarrow 'perfect data surface'

(edge lengths, rates, etc.) (pattern frequencies)

Parameters are **identifiable** when **no self intersection** in 'surface'.

Parameters are **generically identifiable** if **self-intersection** is of **lower-dimension** than surface.

Consider two *different* tree topologies:



Tree topologies are **identifiable** when **no intersection** of two surfaces.

Tree topologies are **generically identifiable** if **intersection is of lower dimension** than surfaces.

Known Identifiability results

Negative:

- For sufficiently complicated rate-across-sites models (non-explicit), tree identifiability can fail (Steel-Székely-Hendy, J. Comp. Biol., 1994)
- explicit **non-generic** examples (not r-a-s) of non-identifiability of mixtures (Štefankovič-Vigoda, Sys. Biol., 2007; J. Comp. Biol., 2007)
- **non-generic** 2-class mixtures on one tree can exactly agree with 1-class model on different tree (Matsen-Steel, preprint)
- more general study of many-class non-identifiable mixtures under 2-state symmetric model (Matsen-Mossel-Steel, preprint)

Positive:

- GTR is identifiable — (e.g., use log-det distance to identify tree, etc.)
- GM is identifiable (Chang, Math. Biosci., 1996)
- general result on mixture models on one tree with ‘small’ number of classes (Allman-Rhodes, J. Comp. Biol., 2006)
 - For DNA models, tree is **generically** identifiable for:
 - GTR+I
 - GTR with 3 rate-across-sites classes
 - GTR+GTR+GTR
 - GM+GM+GM
 - covarion
- 2-tree mixtures on 4-leaf trees are **generically** identifiable for GM, GTR (Allman-Rhodes, in preparation)

But **none** of this work applies to $GTR+\Gamma$ or $GTR+\Gamma+I$, since:

- continuous rate distribution prevents application of Allman-Rhodes positive results (or methods of proof)
- specifying a particular form of rate distribution prevents application of negative Steel or Matsen-Mossel-Steel results.

New result:

Allman, Ané, Rhodes (2007):

For 4-state (DNA) models, $GTR+\Gamma$ is identifiable.

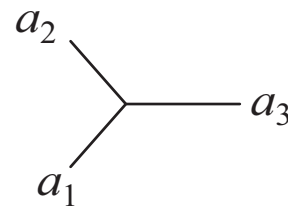
For κ -state models, $GTR+\Gamma$ is generically identifiable.

Note:

1. This is the first proof of identifiability for a rate-across-sites model with a continuous distribution of rates.
2. Identifiability for *all* parameters, not just generic ones.
3. Proof does not follow Rogers' approach.

Main points of GTR+ Γ proof:

- Focus on 3-leaf tree to identify parameters



Result for n -leaf tree then follows from **combinatorial** arguments.

- Use **algebraic** arguments to extract information from 3-dimensional data tensor.
- For generic parameters, **analytic** arguments, using convexity, give identifiability.
- More detailed analysis of non-generic cases completes proof.

Note: We still lack a proof that the tree is identifiable for GTR+ Γ +I.

This is likely to be significantly harder to prove since:

- Γ introduces only 1 parameter (shape parameter α),
- Γ +I introduces 2 parameters (α , proportion of invar. sites p_{inv})

(For the mathematical details, a draft proof is available upon request.)