Identifiability of Models from Parsimony-Informative Pattern Frequencies

John A. Rhodes

University of Alaska

Fairbanks

June 10, 2008

MIEP

Joint work with

Elizabeth Allman (UAF)
Mark Holder (U Kansas)

Thanks to the Isaac Newton Institute

I: Parsimony-informative models:

- Variants of standard Markov substitution models on trees where *only* parsimony-informative patterns are observed

- Useful for phenotypic datasets — acquisition bias prevents appropriate sampling of non-informative character patterns (e.g., all equal, all different)

- Despite shortcomings of simple models for phenotypic datasets, statistical approaches such as ML, Bayesian inference might still be preferable to parsimony

- Model proposed by P. Lewis (2001) omits constant patterns; model of Ronquest–Hulsensebeck (2004?) omits parsimony-noninformative patterns; used for combined analysis of sequence and morphological data by Nylander–Ronquest–Hulsenbeck–Nieves-Aldrey (2004)

For this talk focus on

$\mathrm{GM2}_{\text{pars-inf}}$: 2-state General Markov model, with only parsimony-informative characters observed

Parameters: Tree, 2×2 Markov matrix on each edge,

arbitrary root distribution

$\mathrm{CFN}_{\text{pars-inf}}$: Cavender-Farris-Neyman model, with only parsimony-informative characters observed

Submodel of $\mathrm{GM2}_{\text{pars-inf}}$ with symmetric Markov matrics,

uniform root distribution

But much generalizes to k-state models, $k > 2$ (in progress...)

II: Identifiability:

For a fixed model,

Given an exact distribution of site-patterns arising from the model

— infinite amounts of ‘perfect’ data —

can we determine all model parameters?

Identifiability is necessary for statistical consistency of inference

Tree identifiability:

Theorem (Steel–Hendy–Penny, 1993): Identifiability of 4-taxon tree topologies fails for $\mathrm{CFN}_{\textsf{pars-inf}}$ (and hence for $\mathrm{GM2}_{\textsf{pars-inf}}$).

Proof is to explicitly give two parameter sets leading to same distribution of parimony-informative patterns.

Theorem (Allman-Holder-R): Suppose all Markov matrix parameters are non-singular and have all positive entries. Then topologies of n-taxon trees are identifiable for $\mathrm{GM2}_{\text{pars-inf}}$ (and hence $\mathrm{CFN}_{\text{pars-inf}}$) for $n \geq 8$.

Proof:

- Enough to identify all 4-taxon subtrees.

- For subtree relating taxa a_1, a_2, a_3, a_4, fix some choice of parsimony-informative pattern at all *other* taxa

- Consider only patterns extending this choice to $a_1, \ldots, a_4$.

- Observed frequencies of these extended patterns satisfy certain phylogenetic invariants depending on the 4-taxon topology.

(Invariants are inspired by the 4-point condition using a log-det distance – Cavender-Felsenstein, Steel)

Note: Identifiability of topologies for 5-, 6-, 7-taxon trees unknown.

Numerical parameter identifiability:

Suppose

- the tree topology is known,
- all Markov matrix parameters are non-singular, and
- some parsimony-informative pattern has positive probability of being observed

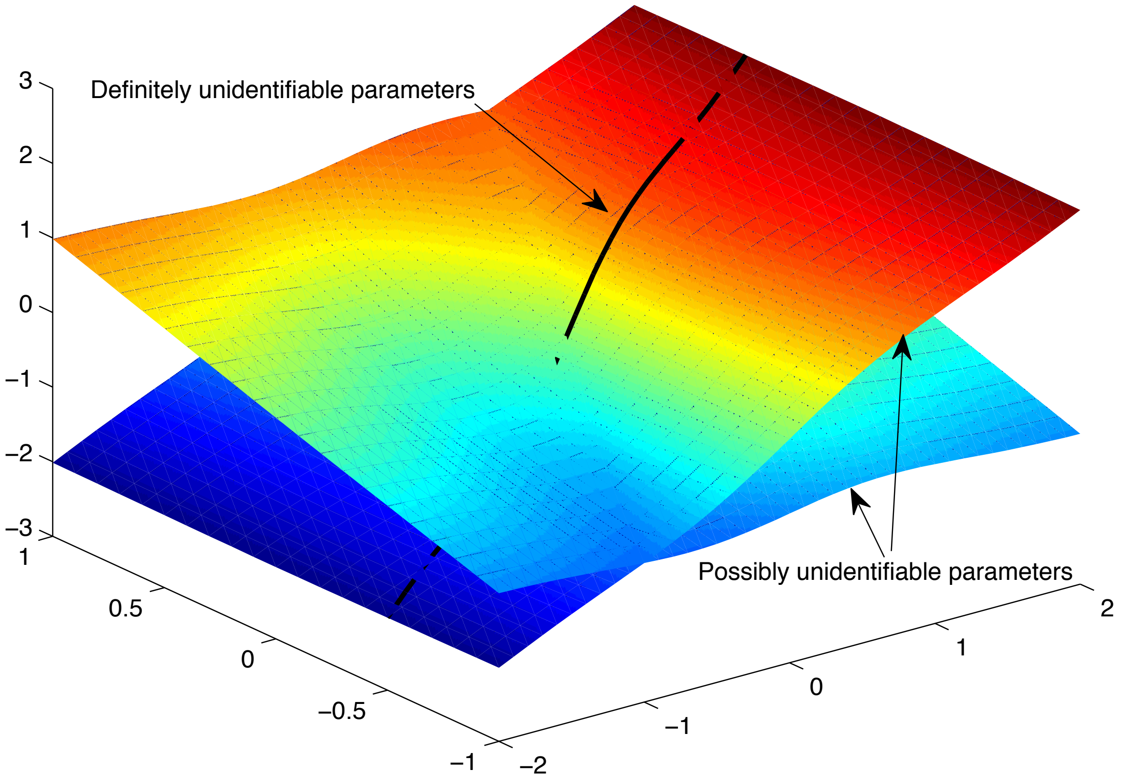
Theorem (Allman-Holder-R): For an n-taxon tree with $n \geq 7$, all numerical parameters of $\mathrm{GM2}_{\textsf{pars-inf}}$ are identifiable, up to 'label-swapping' at internal nodes. Hence numerical parameters of $\mathrm{CFN}_{\textsf{pars-inf}}$ are identifiable.

Theorem (Allman-Holder-R): For a 5-taxon tree generic numerical parameters of $\mathrm{GM2}_{\textsf{pars-inf}}$ are identifiable, up to 'label-swapping' at internal nodes.

However, there exists a subset of codimension 1 in the parameter space for which identifiability may fail.

Within this subset of potentially non-identifiable parameters, there is a smaller subset of codimension 2 in the full parameter space for which identifiability definitely fails.

Cartoon of parameter space for 5-taxon trees:

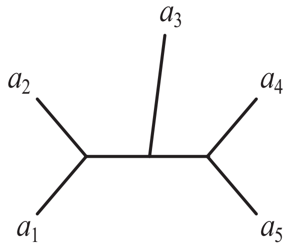


Specializing to $\mathrm{CFN}_{\textsf{pars-inf}}$, generic parameters are identifiable.

However, the potentially non-identifiable parameters for 5-taxon trees include those from ultrametric (molecular clock) trees!

Sketch of method of proof of identifiabilty of numerical parameters: We use

Theorem (Allman–R, 2008): For the 2-state General Markov model on a 5-taxon binary tree as shown, let $\{0, 1\}$ denote the set of character states. Let $p_{i_1 i_2 i_3 i_4 i_5}$ denote the joint probability of observing state i_j in the sequence at leaf a_j, $j = 1, \dots, 5$.



Then the ideal of phylogenetic invariants for this model are generated by the 3×3 minors of the following two matrices:

$$\begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} & p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} & p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} & p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} & p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}$$

and

$$\begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} \\ p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} \\ p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} \\ p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} \\ p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}.$$

If we have only probabilities q of patterns conditioned on parsimony-informativeness, then we know only *some* of these entries, but rescaled by an unknown factor.

$$\begin{pmatrix}
\mathbf{q}_{00000} & \mathbf{q}_{00001} & \mathbf{q}_{00010} & q_{00011} & \mathbf{q}_{00100} & q_{00101} & q_{00110} & q_{00111} \\
\mathbf{q}_{01000} & q_{01001} & q_{01010} & q_{01011} & q_{01100} & q_{01101} & q_{01110} & \mathbf{q}_{01111} \\
\mathbf{q}_{10000} & q_{10001} & q_{10010} & q_{10011} & q_{10100} & q_{10101} & q_{10110} & \mathbf{q}_{10111} \\
q_{11000} & q_{11001} & q_{11010} & \mathbf{q}_{11011} & q_{11100} & \mathbf{q}_{11101} & \mathbf{q}_{11110} & \mathbf{q}_{11111}
\end{pmatrix}$$

Red entries are unknown; 3×3 minors must still be zero.

Judicious choices of 3×3 minors allows for determination of unknown entries, provided certain 2×2 minors don't vanish. E.g.,

$$\begin{vmatrix} q_{01001} & q_{01010} & q_{01011} \\ q_{10001} & q_{10010} & q_{10011} \\ q_{11001} & q_{11010} & \mathbf{q}_{11011} \end{vmatrix} = 0,$$

Expanding the determinant in cofactors by the last column we have

$$q_{01011} \begin{vmatrix} q_{10001} & q_{10010} \\ q_{11001} & q_{11010} \end{vmatrix} - q_{10011} \begin{vmatrix} q_{01001} & q_{01010} \\ q_{11001} & q_{11010} \end{vmatrix} + \mathbf{q}_{11011} \begin{vmatrix} q_{01001} & q_{01010} \\ q_{10001} & q_{10010} \end{vmatrix} = 0$$

Thus provided

$$\begin{vmatrix} q_{01001} & q_{01010} \\ q_{10001} & q_{10010} \end{vmatrix} \neq 0$$

we can determine $\mathbf{q}_{11011}$ from other $q_{\mathbf{i}}$ where $\mathbf{i} \in S$.

For 5-taxon trees, enough 2×2 minors may be zero to defeat this approach, but still gives understanding of potential non-identifiability.

For trees with at least 7 taxa, enough 2×2 minors must be non-zero to determine all unknown entries.

Determining scaling factor is easy – sum of $p_{\mathbf{i}}$ is 1.