Model selection for mixture models

Anna Magdalena Kedzierska
joint work with Marta Casanellas and Jesus Fernandez-Sanchez

Department de Matemàtica Aplicada I
Universitat Politècnica de Catalunya
Centre de Regulació Genòmica

Special Sessions on Advances in Algebraic Statistics
March 28, 2010

Content

1 Sequence evolution

2 Mixture models

3 Identifiability

4 Summary

Markov process along a tree

- T- unrooted tree on random variables with states $B = \{A, C, G, T\}$
- $L(T)$- observed bases at the taxa, $n = \mid L(T) \mid$
- (M_e)- transition matrices associated to the edges $e \in E(T)$
- $\mathrm{p} = \mathrm{p}_{x_1 \cdots x_n} := \mathbb{P}(X_1 = x_1, \cdots, X_n = x_n)$ is the joint probability of the observed states at $L(T)$

Equivariant Evolutionary Models

Equivariant models are in agreement with certain symmetries of the transition matrices imposed by biological assumptions.

general Markov model (GM), general Kimura 3-parameter (K81), Kimura 2-parameter (K80), general Jukes-Cantor (JC69), strand symmetric model (CS05)

no restrictions

$$\begin{pmatrix} a & b & c & d \\ e & f & g & h \\ j & k & l & m \\ n & o & p & q \end{pmatrix}$$

$\cdots$

a+3b=1

$$\begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix}$$

Group structure

Let $\mathbf{G} \subseteq \mathfrak{S}_4$ be group of permutations on the set of nucleotides B and $W := \mathbb{C}^4 = \langle B \rangle_{\mathbb{C}}$ a vector space over $\mathbb{C}$ generated by B.

Definition

An equivariant evolutionary model is a pair $\mathcal{M}^{\mathbf{G}} = (\mathbf{G}, W)$ such that the transition matrices preserve the action of $\mathbf{G}$ on W.

Group action

Example ($JC69, \mathbf{G}^{JC69} = \mathfrak{S}_4 = \langle (ACGT), (AC) \rangle$)

$$\begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix} \qquad (AC) \mapsto \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Permutations of rows and columns leave the matrix unchanged.

JC69 transition matrices coincide with the matrices invariant under the group action $\mathbf{G}^{JC69}$. We call such set of matrices $\mathtt{Hom}_{\mathbf{G}^{JC}}(W, W)$

$\mathbf{G}^{GM} = id, \mathbf{G}^{K81} = \langle (AC)(GT), (AG)(CT) \rangle, \mathbf{G}^{K80} = \langle (ACGT), (AG) \rangle,$
$\mathbf{G}^{JC69} = \mathfrak{S}_4$

The entries of M_e are the model parameters and depend on the chosen model of evolution.
For a given tree, we have a map

$$\Psi_T : \text{parameters} \quad \mapsto \quad \mathcal{L} = \{\text{joint probabilities } \mathrm{p} \in \mathbb{R}^{4^n}\}$$

If we consider an equivariant model $\mathcal{M}^{\mathbf{G}}$, then the parametrization map Ψ_T restricts to:

$$\Psi_T^{\mathbf{G}} : \prod_{e \in E(T)} \mathtt{Hom}_{\mathbf{G}}(W, W) \quad \mapsto \quad \mathcal{L}^G := \{\mathrm{p} : \mathrm{p}_{x_1 \ldots x_n} = \mathrm{p}_{g(x_1) \cdots g(x_n)} \forall g \in \mathbf{G}\}$$

$$\mathcal{L}^G \subseteq \mathcal{L} \cong \mathbb{R}^{4^n}$$

$\mathcal{L}^G$ depends on the model, NOT on a particular choice of T.

Description of $\mathcal{L}^G$

We will see that the inclusion

$$\mathcal{L}^G := \{\mathrm{p} : \mathrm{p}_{x_1 \ldots x_n} = \mathrm{p}_{g(x_1)\cdots g(x_n)}\} \subseteq \mathcal{L}$$

is defined by a set of linear polynomial equations. Furthemore, those are the (linear) invariants for any T on $\mathcal{M}^{\mathbf{G}}$.

$\mathcal{L}^G$ for $\mathbf{G}^{K81}$

$\mathbf{G}^{K81} = \{id, (AC)(GT), (AG)(CT), (AT)(CG)\}$. For each element $g \in \mathbf{G}^{K81}$ consider $\{\mathtt{p}_x : \mathtt{p}_x = \mathtt{p}_{g(x)}, x = (x_1, \cdots, x_n)\}$.

$S = 4^{n-1}$

$$
\begin{array}{ll}
p_{\mathtt{AAA}} = p_{\mathtt{CCC}} = p_{\mathtt{GGG}} = p_{\mathtt{TTT}} & p_{\mathtt{AGA}} = p_{\mathtt{CTC}} = p_{\mathtt{GAG}} = p_{\mathtt{TCT}} \\
p_{\mathtt{AAC}} = p_{\mathtt{CCA}} = p_{\mathtt{GGT}} = p_{\mathtt{TTG}} & p_{\mathtt{AGC}} = p_{\mathtt{CTA}} = p_{\mathtt{GAT}} = p_{\mathtt{TCG}} \\
p_{\mathtt{AAG}} = p_{\mathtt{CCT}} = p_{\mathtt{GGA}} = p_{\mathtt{TTC}} & p_{\mathtt{ACC}} = p_{\mathtt{CAA}} = p_{\mathtt{GTT}} = p_{\mathtt{TGG}} \\
p_{\mathtt{AAT}} = p_{\mathtt{CCG}} = p_{\mathtt{GGC}} = p_{\mathtt{TTA}} & p_{\mathtt{ACT}} = p_{\mathtt{CAG}} = p_{\mathtt{GTC}} = p_{\mathtt{TGA}} \\
p_{\mathtt{ACA}} = p_{\mathtt{CAC}} = p_{\mathtt{GTG}} = p_{\mathtt{TGT}} & p_{\mathtt{ACG}} = p_{\mathtt{CAT}} = p_{\mathtt{GTA}} = p_{\mathtt{TGC}} \\
p_{\mathtt{AGT}} = p_{\mathtt{CTG}} = p_{\mathtt{GAC}} = p_{\mathtt{TCA}} & p_{\mathtt{ATA}} = p_{\mathtt{CGC}} = p_{\mathtt{GCG}} = p_{\mathtt{TAT}} \\
p_{\mathtt{ATC}} = p_{\mathtt{CGA}} = p_{\mathtt{GCT}} = p_{\mathtt{TAG}} & p_{\mathtt{ATG}} = p_{\mathtt{CGT}} = p_{\mathtt{GCA}} = p_{\mathtt{TAC}} \\
p_{\mathtt{ATT}} = p_{\mathtt{CGG}} = p_{\mathtt{GCC}} = p_{\mathtt{TAA}} & p_{\mathtt{AGG}} = p_{\mathtt{CTT}} = p_{\mathtt{GAA}} = p_{\mathtt{TCC}}
\end{array}
$$

Denote by $O^{\mathbf{G}} = (O_1, \cdots, O_S)$ the sets of indices of $\mathtt{p}$ of the elements $\mathtt{p}$ assigned to the collection of these equivalence classes. We call $O^{\mathbf{G}}$ the `orbits` under the action of $\mathbf{G}$.

Example: K81

$S = 4^{n-1}$

$$p_{\mathtt{AAA}} = p_{\mathtt{CCC}} = p_{\mathtt{GGG}} = p_{\mathtt{TTT}} \quad p_{\mathtt{AGA}} = p_{\mathtt{CTC}} = p_{\mathtt{GAG}} = p_{\mathtt{TCT}}$$
$$p_{\mathtt{AAC}} = p_{\mathtt{CCA}} = p_{\mathtt{GGT}} = p_{\mathtt{TTG}} \quad p_{\mathtt{AGC}} = p_{\mathtt{CTA}} = p_{\mathtt{GAT}} = p_{\mathtt{TCG}}$$
$$\ldots$$

We select a representant from each orbit O_s, denoted by x^s. A set of such representants is called a `transversal` set,

e.g. $\{\mathtt{AAA}, \mathtt{AAC}, \mathtt{AAG}, \mathtt{AAT}, \mathtt{ACA}, \mathtt{ACC}, \mathtt{ACG}, \mathtt{ACT}, \mathtt{AGA}, \mathtt{AGC}, \mathtt{AGG}, \mathtt{AGT}, \mathtt{ATA}, \mathtt{ATC}, \mathtt{ATG}, \mathtt{ATT}\}$

Recursive procedure for generating orbits

Since $\mathbf{G}^{K81} \subset \mathbf{G}^{K80}$, can we use $\mathcal{L}^{K81}$ to generate $\mathcal{L}^{K80}$?

Note: $\mathbf{G}^{K80} = \langle \mathbf{G}^{K81}, (AG) \rangle$.

Consider the action of (AG) on the elements of the transversal set of $O^{\mathbf{G}^{K81}}$ (**significant reduction in computation**),

e.g $(AG)\mathrm{p}_{\mathtt{ACA}} = \mathrm{p}_{\mathtt{GCG}}$, so the orbits of $\mathbf{G}^{K81}$ containing $\mathrm{p}_{\mathtt{ACA}}$ and $\mathrm{p}_{\mathtt{GCG}}$ are combined into a new orbit of $\mathbf{G}^{K80}$.

Similarly, $\mathbf{G}^{JC69} = \langle \mathbf{G}^{K80}, (AC) \rangle$.

Motivation

We have been looking the spaces $\mathcal{L}^G$ and provided a convenient description in terms of linear equations.

Can we say more? Do the spaces $\mathcal{L}^G$ have a phylogenetic interpretation?
We will see that they are related to the phylogenetic mixture models.

Content

1 Sequence evolution

2 Mixture models

3 Identifiability

4 Summary

Mixture models

Let n be the number of taxa, $[n]$ the set of taxa and
$\tau := \{\text{tree topologies on } [n] \text{ up to isomorphism}\}$.
Let Ψ_T be a parametrization of a model of evolution for any $T \in \tau$:

$$\Psi_T : \prod_{e \in E(T)} \mathrm{Hom}_{\mathbf{G}}(W, W) \to \mathcal{L}, \quad W = \langle B \rangle_{\mathbb{R}}$$

Definition (in the notation of Matsen-Mosztel-Steel '07)

A phylogenetic mixture (on h classes) is any vector p in $\mathcal{L}$ of the form:

$$\mathrm{p} = \sum_{i=1}^{h} \alpha_i \Psi_{T_i}(\zeta_i),$$

where $T_i \in \tau$, ζ_i-edge parameters, $\alpha_i \in \mathbb{C}$ (or $\mathbb{R}$).

Main result

We denote by $\mathcal{D}$ the set of all phylogenetic mixtures, $\mathcal{D} \subseteq \mathcal{L}$.

Lemma

The set of all phylogenetic mixtures $\mathcal{D}$ is a vector subspace of $\mathcal{L}$

Proposition

If $\Psi_T^{\mathbf{G}}$ are equivariant maps for a certain equivariant model corresponding to a group $\mathbf{G} \subseteq \mathfrak{S}_4$, then $\mathcal{D}$ coincides with $\mathcal{L}^G$.

Model selection

Model selection is important: phylogenetic analysis and characterization of the evolutionary process (estimating $\mathcal{M}_e$, hypothesis testing).

- incorrect model affects the results: incorrect MLE, branch lengths, lower accuracy, inconsistency (Sullivant and Swofford, 1997; Cuningham et.al.1998; Kelsey et.al.1999; Yang et.al. 1994; Buckley et.al 2000; Felstenstein 1978; Bruno and Halpren, 1999; Penny et.al.1994; Huelsenbeck and Hillis 1993)
- Likelihood Ratio (LR) Test and Akaike Information Criterion rely on the candidate input tree to calculate the likelihood-approximation by a single 'best' tree.
- Asymptotic distribution of the LR statistics for mixture models is unknown.

Model selection: alternative outlook

Prior to performing phylogenetic inference, one should assess the fitness of a model to the data for *ALL* possible trees (including their mixtures!).

GOAL:

The linear equations of $\mathcal{L}^G$ are the invariants that vanish on the joint distributions for any T. Therefore, they can be used to assess the goodness-of-fit of $\mathcal{M}^{\mathbf{G}}$ to the data.

Empirical data

-$D = (D_1, \cdots, D_N)$- multiple alignment on n taxa;
-$K = 4^n$;
-$(n_1, \cdots, n_K)$- empirical frequencies, $\sum_{k=1}^{K} n_k = N$;
-$(\mathrm{p}_{x_1}, \cdots, \mathrm{p}_{x_K})$- joint probabilities, $\sum_{k=1}^{K} \mathrm{p}_{x_k} = 1$;

Let $\mathcal{M}$ and $\mathcal{N}$ be two equivariant models, such that $\mathcal{M} \subseteq \mathcal{N}$:

-$O_{\mathcal{M}} = \{O_1, \cdots, O_R\}$ and $O_{\mathcal{N}} = \{Q_1, \cdots, Q_S\}$ - orbits
-$(x^s)_{s=1}^{R}$ and $(y^s)_{s=1}^{S}$ - transversal sets;
-$o^s = \mid O_s \mid, s = 1, \cdots, R$ and $q^s = \mid Q_s \mid, s = 1, \cdots, S$ - cardinalities of the orbits;
-$(n^s)_{s=1}^{R}$ and $(m^s)_{s=1}^{S}$ - cardinalities of the elements of O and Q.

Maximum Likelihood Estimate

Under $\mathcal{M}$ the likelihood takes the form:

$$\mathbf{L}(D \mid \text{parameters}) = \prod_{s=1}^{R} (\mathrm{p}_{x^s})^{n^s}$$

$\hat{\mathrm{p}}_{x^s} = \dfrac{n^s}{o^s R}$ is the MLE of p_{x^s} for all $s = 1, \cdots, R$

Therefore if $x_k \in O_s$, $\hat{\mathrm{p}}_{x_k} = \frac{n^s}{o^s R}$.

Since $\mathcal{L}^G$ are linear spaces, $\hat{\mathrm{p}}_{x_k}$ is the global MLE under $\mathcal{M}^{\mathbf{G}}$.
In fact, $\hat{\mathrm{p}} = (\hat{\mathrm{p}}_{x_1}, \cdots, \hat{\mathrm{p}}_{x_K})$ corresponds to the orthogonal projection of p onto $\mathcal{M}^{\mathbf{G}}$.

Model selection

$H_0 : \mathrm{p} \in \mathcal{L}^{\mathbf{G}_1}$ vs. $H_1 : \mathrm{p} \in \mathcal{L}^{\mathbf{G}_2}$

$$\lambda = -2\log(\mathbf{L}_0 - \mathbf{L}_1) = -2\sum_{k=1}^{K} n_k \log e_k \sim \chi^2(S-R)$$

where $e_k = \sum_{s=1}^{R}\sum_{l=1}^{S} \frac{n^s o^l}{m^l q^s}\delta_{sl}(k), \delta_{sl}(k) = 1$ for $\mathrm{p}_{x_k} \in O_s \cap Q_l$ and 0 otherwise

Alternatives: AIC, MCMC using Markov bases, parametric bootstrap

Content

1 Sequence evolution

2 Mixture models

3 Identifiability

4 Summary

Identifiability

Definition

A statistical model is identifiable if the model parameters (tree topology and the continuous parameters) can be uniquely determined from the distribution of the observed random variables.

From no on, we will think of the **generic** identifiability.

Given a phylogenetic mixture, $\mathrm{p} = \sum_{i=1}^{\mathrm{h}} \alpha_i \Psi_{T_i}^{\mathbf{G}}(\zeta_i)$, are we always guaranteed to recover (T_i, ζ_i)? NOT

Identifiablity is essential for the consistency of statistical inference

Proving consistency of the ML estimate is based on the argument of the identifiability of the parameters (Wald, 1949)

Recovering a tree from the leaf colourations it generates under a Markov model, M. Steel, 1994
Identifiability of parameters in MCMC Bayesian inference of phylogeny B. Rannala, 2002
Phylogenetic MCMC algorithms are misleading on mixtures of trees, E. Mossel and E. Vigoda, 2005
Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. B. Kolaczkowski and J.W. Thornton, 2004
Identifiability of two-tree mixtures under group-based models, E. Allman, S. Petrovic, J. Rhodes, and S. Sullivant, 2009

Identifiability

Question: Is an $\mathtt{h}$-mixture identifiable for any $\mathtt{h}$?

Let $\mathcal{D}^{\mathtt{h}}$ denote the space of all $\mathtt{h}-$mixtures.
if $\mathcal{D}^{\mathtt{h}} = \mathcal{L}^G$, then m-mixtures are not identifiable for $m \geq \mathtt{h}$.

[MMS'07] show that

any $\mathtt{p} \in \mathcal{D}$ is a phylogenetic mixture of class $\mathtt{h} \leq d+1$, $d := dim\mathcal{D}$.

In other words, $\mathcal{D}^{d+1} = \mathcal{D} = \mathcal{L}^G$.
Since now we know that $\mathcal{D} = \mathcal{L}^G$, we can compute $d = dim\mathcal{D} = dim\mathcal{L}^G$ for any equivariant model.

Dimension of $\mathcal{L}^G$

Tools from representation theory allows us to calculate the dimension of the space of mixtures $\mathcal{L}^G$ for any number of taxa under the equivariant models.

Proposition

Dimension of the space of all phylogenetic mixtures for $K81, K80, JC69$ for n taxa is:

$$\begin{aligned} dim_{K81} &= 4^{n-1} \\ dim_{K80} &= 2^{2n-3} + 2^{n-2} \\ dim_{JC69} &= \tfrac{1}{3}2^{2n-3} + 2^{n-2} + \tfrac{1}{3} \end{aligned}$$

Upper bound on the number of mixtures

Proposition

Let $(\mathbf{G}, W)$ *be an equivariant model, n be a number of taxa and let* d *be the dimension of the algebraic varieties associated to trivalent trees on n taxa under this model,* $d = Im\Psi^{\mathbf{G}}_{T}$.

For $\mathtt{h} \geq \frac{dim\mathcal{L}^{G}+1}{d+1}$, $\mathtt{h}$*-tree mixtures are not identifiable.*

Can this bound be lowered?

Upper bound on the number of mixtures: example

Example

K81

$n = 4$: $dim\mathcal{L}^G = 4^{n-1} = 64$, $d = 3(2n-3)+1 = 16$,

$4-$*tree mixtures are non-identifiable*

$n = 5$: $dim\mathcal{L}^G = 256$, $d = 22$,

$12-$*tree mixtures are non-identifiable.*

JC69

$n = 4$: $dim\mathcal{L}^G = \frac{1}{3}2^{2n-3} + 2^{n-2} + \frac{1}{3} = 15$, $d = 2n-2 = 6$,

$3-$*tree mixtures are non-identifiable*

$n = 5$: $dim\mathcal{L}^G = 51$, $d = 8$,

$7-$*tree mixtures are non-identifiable*

Content

1 Sequence evolution

2 Mixture models

3 Identifiability

4 Summary

Ongoing and future work

Development of a web-based tool for model selection on the real-life data.

We showed that there exists a bound on $\mathtt{h}$, the number of mixtures, for which model parameters are non-identifiable.

Are the $\mathtt{h}-1$ mixtures identifiable?

Fix n and h. Let $\mathcal{M}^0 \subseteq \mathcal{M}^1$.
Assume the topology (continuous parameters) to be non-identifiable under $\mathcal{M}^0$ for $\mathtt{h}$ mixtures. Is it non-identifiable under $\mathcal{M}^1$?
Fix $\mathcal{M}^{\mathbf{G}}$ and h. What is the smallest n such that for $k > n$ $\mathcal{M}^{\mathbf{G}}$ is identifiable?

Acknowledgements:
Sonja Petrović
Mathias Drton

Thank you