Linear invariants and the space of phylogenetic mixtures for Felsenstein'81 and other models

Marta Casanellas Rius
(mainly joint work with M. Steel)

Departament de Matemàtiques
Universitat Politècnica de Catalunya

MCEB
Hammeau de l'Etoile
June 16th, 2016

Phylogenetic invariants

- Introduced by Lake, Felsenstein and Cavender '87.
- First attempts to use invariants: Steel et al. '93, Ferretti-Sankoff '95.
- Main problems: not all invariants known, not clear how to use them, evaluation of fit of model and tree topology at the same time.
- Felsenstein's book: "invariants are worth attention, not for what they do for us now, but what they might lead to in the future".
- Invariants can deal with very general models as GMM (which does NOT assume a stationary distribution, NOR time-reversibility, NOR global homogeneity, NOT even local homogeneity along each edge), mixtures (or fully partitioned model) or the coalescent.
- Nowadays:
  - theoretical issues solved for the basic models.
  - widely used in identifiability problems (Allman-Rhodes '06,'07,'12...)
  - successful reconstruction methods for quartets based on invariants: Erik+2 (Sys. Bio. 2016, Fernández-Sánchez' poster)

Phylogenetic invariants

- Introduced by Lake, Felsenstein and Cavender '87.
- First attempts to use invariants: Steel et al. '93, Ferretti-Sankoff '95.
- Main problems: not all invariants known, not clear how to use them, evaluation of fit of model and tree topology at the same time.
- Felsenstein's book: "invariants are worth attention, not for what they do for us now, but what they might lead to in the future".
- Invariants can deal with very general models as GMM (which does NOT assume a stationary distribution, NOR time-reversibility, NOR global homogeneity, NOT even local homogeneity along each edge), mixtures (or fully partitioned model) or the coalescent.
- Nowadays:
  - theoretical issues solved for the basic models.
  - widely used in identifiability problems (Allman-Rhodes '06,'07,'12...)
  - successful reconstruction methods for quartets based on invariants: Erik+2 (Sys. Bio. 2016, Fernández-Sánchez' poster)

Phylogenetic invariants

- Introduced by Lake, Felsenstein and Cavender '87.
- First attempts to use invariants: Steel et al. '93, Ferretti-Sankoff '95.
- Main problems: not all invariants known, not clear how to use them, evaluation of fit of model and tree topology at the same time.
- Felsenstein's book: "invariants are worth attention, not for what they do for us now, but what they might lead to in the future".
- Invariants can deal with very general models as GMM (which does NOT assume a stationary distribution, NOR time-reversibility, NOR global homogeneity, NOT even local homogeneity along each edge), mixtures (or fully partitioned model) or the coalescent.
- Nowadays:
  - theoretical issues solved for the basic models.
  - widely used in identifiability problems (Allman-Rhodes '06,'07,'12...)
  - successful reconstruction methods for quartets based on invariants: Erik+2 (Sys. Bio. 2016, Fernández-Sánchez' poster)

Phylogenetic invariants

- Introduced by Lake, Felsenstein and Cavender '87.
- First attempts to use invariants: Steel et al. '93, Ferretti-Sankoff '95.
- Main problems: not all invariants known, not clear how to use them, evaluation of fit of model and tree topology at the same time.
- Felsenstein's book: "invariants are worth attention, not for what they do for us now, but what they might lead to in the future".
- Invariants can deal with very general models as GMM (which does NOT assume a stationary distribution, NOR time-reversibility, NOR global homogeneity, NOT even local homogeneity along each edge), mixtures (or fully partitioned model) or the coalescent.
- Nowadays:
  - theoretical issues solved for the basic models.
  - widely used in identifiability problems (Allman-Rhodes '06,'07,'12...)
  - successful reconstruction methods for quartets based on invariants: Erik+2 (Sys. Bio. 2016, Fernández-Sánchez' poster)

Phylogenetic invariants

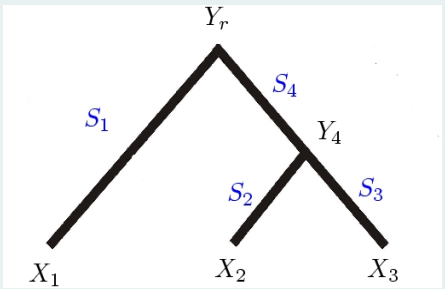
- Introduced by Lake, Felsenstein and Cavender '87.
- First attempts to use invariants: Steel et al. '93, Ferretti-Sankoff '95.
- Main problems: not all invariants known, not clear how to use them, evaluation of fit of model and tree topology at the same time.
- Felsenstein's book: "invariants are worth attention, not for what they do for us now, but what they might lead to in the future".
- Invariants can deal with very general models as GMM (which does NOT assume a stationary distribution, NOR time-reversibility, NOR global homogeneity, NOT even local homogeneity along each edge), mixtures (or fully partitioned model) or the coalescent.
- Nowadays:
  - theoretical issues solved for the basic models.
  - widely used in identifiability problems (Allman-Rhodes '06,'07,'12...)
  - successful reconstruction methods for quartets based on invariants: Erik+2 (Sys. Bio. 2016, Fernández-Sánchez' poster)

Phylogenetic invariants

- Introduced by Lake, Felsenstein and Cavender '87.
- First attempts to use invariants: Steel et al. '93, Ferretti-Sankoff '95.
- Main problems: not all invariants known, not clear how to use them, evaluation of fit of model and tree topology at the same time.
- Felsenstein's book: "invariants are worth attention, not for what they do for us now, but what they might lead to in the future".
- Invariants can deal with very general models as GMM (which does NOT assume a stationary distribution, NOR time-reversibility, NOR global homogeneity, NOT even local homogeneity along each edge), mixtures (or fully partitioned model) or the coalescent.
- Nowadays:
  - theoretical issues solved for the basic models.
  - widely used in identifiability problems (Allman-Rhodes '06,'07,'12...)
  - successful reconstruction methods for quartets based on invariants: Erik+2 (Sys. Bio. 2016, Fernández-Sánchez' poster)

Hidden Markov process

$p_{x_1x_2x_3} = Prob(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ = joint probability of the observed variables X_1, X_2, X_3. Then,

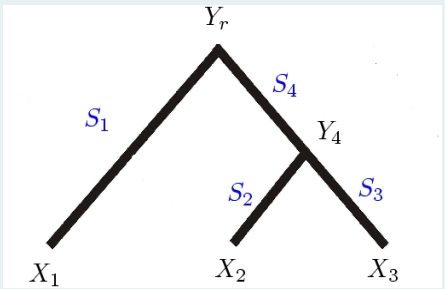


$$p_{x_1x_2x_3} = \sum_{y_4,y_r \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_{y_r} S_1(y_r, x_1) S_4(y_r, y_4) S_2(y_4, x_2) S_3(y_4, x_3)$$

- Parameters: entries of S_i and π.
- Constraints on S_i and π specify the evolutionary model: (non-homogeneous) JC69, K80, K81, F81, SSM, GMM, ...
- S_i does need to be of type $exp(t_i Q_i)$ (and if it is, the rate matrices Q_i can be different for different branches).
- NO global instantaneous mutation rate matrix assumed $\Rightarrow$ "non-homogeneous across lineages".

Hidden Markov process

$p_{x_1x_2x_3} = Prob(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ = joint probability of the observed variables X_1, X_2, X_3. Then,

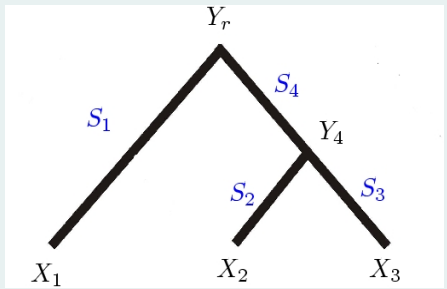


$$p_{x_1x_2x_3} = \sum_{y_4,y_r \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_{y_r} S_1(y_r, x_1) S_4(y_r, y_4) S_2(y_4, x_2) S_3(y_4, x_3)$$

- Parameters: entries of S_i and π.
- Constraints on S_i and π specify the evolutionary model: (non-homogeneous) JC69, K80, K81, F81, SSM, GMM, ...
- S_i does need to be of type $exp(t_i Q_i)$ (and if it is, the rate matrices Q_i can be different for different branches).
- NO global instantaneous mutation rate matrix assumed $\Rightarrow$ "non-homogeneous across lineages".

Hidden Markov process

$p_{x_1x_2x_3} = Prob(X_1 = x_1, X_2 = x_2, X_3 = x_3) =$ joint probability of the observed variables X_1, X_2, X_3. Then,

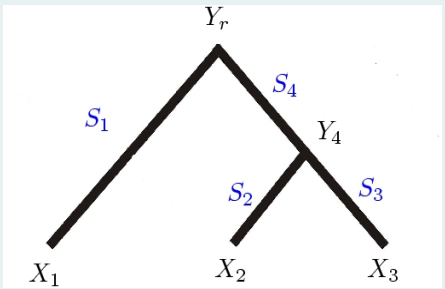


$$p_{x_1x_2x_3} = \sum_{y_4,y_r \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_{y_r} S_1(y_r, x_1) S_4(y_r, y_4) S_2(y_4, x_2) S_3(y_4, x_3)$$

- Parameters: entries of S_i and π.
- Constraints on S_i and π specify the evolutionary model: (non-homogeneous) JC69, K80, K81, F81, SSM, GMM, ...
- S_i does need to be of type $exp(t_i Q_i)$ (and if it is, the rate matrices Q_i can be different for different branches).
- NO global instantaneous mutation rate matrix assumed $\Rightarrow$ "non-homogeneous across lineages".

Hidden Markov process

$p_{x_1x_2x_3} = Prob(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ = joint probability of the observed variables X_1, X_2, X_3. Then,

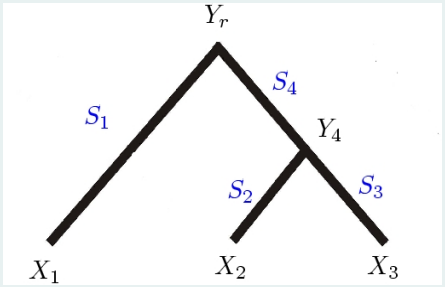


$$p_{x_1x_2x_3} = \sum_{y_4, y_r \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_{y_r} S_1(y_r, x_1) S_4(y_r, y_4) S_2(y_4, x_2) S_3(y_4, x_3)$$

- Parameters: entries of S_i and π.
- Constraints on S_i and π specify the evolutionary model: (non-homogeneous) JC69, K80, K81, F81, SSM, GMM, ...
- S_i does need to be of type $exp(t_i Q_i)$ (and if it is, the rate matrices Q_i can be different for different branches).
- NO global instantaneous mutation rate matrix assumed $\Rightarrow$ "non-homogeneous across lineages".

Hidden Markov process

$p_{x_1x_2x_3} = Prob(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ = joint probability of the observed variables X_1, X_2, X_3. Then,

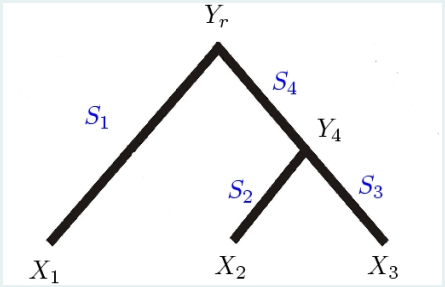


$$p_{x_1x_2x_3} = \sum_{y_4,y_r \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_{y_r} S_1(y_r, x_1) S_4(y_r, y_4) S_2(y_4, x_2) S_3(y_4, x_3)$$

- Joint distribution at the leaves $p_{x_1 \ldots x_n}$ can be expressed as a polynomial in terms of the parameters of the model.
- Are there polynomial relations among these probabilities that are satisfied no matter what the parameters are? Why should we care about them?

Hidden Markov process

$p_{x_1x_2x_3} = Prob(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ = joint probability of the observed variables X_1, X_2, X_3. Then,



$$p_{x_1x_2x_3} = \sum_{y_4,y_r \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_{y_r} S_1(y_r, x_1) S_4(y_r, y_4) S_2(y_4, x_2) S_3(y_4, x_3)$$

- Joint distribution at the leaves $p_{x_1 \ldots x_n}$ can be expressed as a polynomial in terms of the parameters of the model.
- Are there polynomial relations among these probabilities that are satisfied no matter what the parameters are? Why should we care about them?

Example: Jukes-Cantor

- Some relations depend on the model chosen (not on the *tree topology*): $\sum p_{x_1 x_2 \ldots x_n} = 1$ (trivial) and

$$p_{AA\ldots A} = p_{CC\ldots C} = p_{GG\ldots G} = p_{TT\ldots T}$$

$$p_{A\ldots AC} = p_{A\ldots AG} = p_{A\ldots AT} = \cdots = p_{T\ldots TG}$$

$$\ldots$$

- These are called model invariants (relations satisfied on all tree topologies)
- But some relations depend on the tree topology.

Definition

- Phylogenetic Invariants of a tree T_0: polynomial relations satisfied by any joint distribution that has evolved under an evolutionary model $\mathcal{M}$ on T_0.
- *Topology* invariants: invariants of T_0 that are not satisfied by some joint distributions on some other tree topologies (hence, they could be used to distinguish between different topologies).

Example: Jukes-Cantor

- Some relations depend on the model chosen (not on the *tree topology*): $\sum p_{x_1 x_2 \ldots x_n} = 1$ (trivial) and

$$p_{AA\ldots A} = p_{CC\ldots C} = p_{GG\ldots G} = p_{TT\ldots T}$$
$$p_{A\ldots AC} = p_{A\ldots AG} = p_{A\ldots AT} = \cdots = p_{T\ldots TG}$$
$$\ldots$$

- These are called model invariants (relations satisfied on all tree topologies)
- But some relations depend on the tree topology.

Definition

- Phylogenetic Invariants of a tree T_0: polynomial relations satisfied by any joint distribution that has evolved under an evolutionary model $\mathcal{M}$ on T_0.
- *Topology* invariants: invariants of T_0 that are not satisfied by some joint distributions on some other tree topologies (hence, they could be used to distinguish between different topologies).

Example: Jukes-Cantor

- Some relations depend on the model chosen (not on the *tree topology*): $\sum p_{x_1 x_2 \ldots x_n} = 1$ (trivial) and

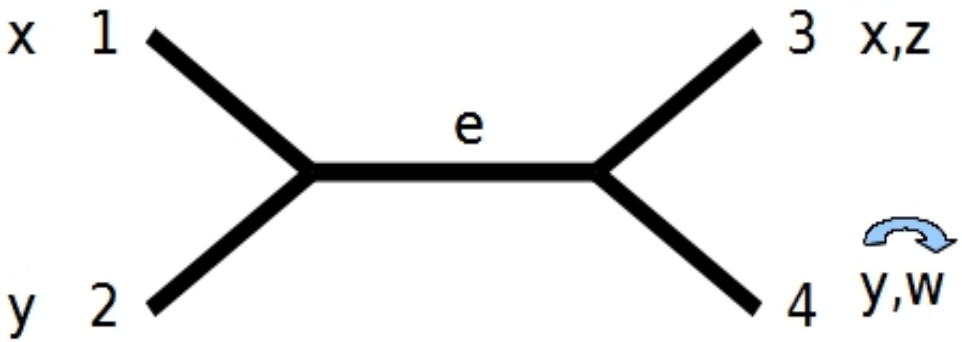
$$p_{AA\ldots A} = p_{CC\ldots C} = p_{GG\ldots G} = p_{TT\ldots T}$$
$$p_{A\ldots AC} = p_{A\ldots AG} = p_{A\ldots AT} = \cdots = p_{T\ldots TG}$$
$$\ldots$$

- These are called model invariants (relations satisfied on all tree topologies)
- But some relations depend on the tree topology.

Definition

- Phylogenetic Invariants of a tree T_0: polynomial relations satisfied by any joint distribution that has evolved under an evolutionary model $\mathcal{M}$ on T_0.
- *Topology* invariants: invariants of T_0 that are not satisfied by some joint distributions on some other tree topologies (hence, they could be used to distinguish between different topologies).

Example: Lake's linear invariants



For the JC60 model on the tree 12|34 the following are linear topology invariants known as Lake's invariants:

$$H_1: \quad p_{xyxy} + p_{xyzw} = p_{xyzy} + p_{xyxw}$$

$$H_2: \quad p_{xyyx} + p_{xywz} = p_{xyyz} + p_{xywx}$$

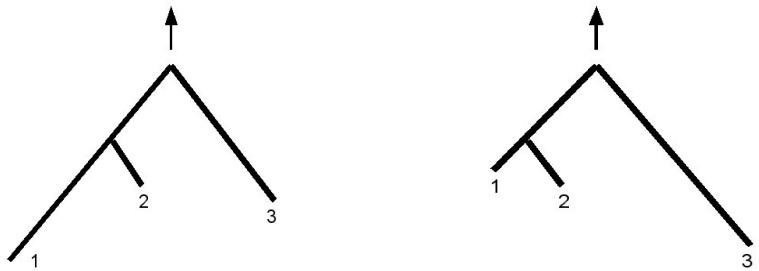
for any x, y, z, w in $\{A, C, G, T\}$.

- Any other linear topology invariant is a linear combination of these two.
- Degree 1 (linear) ... interesting enough?

Linear invariants can deal with mixtures

- Mixture of Markov processes on a tree T: sites undergo the same model on T (but not with the same instantaneous mutation rates); m partitions in the alignment;
  NO constraint between the substitution matrices of different partitions.

AACTTCGAGGCTTACCGCTG...
AAGGTCGATGCTCACCGATG...
AACGTCTATGCTCACCGATG...

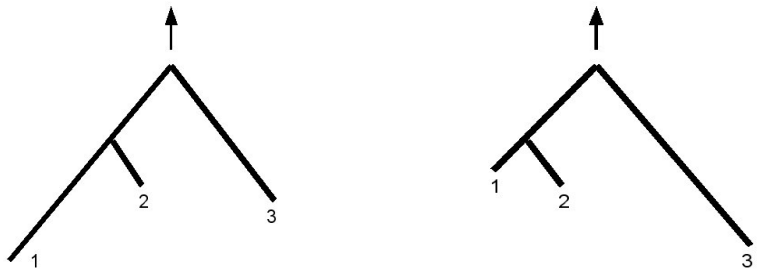


- In general, it becomes impossible to decide which tree topology generated the data (even for ∞ data): the tree topology is not identifiable any more. But for some models it is possible.

Linear invariants can deal with mixtures

- Mixture of Markov processes on a tree T: sites undergo the same model on T (but not with the same instantaneous mutation rates); m partitions in the alignment;
  NO constraint between the substitution matrices of different partitions.

AACTTCGAGGCTTACCGCTG...
AAGGTCGATGCTCACCGATG...
AACGTCTATGCTCACCGATG...



- In general, it becomes impossible to decide which tree topology generated the data (even for ∞ data): the tree topology is not identifiable any more. But for some models it is possible.

The space of mixtures for a given model $\mathcal{M}$

Definition

Fix an evolutionary model $\mathcal{M}$. $\mathbb{P}_{T,\Theta}$ = distribution at the tips of of T, T evolving under $\mathcal{M}$ with parameters Θ .
The *space of mixtures on* T is the affine linear variety

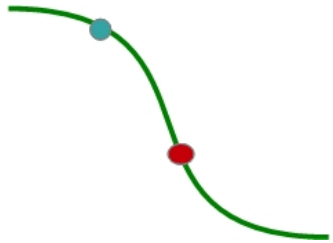
$$\mathcal{D}_T = \left\{ p = \sum_i \lambda_i \mathbb{P}_{T,\Theta_i} \,\middle|\, \sum_i \lambda_i = 1 \right\}.$$

If $\mathcal{T}$ = phylogenetic trees on the set X of taxa, *the space of phylogenetic mixtures* on X is the affine linear variety

$$\mathcal{D} = \left\{ p = \sum_i \lambda_i \mathbb{P}_{T_i,\Theta_i} \,\middle|\, \sum_i \lambda_i = 1\,,\ T_i \in \mathcal{T} \right\}$$

Linear invariants and mixtures

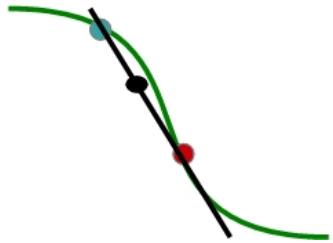
If linear topology invariants exist for the model considered, then the tree topology that generated the mixture is identifiable:



- 2-states, Neyman symmetric model: there are no linear topology invariants (Matsen-Mossel-Steel), so tree topology *cannot* be identified for mixtures on *an arbitrary* number of categories.
- (nonhomogeneous) JC69, K80 have linear topology invariants but K81, SSM and GMM no (tree topology can only be identified for mixtures on a certain number of categories).

Linear invariants and mixtures

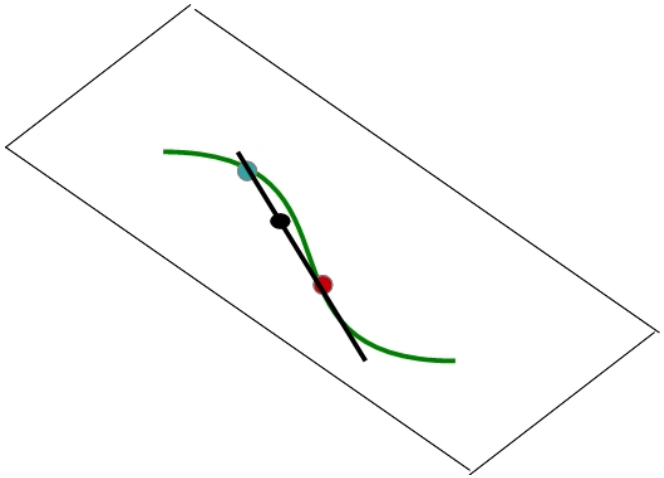
If linear topology invariants exist for the model considered, then the tree topology that generated the mixture is identifiable:



- 2-states, Neyman symmetric model: there are no linear topology invariants (Matsen-Mossel-Steel), so tree topology *cannot* be identified for mixtures on *an arbitrary* number of categories.
- (nonhomogeneous) JC69, K80 have linear topology invariants but K81, SSM and GMM no (tree topology can only be identified for mixtures on a certain number of categories).

Linear invariants and mixtures

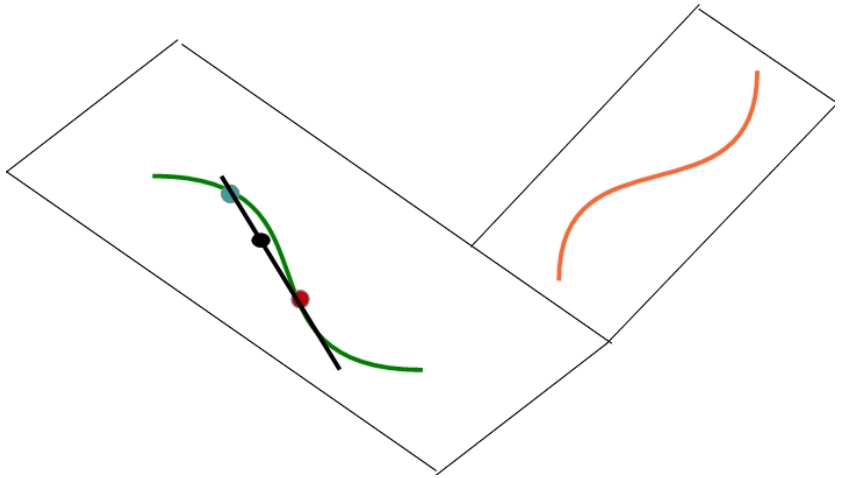
If linear topology invariants exist for the model considered, then the tree topology that generated the mixture is identifiable:



- 2-states, Neyman symmetric model: there are no linear topology invariants (Matsen-Mossel-Steel), so tree topology *cannot* be identified for mixtures on *an arbitrary* number of categories.
- (nonhomogeneous) JC69, K80 have linear topology invariants but K81, SSM and GMM no (tree topology can only be identified for mixtures on a certain number of categories).

Linear invariants and mixtures

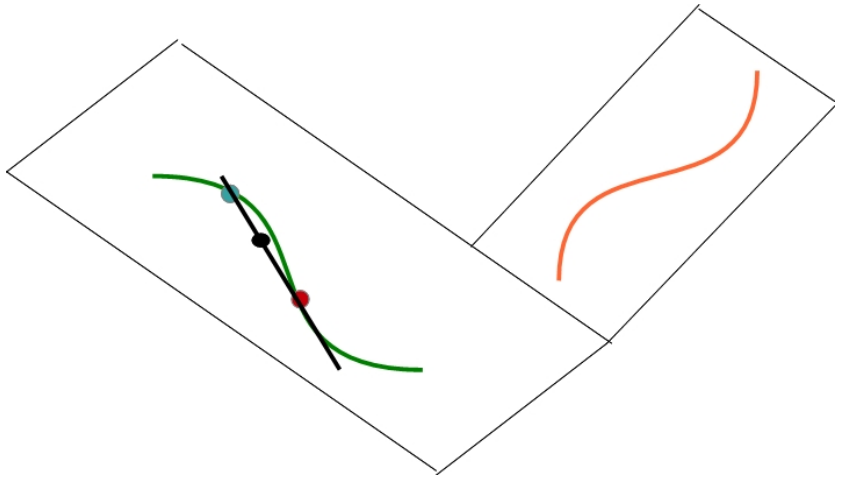
If linear topology invariants exist for the model considered, then the tree topology that generated the mixture is identifiable:



- 2-states, Neyman symmetric model: there are no linear topology invariants (Matsen-Mossel-Steel), so tree topology *cannot* be identified for mixtures on *an arbitrary* number of categories.
- (nonhomogeneous) JC69, K80 have linear topology invariants but K81, SSM and GMM no (tree topology can only be identified for mixtures on a certain number of categories).

Linear invariants and mixtures

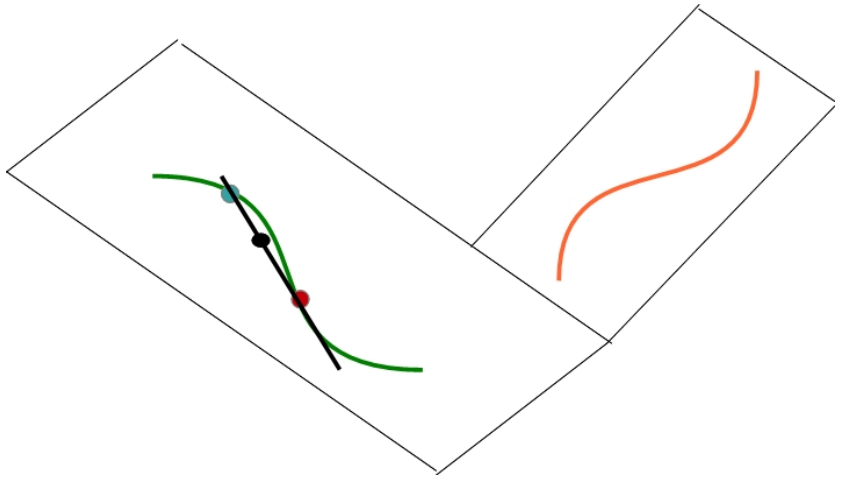
If linear topology invariants exist for the model considered, then the tree topology that generated the mixture is identifiable:



- 2-states, Neyman symmetric model: there are no linear topology invariants (Matsen-Mossel-Steel), so tree topology *cannot* be identified for mixtures on *an arbitrary* number of categories.
- (nonhomogeneous) JC69, K80 have linear topology invariants but K81, SSM and GMM no (tree topology can only be identified for mixtures on a certain number of categories).

Linear invariants and mixtures

If linear topology invariants exist for the model considered, then the tree topology that generated the mixture is identifiable:



- 2-states, Neyman symmetric model: there are no linear topology invariants (Matsen-Mossel-Steel), so tree topology *cannot* be identified for mixtures on *an arbitrary* number of categories.
- (nonhomogeneous) JC69, K80 have linear topology invariants but K81, SSM and GMM no (tree topology can only be identified for mixtures on a certain number of categories).

Space of Mixtures and linear phylogenetic invariants

Relationship between mixtures and linear invariants:

$$\mathcal{D}_T = \underbrace{\left\langle \vec{p} \,\middle|\, p = \mathbb{P}_{T,\Theta} \right\rangle}_{E_T} \cap \{\text{trivial invariant}\}$$

$$\mathcal{D} = \underbrace{\left\langle \vec{p} \,\middle|\, p = \mathbb{P}_{T,\Theta}\ T \in \mathcal{T} \right\rangle}_{E} \cap \{\text{trivial invariant}\}$$

- Linear model invariants: L orthogonal (dual) space to E.
- Linear invariants for a tree T: L_T orthogonal (dual) space to E_T.
- Linear topology invariants: quotient space L_T/L

So far, only linear invariants for models with uniform stationary distribution and 4 states had been studied.

Equal Input model on k-states (EI-model)

- $\pi =$ a stationary distribution on k states; for each edge e, consider a parameter $\theta_e \in [0,1]$ and let conditional probabilities satisfy:

$$Prob(\beta|\alpha, e) = \pi_\beta \cdot \theta_e,$$

  for any states α, β.

- For $k = 4$, this is Felsenstein'81 model:

$$S^e = \begin{pmatrix} 1-(1-\pi_A)\theta_e & \pi_C\theta_e & \pi_G\theta_e & \pi_T\theta_e \\ \pi_A\theta_e & 1-(1-\pi_C)\theta_e & \pi_G\theta_e & \pi_T\theta_e \\ \pi_A\theta_e & \pi_C\theta_e & 1-(1-\pi_G)\theta_e & \pi_T\theta_e \\ \pi_A\theta_e & \pi_C\theta_e & \pi_G\theta_e & 1-(1-\pi_T)\theta_e \end{pmatrix}$$

- Fully symmetric model: π uniform distribution. $k = 4 \rightarrow$ Jukes-Cantor model.
- Coincides with the random cluster model on k states.
- For $k = \infty$: Kimura's infinite alleles model.

$k \geq 4$, Generalizing Lake's invariants

- π fixed (inferred from data?)
- Felsensten'81

$$H_1: \quad \frac{p_{xyxy}}{\pi(x)\pi(y)} + \frac{p_{xyzw}}{\pi(z)\pi(w)} = \frac{p_{xyzy}}{\pi(z)\pi(y)} + \frac{p_{xyxw}}{\pi(x)\pi(w)}$$

$$H_2: \quad \frac{p_{xyyx}}{\pi(x)\pi(y)} + \frac{p_{xywz}}{\pi(z)\pi(w)} = \frac{p_{xyyz}}{\pi(z)\pi(y)} + \frac{p_{xywx}}{\pi(x)\pi(y)}$$

- Generalized to k states $EI-$models, $k \geq 3$ (and more general models).

Model invariants for $EI-$models, π fixed

Set of taxa $[n] = \{1, 2, \ldots, n\}$ $k \geq 2$.

Theorem (C-Steel)

We provide a set of linearly independent points that span the space of mixtures $\mathcal{D}^\pi$ for almost any π . The dimension of $\mathcal{D}^\pi$ equals

$$\sharp\{partitions\ of\ [n]\ of\ size\ \leq k\} - 1$$

(if $k \geq n$ it equals $B_n - 1$ where B_n: Bell number.)

Consequence: easy way to obtain a set of generators for the space of linear model invariants.

- $k = 4$, $dim\mathcal{D}^\pi = \frac{2^{2n-3}+1}{3} + 2^{n-2} - 1$.
- Results are also valid for $k = \infty$, Kimura's infinite alleles model.

Topology invariants for $EI-$models, π fixed

Theorem (C-Steel)

We provide a set of linearly independent points that span the space of mixtures $\mathcal{D}_T^\pi$, for any π fixed, any k, any $n \geq 3$.
The dimension of $\mathcal{D}_T^\pi$ equals $|co(T)| - 1$, where $co(T)$ is the set of partitions of $\{1, \ldots, n\}$ compatible with T.

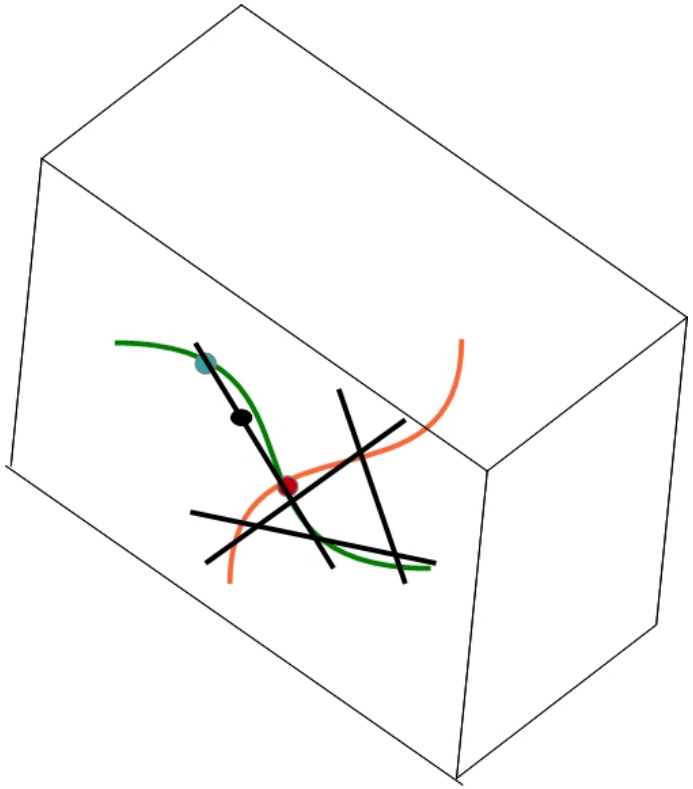
Corollary

Algorithm to obtain a basis of the set of **linear topology invariants** *for the EI model on $n \geq 4$ leaves, any k; dimension $= |\Sigma_k| - |co(T)| =$ number of partitions of $\{1, \ldots, n\}$ of size $\leq k$ incompatible with T.*

Corollary

For $k \geq n$ or for $n = 4$, Lake-type invariants generate all linear topology invariants. For $k < n$, NO.

What if there are NO linear topology invariants?



But still, different models give rise to different linear spaces $\Rightarrow$ use linear **model invariants**.

Why should one care? Model selection:

- Linear model invariants are satisfied by any number of mixtures on (possibly) different topologies.
- They allow to distinguish between different models and can be used for model selection.
- C-Fernández-Sánchez-Kedzierska'12: algorithm that provides all linear model invariants for any number of taxa for JC69, K80, K81, SSM.
- This has been used in `SPIn` (Kedzierska-Drton-Guigó-C, MBE 2012): it tells you wether your data is likely to come from a mixture (possibly on a collection of different tree topologies) of (nonhomogeneous) JC69, K80, K81 or SSM processes.
- Not many models ... but good results compared to `jmodeltest` because we consider nonhomogeneous models across lineages (different instantaneous mutation rates at different lineages) and allow mixtures!

Why should one care? Model selection:

- Linear model invariants are satisfied by any number of mixtures on (possibly) different topologies.
- They allow to distinguish between different models and can be used for model selection.
- C-Fernández-Sánchez-Kedzierska'12: algorithm that provides all linear model invariants for any number of taxa for JC69, K80, K81, SSM.
- This has been used in `SPIn` (Kedzierska-Drton-Guigó-C, MBE 2012): it tells you wether your data is likely to come from a mixture (possibly on a collection of different tree topologies) of (nonhomogeneous) JC69, K80, K81 or SSM processes.
- Not many models ... but good results compared to `jmodeltest` because we consider nonhomogeneous models across lineages (different instantaneous mutation rates at different lineages) and allow mixtures!

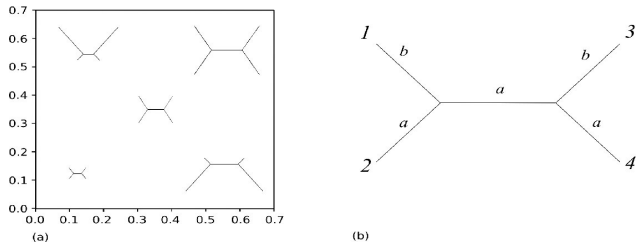
Why should one care? Model selection:

- Linear model invariants are satisfied by any number of mixtures on (possibly) different topologies.
- They allow to distinguish between different models and can be used for model selection.
- C-Fernández-Sánchez-Kedzierska'12: algorithm that provides all linear model invariants for any number of taxa for JC69, K80, K81, SSM.
- This has been used in `SPIn` (Kedzierska-Drton-Guigó-C, MBE 2012): it tells you wether your data is likely to come from a mixture (possibly on a collection of different tree topologies) of (nonhomogeneous) JC69, K80, K81 or SSM processes.
- Not many models ... but good results compared to `jmodeltest` because we consider nonhomogeneous models across lineages (different instantaneous mutation rates at different lineages) and allow mixtures!

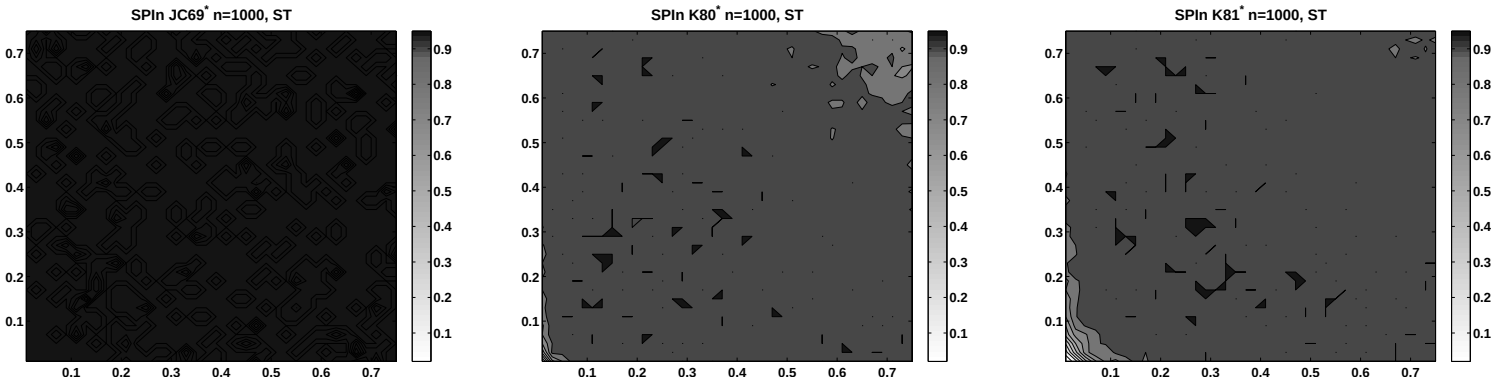
Why should one care? Model selection:

- Linear model invariants are satisfied by any number of mixtures on (possibly) different topologies.
- They allow to distinguish between different models and can be used for model selection.
- C-Fernández-Sánchez-Kedzierska'12: algorithm that provides all linear model invariants for any number of taxa for JC69, K80, K81, SSM.
- This has been used in `SPIn` (Kedzierska-Drton-Guigó-C, MBE 2012): it tells you wether your data is likely to come from a mixture (possibly on a collection of different tree topologies) of (nonhomogeneous) JC69, K80, K81 or SSM processes.
- Not many models ... but good results compared to `jmodeltest` because we consider nonhomogeneous models across lineages (different instantaneous mutation rates at different lineages) and allow mixtures!

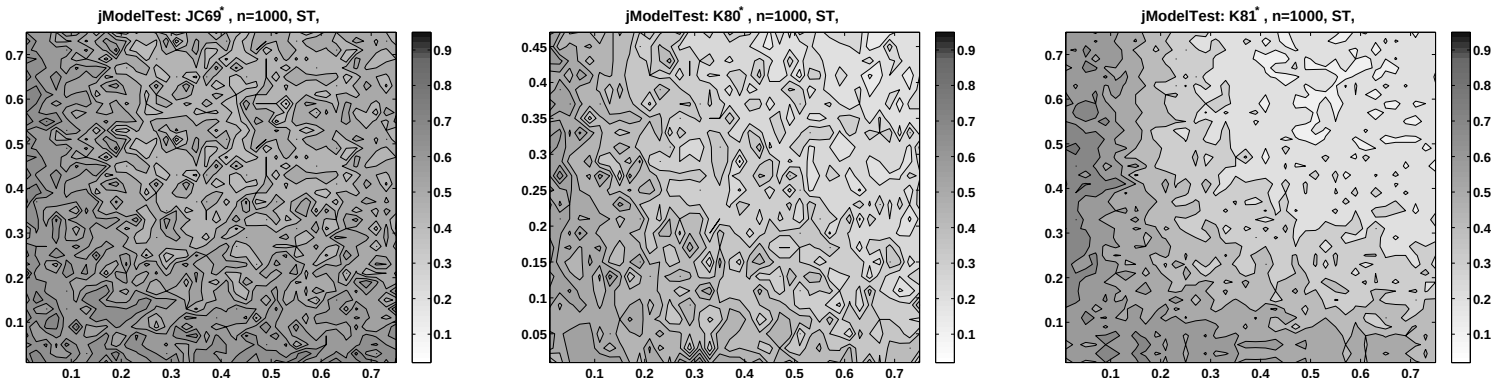
Results on simulated (unmixed) data



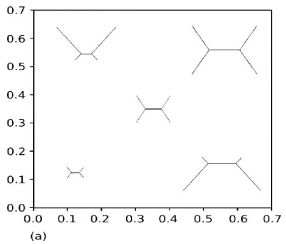
SPIn

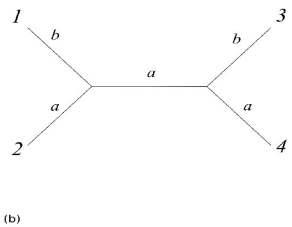


jModelTest

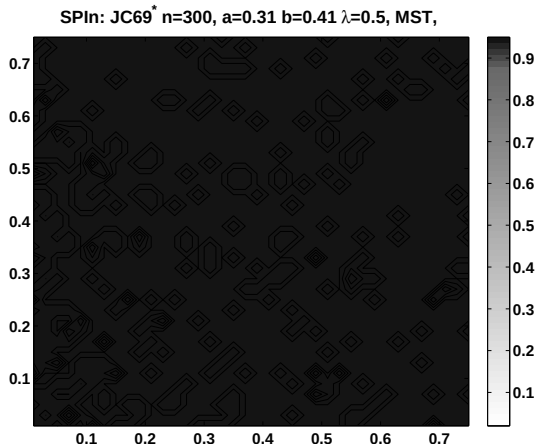


Mixture data

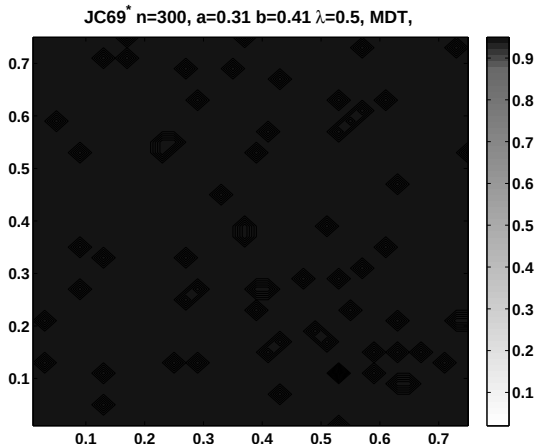




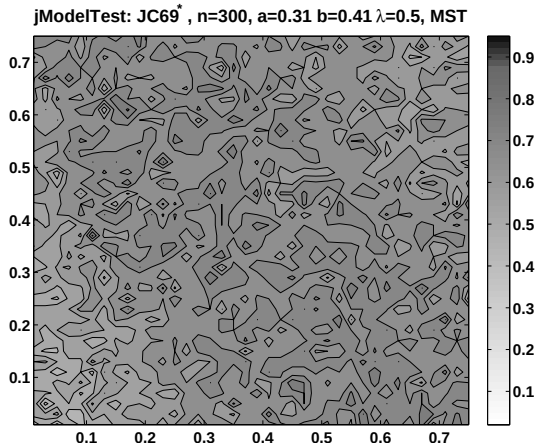
SPIn



SPIn

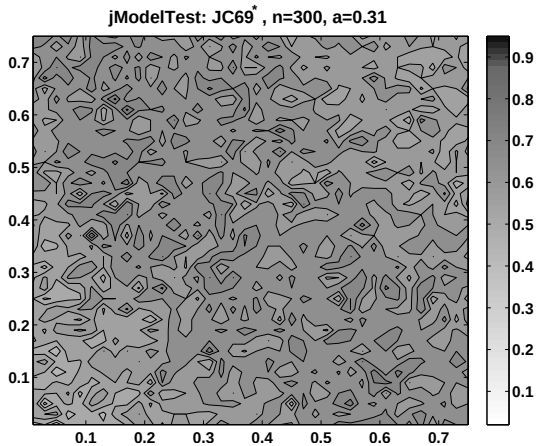


jModelTest



2 categories, 1 tree

jModelTest



2 categories, 2 trees

Thank you for your attention!

Advertisement:

ALGEBRAIC AND COMBINATORIAL PHYLOGENETICS

whole MONTH research program

Barcelona June 12th - July 10th (approx.) 2017.

A WORKSHOP + COURSES by:

Mike Steel

Arndt von Haeseler

Piotr Zwiernik

Some funding available for participants, check the "webpage" (coming soon...)