The embedding problem for Markov processes

Marta Casanellas, Jesús Fernández-Sánchez, **Jordi Roca-Lacostena**

Seminari de geometria algebraica de Barcelona
05/03/2021

The embedding problem

Definition

- A *Markov matrix* is a non-negative square matrix with row sum equal to one.
- A *rate matrix* is a real square matrix with row sum equal to zero and non-negative off-diagonal entries.

A Markov matrix M is said to be embeddable if $M = \exp(Q)$ for some rate matrix Q. In this case, we say that Q is a *Markov generator* for M.

Embedding Problem (Elfving 1937)

Given a Markov matrix M, decide whether it is embeddable or not.

The embedding problem

2×2 M embeddable $\Leftrightarrow \det(M) > 0$ (Kingman 1962).

3×3 Characterization split in cases depending on eigenvalues.
(Cuthbert 1973, Johansen 1974, Chen 2011).

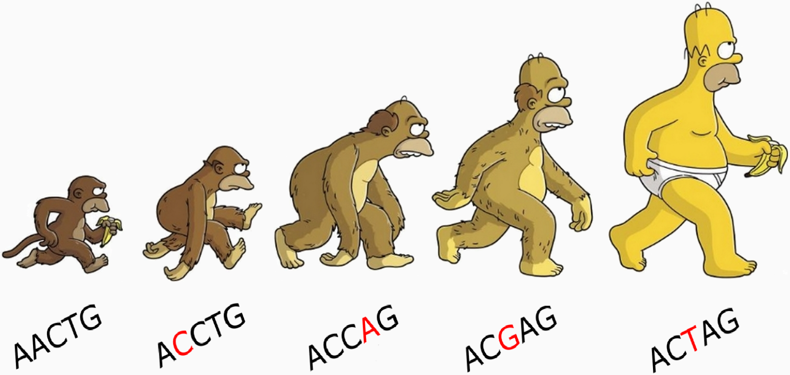
$n \times n$ Solved for some particular cases:

- Different and real eigenvalues (Singer 1976).
- Double-stochastic matrices (Jia 2016).
- Equal-input model (Baake-Sumner 2019).

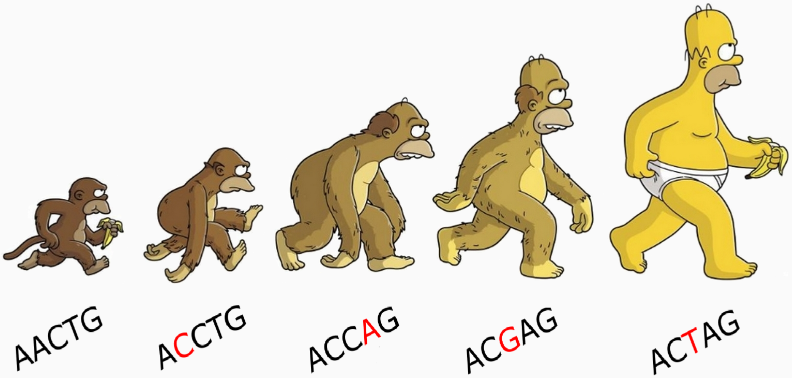
In this talk:

- Different eigenvalues (real or not).
- 4×4 Markov matrices.

Modelling evolution



Modelling evolution



- Nucleotides are represented as the states of random variables.
- Nucleotide substitution is modelled as a *Markov process*.

Markov Processes

Markov matrices encode the conditional substitution probabilities between states:

$$M = \begin{pmatrix} P(\mathtt{A} \to \mathtt{A}) & P(\mathtt{A} \to \mathtt{G}) & P(\mathtt{A} \to \mathtt{C}) & P(\mathtt{A} \to \mathtt{T}) \\ P(\mathtt{G} \to \mathtt{A}) & P(\mathtt{G} \to \mathtt{G}) & P(\mathtt{G} \to \mathtt{C}) & P(\mathtt{G} \to \mathtt{T}) \\ P(\mathtt{C} \to \mathtt{A}) & P(\mathtt{C} \to \mathtt{G}) & P(\mathtt{C} \to \mathtt{C}) & P(\mathtt{C} \to \mathtt{T}) \\ P(\mathtt{T} \to \mathtt{A}) & P(\mathtt{T} \to \mathtt{G}) & P(\mathtt{T} \to \mathtt{C}) & P(\mathtt{T} \to \mathtt{T}) \end{pmatrix}$$

Markov Processes

Markov matrices encode the conditional substitution probabilities between states:

$$M = \begin{pmatrix} P(\mathtt{A} \to \mathtt{A}) & P(\mathtt{A} \to \mathtt{G}) & P(\mathtt{A} \to \mathtt{C}) & P(\mathtt{A} \to \mathtt{T}) \\ P(\mathtt{G} \to \mathtt{A}) & P(\mathtt{G} \to \mathtt{G}) & P(\mathtt{G} \to \mathtt{C}) & P(\mathtt{G} \to \mathtt{T}) \\ P(\mathtt{C} \to \mathtt{A}) & P(\mathtt{C} \to \mathtt{G}) & P(\mathtt{C} \to \mathtt{C}) & P(\mathtt{C} \to \mathtt{T}) \\ P(\mathtt{T} \to \mathtt{A}) & P(\mathtt{T} \to \mathtt{G}) & P(\mathtt{T} \to \mathtt{C}) & P(\mathtt{T} \to \mathtt{T}) \end{pmatrix}$$

The change between probability distributions π_i is computed as

$$\pi_1 = \pi_0 \cdot M$$

$$\pi_x^1 = \pi_{\mathtt{A}}^0 P(\mathtt{A} \to x) + \pi_{\mathtt{G}}^0 P(\mathtt{G} \to x) + \pi_{\mathtt{C}}^0 P(\mathtt{C} \to x) + \pi_{\mathtt{T}}^0 P(\mathtt{T} \to x)$$

Markov Processes: Continuous Approach

Hypothesis

- Substitution events ruled by the same instantaneous transition matrix R.
- Substitution events follow a homogeneous Poisson distribution.

Then the transition matrix corresponding to time t is:

$$M(t) = \sum_{k=0}^{\infty} \frac{(\mu t)^k \cdot e^{-\mu t}}{k!} R^k = \cdots = e^{Qt}, \qquad \text{where } Q = \mu(R - Id)$$

The matrix Q is the instantaneous rate matrix ruling the Markov process.

Nucleotide substitution models

Each evolutionary model assumes meaningful constraints on the set of substitution probabilities or rates :

- Algebraic models : Constraints on probabilities (M).
- Continuous-time models: Constraints on rates (Q) .

A matrix structure provides two different but related models.

Embedding Problem (Related questions)

- *Which Markov matrices are rejected/considered for each approach?*
- *Are we allowed to concatenate homogeneous processes?*

Drawbacks

- Algebraic models consider unrealistic matrices.

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

- The product of embeddable matrices is not necessarily an embeddable matrix.
- There are Markov matrices close to Id that are not the exponential of any rate matrix.

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1-\varepsilon_1 & \varepsilon_1 \\ 0 & \varepsilon_2 & 0 & 1-\varepsilon_2 \end{pmatrix}$$

Rate identifiability

If the determinant is small enough there might be more than one generator:

$$Q1 = \begin{pmatrix} -\pi & 0 & 0 & \pi \\ 0 & -\pi & \pi & 0 \\ \pi & 0 & -\pi & 0 \\ 0 & \pi & 0 & -\pi \end{pmatrix} \qquad Q_2 = \begin{pmatrix} -\pi & 0 & \pi & 0 \\ 0 & -\pi & 0 & \pi \\ \pi & 0 & -\pi & 0 \\ 0 & \pi & 0 & -\pi \end{pmatrix}.$$

Identifiability Problem

Given an embeddable Markov matrix, is there a unique Markov generator? If not, how many Markov generators does it admit?

Enumerating all the logarithms of a matrix

All the logarithms of a (non-singular) diagonalizable matrix are given by choosing a diagonalizing basis and a determination of the logarithm for each of its eigenvalues:

Theorem

Given a non-singular matrix $M = P\, diag(\lambda_1, \lambda_2, \dots, \lambda_n)\, P^{-1}$*, then any solution* Q *of the equation* $M = e^Q$ *can be expressed as:*

$$Q = (P\,A)\; diag\big(\log_{k_1}(\lambda_1), \log_{k_2}(\lambda_2), \dots, \log_{k_n}(\lambda_n)\big)\; (P\,A)^{-1}$$

for some $k_1, k_2, \dots, k_n \in \mathbb{Z}$, $A \in Comm^*(diag(\lambda_1, \lambda_2, \dots, \lambda_n))$.

The principal logarithm

Any non-singular matrix $A \in GL_n(\mathbb{R})$ has a unique logarithm, called the **principal logarithm** $Log(A)$, all of whose eigenvalues lie in the strip $\{z \in \mathbb{C} \mid -\pi < Im(z) \leq \pi\}$.

- $Log(M)$ is real iff M has no negative eigenvalues.
- $Log(M)$ is the only possible Markov generator if M is close enough to Id.

  Cuthbert(1972), Singer and Spilerman(1976), Israel et al.,(2001)

Theorem (Singer and Spilerman(1976))

Solution to the embedding problem and the rate identifiability problem for Markov matrices with different real eigenvalues.

Embeddability of $n \times n$ Markov matrices

Lemma

Let Q be a rate matrix. Then for any eigenvalue $\lambda \in \sigma(Q)$ we have

$$|\mathrm{Im}(\lambda)| \leq \min\left\{\sqrt{2\,\mathrm{tr}(Q)\,\mathrm{Re}(\lambda) - (\mathrm{Re}(\lambda))^2}, -\frac{\mathrm{Re}(\lambda)}{\tan(\pi/n)}\right\}.$$

We can bound the number of real logarithms with rows summing to zero in terms of the eigenvalues and the determinant of M.

Theorem

(Algorithmic) Solution to the embedding problem for a dense subset of $n \times n$ Markov matrices (for any n).

MC, JFS and **JRL** *The embedding problem for Markov matrices.* arXiv:2005.00818

Embeddability of 4×4 Markov matrices

Theorem

Let $M = P diag(1, \lambda_1, \lambda_2, \lambda_3) P^{-1}$, be a Markov matrix with different eigenvalues $\lambda_1 \in \mathbb{R}$, $\lambda_2, \lambda_3 \in \mathbb{C}$.

$\lambda_2, \lambda_3 \in \mathbb{R}$: *Let V be the zero matrix and $\mathcal{L} = \mathcal{U} = 0$.*

$\lambda_2, \lambda_3 \notin \mathbb{R}$: *Take $V = P\ diag(0, 0, 2\pi i, -2\pi i)\ P^{-1}$*

$$\mathcal{L} := \max_{i \neq j,\ V_{i,j} > 0} \left\lceil -\frac{Log(M)_{i,j}}{V_{i,j}} \right\rceil, \quad \mathcal{U} := \min_{i \neq j,\ V_{i,j} < 0} \left\lfloor -\frac{Log(M)_{i,j}}{V_{i,j}} \right\rfloor.$$

M is embeddable if and only if

- $\lambda_i \notin \mathbb{R}_{\leq 0}$,
- $\{(i,j) : i \neq j,\ V_{i,j} = 0$ *and* $Log(M)_{i,j} < 0\} = \emptyset$,
- $\mathcal{L} \leq \mathcal{U}$.

The Markov generators of M are $Log(M) + kV$ with $k \in [\mathcal{L}, \mathcal{U}]$.

Embeddability of 4×4 Markov matrices

Sketch of proof

$\lambda_2, \lambda_3 \in \mathbb{R}$

M is embeddable $\Leftrightarrow$ Log(M) is a rate (even if $\lambda = 1$).

$\lambda_2, \lambda_3 \notin \mathbb{R}$

All the real logarithms with rows summing to 0 can be written as $Log_k(M) := Log(M) + k \cdot V$ for some $k \in Z$.

$Log_k(M)$ is a rate matrix if and only if $N = \emptyset$ and $\mathcal{L} \leq k \leq \mathcal{U}$.

MC, JFS and **JRL** *The embedding problem for Markov matrices.* arXiv:2005.00818

Embeddability of 4×4 Markov matrices

Theorem

For all $k \in \mathbb{Z}$, there is a non-empty open set of embeddable Markov matrices whose unique Markov generator is Log_k.

MC, JFS and **JRL** *An open set of 4×4 embeddable matrices whose principal logarithm is not a Markov generator.* To appear in Linear and Multilinear Algebra.

In particular, there is a non-empty Euclidean open set of 4×4 Markov matrices that are embeddable and whose principal logarithm is not a rate matrix.

Embeddability of 4×4 Markov matrices

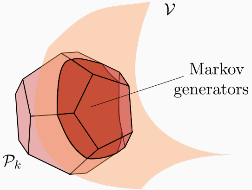
Special case: $M = P\mathrm{diag}(1, \lambda, \mu, \mu)P^{-1}$ with $\lambda \geq 0$, $\mu \neq 0, \lambda$.

- Define $Q_k(x,y,z) = L + (2\pi k + Arg\mu)\ V(x,y,z)$, where

$$L = P\ \mathrm{diag}(0, \log(\lambda), \log|\mu|, \log|\mu|)\ P^{-1}$$

$$V(x,y,z) = P\ \mathrm{diag}\left(0, 0, \begin{pmatrix} -y & x \\ -z & y \end{pmatrix}\right)\ P^{-1}.$$

- $\mathcal{P}_k = \left\{(x,y,z) \in \mathbb{R}^3 : Q_k(x,y,z) \text{ is a rate matrix}\right\}$.
  $\mathcal{V} = \left\{(x,y,z) \in \mathbb{R}^3 :\ x > 0,\ z > 0 \text{ and } xz - y^2 = 1\right\}$.



Embeddability of 4×4 Markov matrices

Special case: $M = P\text{diag}(1, \lambda, \lambda, \lambda)P^{-1}$ with $\lambda \in \mathbb{R}$.

- $\text{Log}(M) = \frac{-\log(\lambda)}{1-\lambda}(M - Id)$.
- The following are equivalent:
  - i) M is embeddable.
  - ii) $\det(M) > 0$.
  - iii) $\text{Log}(M)$ is a rate matrix.

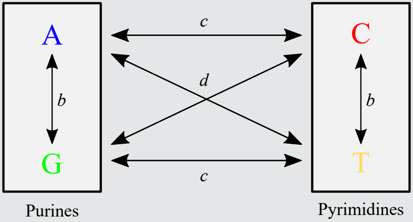
Special case: M does not diagonalize.

$$M \text{ embeddable} \Leftrightarrow \text{Log}(M) \text{ is a rate matrix.}$$

Evolutionary Models: Kimura models

Definition

Kimura 3-parameter model (K3P) assign probabilities/rates depending on the types of the substitution.



$$\begin{pmatrix} \cdot & b & c & d \\ b & \cdot & d & c \\ c & d & \cdot & b \\ d & c & b & \cdot \end{pmatrix}$$

- K2P model: $c = d$.
- JC model: $b = c = d$.

Embedding problem: Kimura 3-parameter model

Theorem

Let M be a Markov K3ST matrix with eigenvalues $1, x, y, z$.

- *If all eigenvalues are positive, then*

$$M \text{ is embeddable } \Leftrightarrow x \geq yz,\ y \geq xz,\ z \geq xy.$$

- *If M has a negative eigenvalue, say x, then*

$$M \text{ is embeddable } \quad \Leftrightarrow \quad x \text{ has algebraic multiplicity 2, and } x^2 \leq y \leq e^{-2\pi}.$$

Relative volume of K3P embeddable matrices $= 0.09375$.

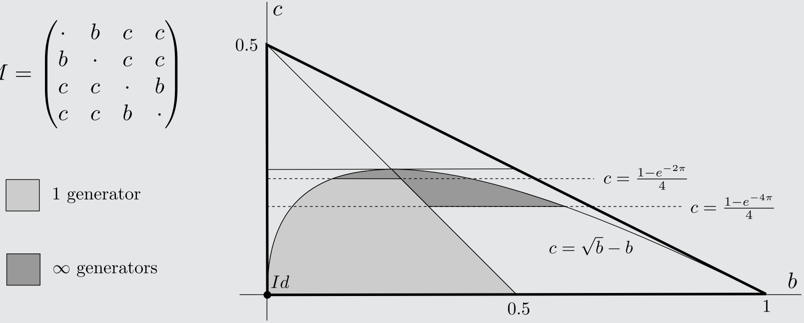
JRL and JFS, *Embeddability of Kimura 3ST Markov matrices*, Journal of theoretical biology 445)

D. Kosta and K. Kubjas. *Geometry of time-reversible group-based models.*

Embedding problem: Kimura 2-parameter model

Theorem

$$M = \begin{pmatrix} \cdot & b & c & c \\ b & \cdot & c & c \\ c & c & \cdot & b \\ c & c & b & \cdot \end{pmatrix}$$



Relative volume of K2P embeddable matrices $= \frac{(1+e^{-3\pi})}{3}$.

MC, JFS and **JRL** *Embeddability and rate identifiability of Kimura 2ST Markov matrices*, Journal of Mathematical Biology Nov-2019.

Some other nucleotide substitution models

Definition

The *strand symmetric model* (SSM) takes into account the double strand structure of DNA (Watson-Crick base pairing `A–T`, `C–G`),

$$\begin{pmatrix} \cdot & b & c & d \\ e & \cdot & g & h \\ h & g & \cdot & e \\ d & c & b & \cdot \end{pmatrix}$$

Definition

The *General Markov model* (GM) allows any 4×4 Markov/rate matrix a nucleotide substitution process.

Embedding problem: Nucleotide substitution models

Corollary

Percentage of embeddable matrices w.r.t all 4×4 Markov matries (GM model), the strand symmetric model (SSM), the K3P model and its submodels.

Model	*Dimension*	*Embeddable*
JC	*1*	75%
K2P	*2*	$33.336\ldots\%$
K3P	*3*	9.375%
SSM	*6*	$\sim 1.745\%$
GM	*12*	$\sim 0.05774\%$

Acknowledgements

AGAUR

Generalitat de Catalunya

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH