ANZIAM J. 59(2017), 240–246
doi:10.1017/S1446181117000359

MULTIPLICATIVELY CLOSED MARKOV MODELS MUST FORM LIE ALGEBRAS

JEREMY G. SUMNER[1]

(Received 17 April, 2017; accepted 31 July, 2017; first published online 23 October 2017)

Abstract

We prove that the probability substitution matrices obtained from a continuous-time Markov chain form a multiplicatively closed set if and only if the rate matrices associated with the chain form a linear space spanning a Lie algebra. The key original contribution we make is to overcome an obstruction, due to the presence of inequalities that are unavoidable in the probabilistic application, which prevents free manipulation of terms in the Baker–Campbell–Haursdorff formula.

2010 *Mathematics subject classification*: primary 17B45; secondary 60J27.

Keywords and phrases: Lie algebras, continuous-time Markov chains, semigroups, phylogenetics.

1. Background

In this note, we prove a result which makes explicit the requirement that a *multiplicatively closed* Markov model must form a Lie algebra (definitions will be provided). We consider continuous-time Markov chains and work under the general assumption that a model is determined by specifying a subset of rate matrices (or rate generators). These models are used in a wide array of scientific modelling problems and have been previously studied in the context of Lie group theory by Johnson [8] and Mourad [9]. Although the results given here are general, we are motivated primarily by applications to phylogenetics.

Phylogenetics consists of the mathematical and statistical methods applied to reconstructing evolutionary history from observed molecular sequences such as DNA [3]. Recent theoretical work [4, 12] has discussed the relevance of Lie groups and algebras to this applied area. The importance of Lie algebras for robust phylogenetic modelling has been demonstrated using simulation [13] as well as a diverse set of biological data sets [15]. The class of Markov models that form Lie algebras is discussed in the recent textbook on mathematical phylogenetics [11], and this approach

[1] University of Tasmania, Hobart 7000, Australia; e-mail: jsumner@utas.edu.au.

https://doi.org/10.1017/S1446181117000359 Published online by Cambridge University Press

also has important applications outside phylogenetic modelling [7]. However, previous work on this topic has not established the necessity of a Lie algebra in the general setting. In Theorem 2.4, we establish that the Lie algebra property is a consequence of modelling assumptions, which we claim are natural, easily understandable and well justified in an applied setting.

Our results fit into the general theory of Lie semigroups and Lie semialgebras as developed by Hilgert and Hofmann [6]. However, the approach we follow here gives the most direct path explicitly tailored to the practical setting of Markov chains, and does so with minimal abstract theory.

2. Main result

Fixing notation, we denote $\mathcal{L} \subset \mathrm{Mat}_n(\mathbb{R})$ as the set of real valued $n \times n$ matrices with zero-column sums, and we denote $\mathcal{L}^+ \subset \mathcal{L}$ as the subset of matrices with nonnegative off-diagonal entries. We then have the interpretation that $Q \in \mathcal{L}$ corresponds to a valid rate matrix for a continuous-time Markov chain if and only if $Q \in \mathcal{L}^+$. To distinguish from members of $\mathcal{L}$, we refer to the members of $\mathcal{L}^+$ as *stochastic*.

The reader who prefers to use row sums for matrices associated to a Markov chain, may simply modify the definitions above appropriately and read what follows without variation.

We assume that a given model is specified as a subset $\mathcal{R}^+ \subseteq \mathcal{L}^+$, which is defined either using a parametrization or by giving some (polynomial) constraints on the matrix entries. In phylogenetics, the former situation is the norm, and it is standard to use methods such as maximum likelihood to provide estimates of these model parameters. However, the former specification can usually be reinterpreted using the latter, which also plays a role in some formulations (such as the "group-based" [10, Ch. 8] and "equivariant" [2] model classes). An example of a popular phylogenetic model will be given in the next section. This motivates the following.

Property 2.1. A model $\mathcal{R}^+$ is expressible as an intersection $\mathcal{R}^+ = \mathcal{R} \cap \mathcal{L}^+$, where $\mathcal{R} \subseteq \mathcal{L}$ is determined by a finite set of polynomial constraints on the matrix entries of members of $\mathcal{L}$. That is, there exist polynomials $f_1, f_2, \ldots, f_r$ on the variables $Q = (q_{ij})$ such that $\mathcal{R} = \{Q \in \mathcal{L} \mid 0 = f_1(Q) = f_2(Q) = \cdots = f_r(Q)\} \subseteq \mathcal{L}$. Additionally, we demand that $\mathcal{R}$ is *minimal* in the sense that there is also no similarly constructed set $\mathcal{R}' \subset \mathcal{R}$ such that $\mathcal{R}^+ = \mathcal{R}' \cap \mathcal{L}^+$.

Although Property 2.1 does not imply that $\mathcal{R}$ is necessarily unique, the minimality condition ensures that $\mathcal{R}$ does not contain any member superfluous to the determination of $\mathcal{R}^+$. A simple example gives a clear motivating precedent for this condition, as follows.

Consider $(x, y) \in \mathbb{R}^2$ and the line $y = x$ restricted to the positive orthant

$$\{(x, y) \in \mathbb{R}^2 \mid x - y = 0,\ x, y \geq 0\}.$$

Clearly, we obtain this set most simply by taking the intersection of the positive orthant with the line $y = x$ (defined as the subset of $(x, y) \in \mathbb{R}^2$ satisfying the polynomial

https://doi.org/10.1017/S1446181117000359 Published online by Cambridge University Press

constraint $x - y = 0$). However, we may also obtain this set by taking the intersection of the positive orthant with the *pair* of lines defined by the quadratic constraint $y^2 = x^2$. In this case, analogous application of Property 2.1 would ensure that we choose the former possibility.

Following general Markov chain theory in the time-homogeneous setting, given some amount of elapsed time t, the probability substitution matrix associated with $Q \in \mathcal{R}^+$ is computed as the matrix exponential $M = e^{Qt}$ (using the power series $e^A = \sum_{m\geq 0}(A^m/m!)$). Since $t \geq 0$ may take on any nonnegative value, it is sensible to consider the following property.

Property 2.2. A model $\mathcal{R}^+$ is closed under nonnegative scalar multiplication. That is, for all $Q \in \mathcal{R}^+$ and $\alpha \geq 0$, it also follows that $\alpha Q \in \mathcal{R}^+$.

If Property 2.1 is assumed, Property 2.2 follows if and only if the polynomial constraints defining $\mathcal{R}$ are homogeneous. Up to conventions of exactly how models are parametrized (possibly obscured by conventions of overall scaling and "normalization"), as far as we are aware *all* phylogenetic models proposed in the literature have Property 2.2. When Property 2.2 is assumed, we may simplify the notation by writing e^Q in place e^{Qt}.

We now place a third reasonable restriction on a model $\mathcal{R}^+$ by imposing what we refer to as the *multiplicative closure.* This property is relevant in any setting that generalizes from the time-homogeneous to time-inhomogeneous formulation of continuous-time Markov chains. In rough terms, what we mean by this is that, if Q, Q' are in a model, then there exists another $\widehat{Q}$ also in the model such that $e^Q e^{Q'} = e^{\widehat{Q}}$. This question rouses the Baker–Campbell–Hausdorff (BCH) [1] formula for all $n \times n$ matrices A, B: that is,

$$\log(e^A e^B) = A + B + \tfrac{1}{2}[A, B] + \tfrac{1}{12}([A, [A, B]] + [B, [B, A]]) + \cdots$$

(where log is the matrix-log power series and $[A, B] = AB - BA$ is the "Lie bracket", or "commutator"). This naturally leads to a discussion of Lie algebras in the context of continuous-time Markov chains. Precisely how this arises is developed in the argument that follows. However, careful attention to detail must be shown since, for certain cases, it is possible that either (i) $\widehat{Q}$ does not belong to $\mathcal{L}^+$ or (ii) the BCH series does not converge. The obstruction we overcome in this note is that there is no immediate means available to isolate terms in the BCH series since, by construction, a model $\mathcal{R}^+$ does not form a linear space.

The definitions and notation required to state and prove our main result are given in the following steps.

(1) Let $\mathcal{M}$ be the semigroup generated by the set $\exp(\mathcal{R}^+) = \{e^Q \mid Q \in \mathcal{R}^+\}$. Equivalently, $\mathcal{M}$ is the intersection of all semigroups that contain $\exp(\mathcal{R}^+)$. (Notice that $\mathcal{M}$ includes the identity matrix since $I = e^{Q\cdot 0}$, so $\mathcal{M}$ is technically a monoid.)

https://doi.org/10.1017/S1446181117000359 Published online by Cambridge University Press

(2) Let $\overline{\mathcal{R}}$ be the set of (scaled) logarithms of the members of $\mathcal{M}$. Specifically, for what follows, it is sufficient to take $\overline{\mathcal{R}} = \{\alpha \log(M) \mid \alpha \geq 0, M \in \mathcal{M}\}$, where log is the matrix-log power series (wherever it converges). Since $\log(e^{Q\epsilon}) = Q\epsilon$ for sufficiently small ϵ, we see that $\mathcal{R}^+ \subseteq \overline{\mathcal{R}}$. In general, this definition allows for the circumstance that $\overline{\mathcal{R}}$ may contain rate matrices that are non-stochastic and/or *not members of* $\mathcal{R}$; the latter is our crucial observation.

We are now in a position to state our third proposed property for continuous-time Markov chains.

Property 2.3. A model $\mathcal{R}^+$ satisfies $\overline{\mathcal{R}} \subseteq \mathcal{R}$.

We claim that Property 2.3 is a very reasonable demand on a model, since it is saying that all expressions of the form $\log(e^{Q}e^{Q'}) = \widehat{Q}$ produce rate matrices $\widehat{Q}$ which satisfy the same constraints as the matrices Q, Q' (up to possible relaxation of the stochastic condition of membership in $\mathcal{L}^+$). When this is the case, we say that the model is *multiplicatively closed*.

Theorem 2.4. *Suppose that a model $\mathcal{R}^+$ satisfies Property 2.1. Then $\mathcal{R}^+$ satisfies Properties 2.2 and 2.3 if and only if $\mathcal{R} = \mathrm{span}_{\mathbb{R}}(\mathcal{R}^+)$, and this space forms a real Lie algebra.*

Proof. Assume, throughout, that $\mathcal{R}^+$ satisfies Property 2.1.

- Assume that $\mathcal{R}^+$ satisfies Properties 2.2 and 2.3. For $a, b \geq 0$ and $Q, Q' \in \mathcal{R}^+$, we also know that $aQ, bQ' \in \mathcal{R}^+$ (by Property 2.2). Then $\log(e^{aQ}e^{bQ'}) = aQ + bQ' + (1/2)ab[Q, Q'] + \cdots \in \overline{\mathcal{R}}$ for some choice of a, b small enough such that the series converges. By Property 2.3, $aQ + bQ' + (1/2)ab[Q, Q'] + \cdots \in \mathcal{R}$. Choosing $a = b$ and rescaling by a^{-1} (using Property 2.2), in the limit $a \to 0$, $Q + Q' \in \mathcal{R}$. Using Property 2.2, we observe that this generalizes to $\alpha Q + \beta Q' \in \mathcal{R}$ for all $\alpha, \beta \geq 0$. More specifically, since $\alpha Q + \beta Q' \in \mathcal{L}^+$ and $\mathcal{R}^+ = \mathcal{R} \cap \mathcal{L}^+$, $\alpha Q + \beta Q' \in \mathcal{R}^+$ for all $\alpha, \beta \geq 0$. Iterating this result, shows that $\alpha_1 Q_1 + \alpha_2 Q_2 + \cdots + \alpha_k Q_k \in \mathcal{R}$ for all choices $Q_i \in \mathcal{R}^+$ and $\alpha_i \geq 0$.
  However, since the constraints defining $\mathcal{R}$ are polynomial, this result must be true more generally for all choices of $\alpha_i \in \mathbb{R}$. Thus
  $$\mathrm{span}_{\mathbb{R}}(\mathcal{R}^+) = \{\alpha_1 Q_1 + \alpha_2 Q_2 + \cdots + \alpha_k Q_k \mid Q_i \in \mathcal{R}^+, \alpha_i \in \mathbb{R}\} \subseteq \mathcal{R},$$
  which yields
  $$\mathcal{R} \cap \mathcal{L}^+ = \mathcal{R}^+ \subseteq \mathrm{span}_{\mathbb{R}}(\mathcal{R}^+) \cap \mathcal{L}^+ \subseteq \mathcal{R} \cap \mathcal{L}^+,$$
  so the equality $\mathcal{R}^+ = \mathcal{R} \cap \mathcal{L}^+ = \mathrm{span}_{\mathbb{R}}(\mathcal{R}^+) \cap \mathcal{L}^+$ follows, and the minimality condition claimed by Property 2.1 shows that $\mathcal{R} = \mathrm{span}_{\mathbb{R}}(\mathcal{R}^+)$. Having established that $\mathcal{R}$ is a linear space, we can now freely isolate terms in the BCH formula and be guaranteed to stay in $\mathcal{R}$. In particular, taking $Q, Q \in \mathcal{R}^+$ and $\epsilon > 0$,
  $$\log(e^{\epsilon Q}e^{\epsilon Q'}) - (\epsilon Q + \epsilon Q') = \tfrac{1}{2}\epsilon^2[Q, Q'] + \cdots \in \mathcal{R}.$$

https://doi.org/10.1017/S1446181117000359 Published online by Cambridge University Press

So rescaling by $2\epsilon^{-2}$ and taking the limit $\epsilon \to 0$, $[Q, Q'] \in \mathcal{R}$ also. Thus $\mathrm{span}_{\mathbb{R}}(\mathcal{R}^+) = \mathcal{R}$ forms a real Lie algebra, as required.

- Assuming that $\mathcal{R} = \mathrm{span}_{\mathbb{R}}(\mathcal{R}^+)$ shows that the constraints defining $\mathcal{R}$ are linear, which implies that Property 2.2 is satisfied. Further, assuming that $\mathrm{span}_{\mathbb{R}}(\mathcal{R}^+)$ is a real Lie algebra and applying the BCH formula shows that each member of $\overline{\mathcal{R}}$ is a member of $\mathcal{R}$. Hence Property 2.3 is satisfied. □

3. Example

We illustrate this process with the Hasegawa–Kishino–Yano (HKY) [5] model of DNA substitutions. This is an example of a time-reversible model [14], and is defined via the parametrization (rows and columns ordered as A, G, C, T)

$$Q = \begin{pmatrix} * & \kappa\alpha_A & \alpha_A & \alpha_A \\ \kappa\alpha_G & * & \alpha_G & \alpha_G \\ \alpha_C & \alpha_C & * & \kappa\alpha_C \\ \alpha_T & \alpha_T & \kappa\alpha_T & * \end{pmatrix},$$

where the missing entries "$*$" are chosen to ensure unit column sums. The parameters $\alpha_i \geq 0$ are proportional to the equilibrium nucleotide frequencies of the Markov chain, and $\kappa \geq 0$ is included to accommodate the "transition/transversion" ratio (distinguishing substitutions within both the "purines" $A \leftrightarrow G$ and the "pyrimidines" $C \leftrightarrow T$ from substitutions across these two groups).

Equivalently, we may express the HKY model as the subset of the rate matrices

$$\mathcal{R}^+_{\mathrm{HKY}} = \left\{ Q \in \mathcal{L}^+ \;\middle|\; \begin{array}{c} q_{13} = q_{14}, q_{23} = q_{24}, q_{31} = q_{32}, q_{41} = q_{42} \\ q_{12}q_{23} = q_{21}q_{13}, q_{34}q_{13} = q_{12}q_{31}, q_{43}q_{13} = q_{12}q_{41}, \ldots \end{array} \right\}$$

(where the displayed constraints are sufficient to determine the model). We immediately see that $\mathcal{R}^+_{\mathrm{HKY}}$ is not multiplicatively closed since the defining constraints are not linear. To illustrate the issue, we give a numerical example.

We chose $Q_1, Q_2 \in \mathcal{R}^+_{\mathrm{HKY}}$ via $(\alpha_A, \alpha_G, \alpha_C, \alpha_T; \kappa) = (0.02, 0.01, 0.005, 0.009; 1.5)$ and $(0.03, 0.01, 0.006, 0.008; 1.4)$, respectively, and computed (using MATHEMATICA)

$$\log(e^{Q_1}e^{Q_2}) = \begin{pmatrix} -0.0571752 & 0.0718248 & 0.0498348 & 0.0498348 \\ 0.0291051 & -0.0998949 & 0.0200951 & 0.0200951 \\ 0.0109967 & 0.0109967 & -0.0947047 & 0.0158953 \\ 0.0170734 & 0.0170734 & 0.0247748 & -0.0858252 \end{pmatrix}.$$

Attempting to find this matrix in the set $\mathcal{R}^+_{\mathrm{HKY}}$, we are immediately led to

$$(\alpha_A, \alpha_G, \alpha_C, \alpha_T) = (0.0498348, 0.0200951, 0.0109967, 0.0170734),$$

but no consistent solution for κ is obtainable (in fact, four different values are required). Therefore $\log(e^{Q_1}e^{Q_2})$ is not a member of $\mathcal{R}^+_{\mathrm{HKY}}$ (or indeed $\mathcal{R}_{\mathrm{HKY}}$ under relaxation of the stochastic conditions).

https://doi.org/10.1017/S1446181117000359 Published online by Cambridge University Press

Following the definitions given in the previous section, the form obtained in this example does, however, suggest that all rate matrices in the closure $\overline{\mathcal{R}}_{\mathrm{HKY}} = \log(\mathcal{M}_{\mathrm{HKY}})$ are of the form

$$\begin{pmatrix} * & \kappa_1 & \alpha & \alpha \\ \kappa_2 & * & \beta & \beta \\ \gamma & \gamma & * & \kappa_3 \\ \delta & \delta & \kappa_4 & * \end{pmatrix}.$$

That this is indeed the case, is confirmed by the following two simple computations.

(i) $\mathrm{span}_{\mathbb{R}}(\mathcal{R}^{+}_{\mathrm{HKY}}) = \left\{ \left(\begin{smallmatrix} * & \kappa_1 & \alpha & \alpha \\ \kappa_2 & * & \beta & \beta \\ \gamma & \gamma & * & \kappa_3 \\ \delta & \delta & \kappa_4 & * \end{smallmatrix} \right) \mid \alpha, \beta, \gamma, \delta, \kappa_1, \ldots, \kappa_4 \in \mathbb{R} \right\}.$

(ii) This set forms a Lie algebra (in fact, this is the Model 8.8 in Lie-Markov hierarchy [4]).

Thus, the span of the HKY model forms a Lie algebra (without the additional need to take closure under Lie brackets).

4. Discussion

The contribution of this work is to lay out conditions on a continuous-time Markov chain (Properties 2.1, 2.2, 2.3) in order to show that multiplicative closure necessitates that the associated rate matrices are minimally contained inside a Lie algebra. The conditions need to be set up carefully in order to, firstly, be convincingly well motivated to the applied setting and, secondly, allow for the relatively elementary proof of the main result (Theorem 2.4). This result provides a solid justification (albeit post-hoc) for recent work exploring the classification, enumeration and application of this class of Markov models [4, 12].

We focussed on continuous-time models and hence naturally assume that a model is defined in terms of its rate matrices (compare with Property 2.1). This does, however, leave open the possibility that a Markov chain defined *solely at the level of substitution matrices* could be multiplicatively closed without necessitating the existence of an associated Lie algebra (constructed as the tangent space at the identity). We conjecture that this is not possible, but leave the details for future work.

Acknowledgements

This research was supported by Australian Research Council (ARC) Discovery Grant DP150100088. I would like to thank Michael Woodhams, Barbara Holland and the anonymous reviewers for their helpful comments on this work.

References

[1] J. E. Campbell, “On a law of combination of operators”, *Proc. Lond. Math. Soc.* (3) **29** (1898) 14–32; doi:10.1112/plms/s1-29.1.14.

[2] J. Draisma and J. Kuttler, “On the ideals of equivariant tree models,”, *Math. Ann.* **344** (2009) 619–644; doi:10.1007/s00208-008-0320-6.

https://doi.org/10.1017/S1446181117000359 Published online by Cambridge University Press

[3] J. Felsenstein, *Inferring phylogenies* (Sinauer Associates, Sunderland, 2004); https://global.oup.com/ushe/product/inferring-phylogenies-9780878931774?.

[4] J. Fernández-Sánchez, J. G. Sumner, P. D. Jarvis and M. D. Woodhams, "Lie Markov models with purine/pyrimidine symmetry", *J. Math. Biol.* **70** (2015) 855–891; doi:10.1007/s00285-014-0773-z.

[5] M. Hasegawa, H. Kishino and T. Yano, "Dating of human-ape splitting by a molecular clock of mitochondrial DNA", *J. Mol. Evol.* **22** (1985) 160–174; doi:10.1007/BF02101694.

[6] J. Hilgert and K. H. Hofmann, "Semigroups in Lie groups, semialgebras in Lie algebras", *Trans. Amer. Math. Soc.* **288** (1985) 481–504; doi:10.1090/S0002-9947-1985-0776389-7.

[7] T. House, "Lie algebra solution of population models based on time-inhomogeneous Markov chains", *J. Appl. Probab.* **49** (2012) 472–481; doi:10.1239/jap/1339878799.

[8] J. E. Johnson, "Markov-type Lie groups in GL(n, R)", *J. Math. Phys.* **26** (1985) 252–257; doi:10.1063/1.526654.

[9] B. Mourad, "On a Lie-theoretic approach to generalized doubly stochastic matrices and applications", *Linear Multilinear Algebra* **52** (2004) 99–113; doi:10.1080/0308108031000140687.

[10] C. Semple and M. Steel, *Phylogenetics* (Oxford University Press, Oxford, 2003); https://global.oup.com/academic/product/phylogenetics-9780198509424?cc=au&lang=en&.

[11] M. Steel, *Phylogeny: Discrete and random processes in evolution* (SIAM, Philadelphia, PA, 2016); doi:10.1137/1.9781611974485.

[12] J. G. Sumner, J. Fernández-Sánchez and P. D. Jarvis, "Lie Markov models", *J. Theoret. Biol.* **298** (2012) 16–31; doi:10.1016/j.jtbi.2011.12.017.

[13] J. G. Sumner, P. D. Jarvis, J. Fernández-Sánchez, B. T. Kaine, M. D. Woodhams and B. R. Holland, "Is the general time-reversible model bad for molecular phylogenetics?", *Syst. Biol.* **61** (2012) 1069–1074; doi:10.1093/sysbio/sys042.

[14] S. Tavaré, "Some probabilistic and statistical problems in the analysis of DNA sequences", *Lect. Math. Life Sci. (American Society)* **17** (1986) 57–86; https://www.scopus.com/record/display.uri?eid=2-s2.0-60649089791&origin=inward&txGid=47ea59cb4b0bb9183bac8afba8e05dc1.

[15] M. D. Woodhams, J. Fernández-Sánchez and J. G. Sumner, "A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates", *Syst. Biol.* **64** (2015) 638–650; doi:10.1093/sysbio/syv021.

https://doi.org/10.1017/S1446181117000359 Published online by Cambridge University Press