EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Humanities

Department of Linguistics

DGfS 2021
Workshop *Model and Evidence in Quantitative Comparative Linguistics* Februa

Models in quantitative comparative linguistics

Gerhard Jäger, Tübingen University

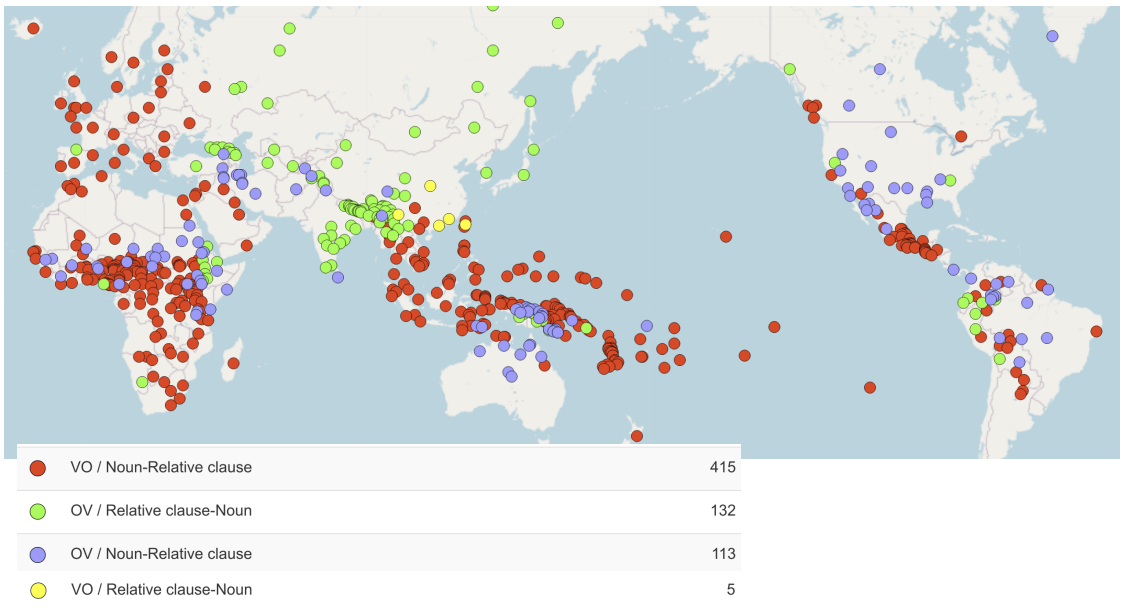
European Research Council
Established by the European Commission

Introduction

- large portion of current work in the field consists of *model fitting*
- common models:
  - continuous time Markov chain
  - mixed-effects regression
  - birth-death tree distributions with relaxed molecular clock
  - pair-Hidden Markov Models (tacitly underly many alignment studies)
  - ...
- comparatively little attention to *model criticism* and *model checking* in our field
- We can learn something from other fields, such as psychology!

A case study: Typological word order correlations

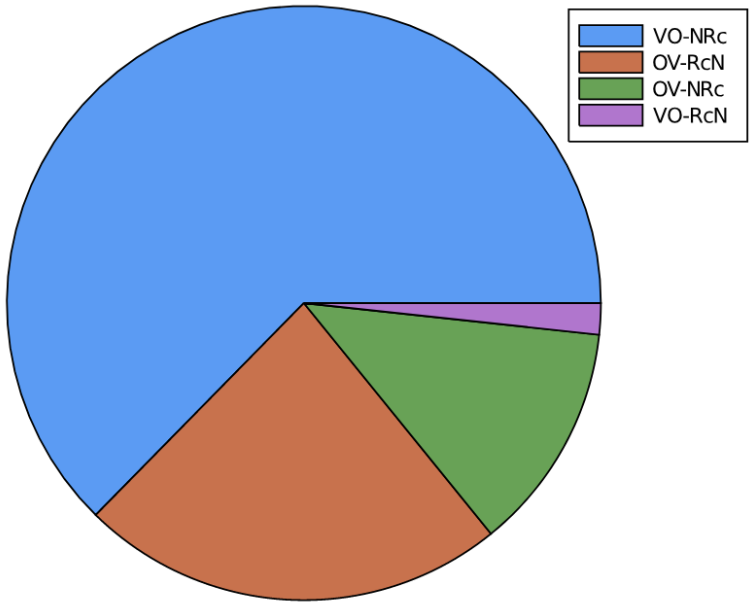
Distribution of verb-object/object verb vs. noun-relative clause/relative clause-noun



VO vs. NRc

this study:

- word-order data from WALS
- 1,060 languages
- 94 families + 81 isolates = 175 lineages



Steps of (Bayesian) model validation

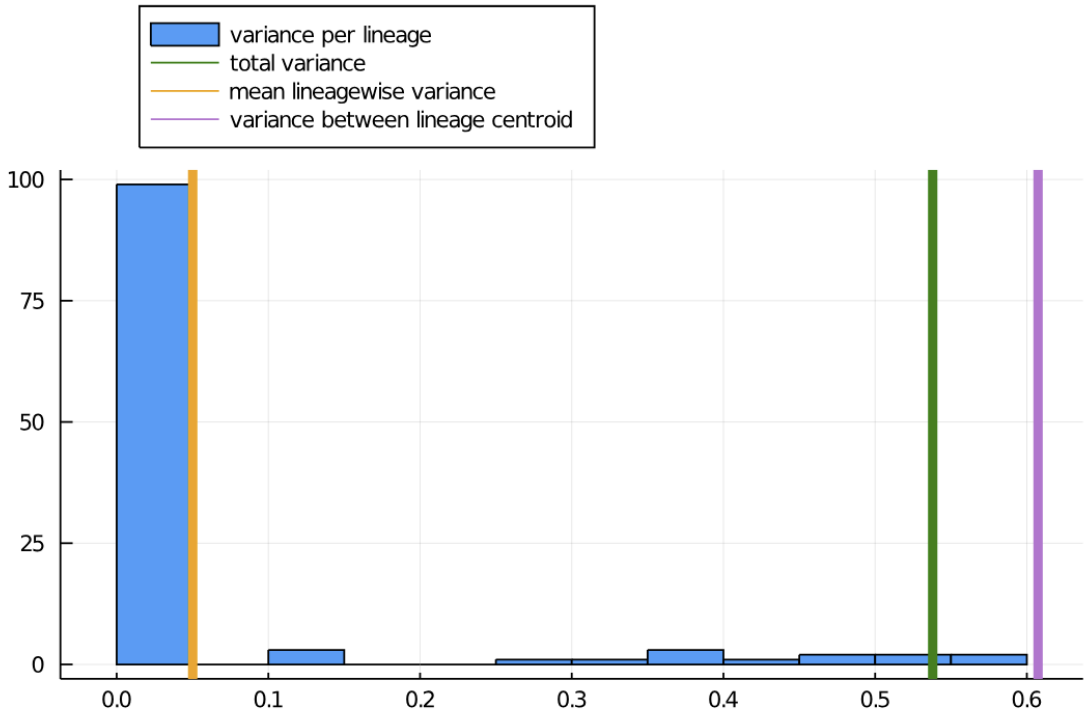
- exploratory data exploration $\rightarrow$ descriptive statistics
- specification of (a) generative probablistic model(s)
- prior predictive simulation
- model fitting
- posterior predictive simulation
- model comparison

(cf., eg., Gelman et al. 2014)

Descriptive statistics

- each language can be represented as a binary vector over 4 variables (for the four combinations of OV/VO and NRc/RcN)
- the **total variance** is the sum of the variance of those for binary variables
- the **mean lineage-wise variance** is the average total variance per lineage
- the **between-family variance** is the total variance between the centroids for each family

Descriptive statistics

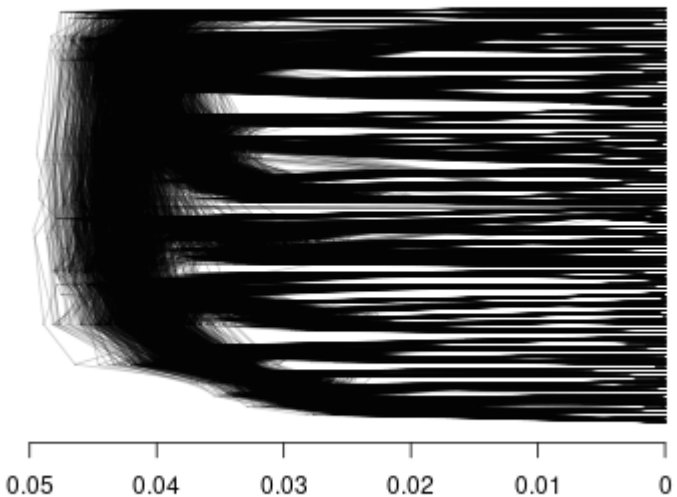


Defining models

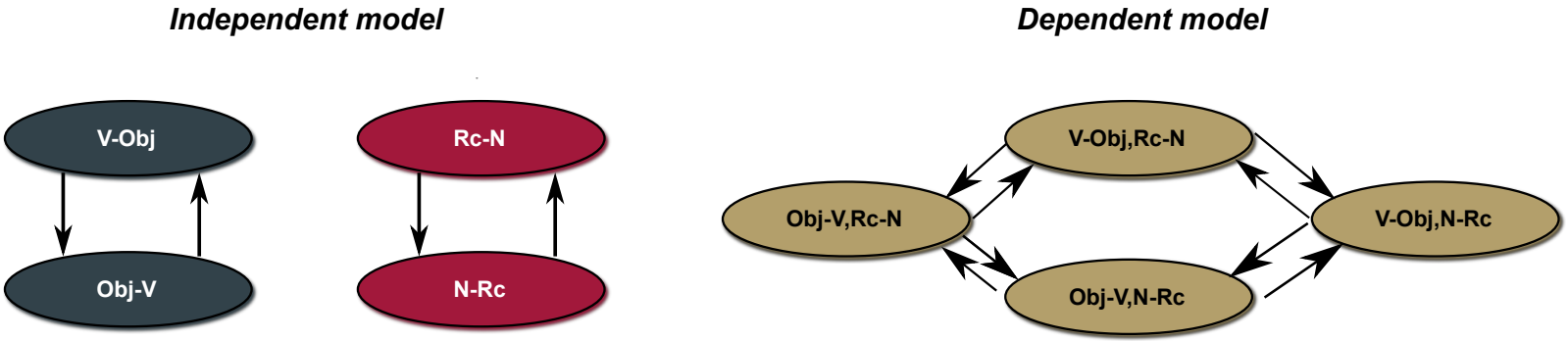
- feature values evolve according to a *continuous time Markov chain* (CTMC)
- evolution along a phylogeny
- phylogenetic tree is only partially known - represented here as posterior distribution of Bayesian phylogenetic inference from lexical data (from ASJP)

Phylogenies

- 1,000 trees from a MrBayes run for each family
- degenerate 1-node tree for isolates



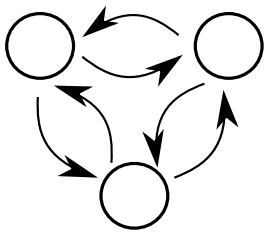
CTMC



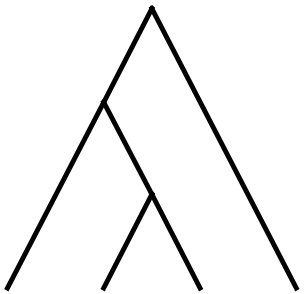
(following Pagel and Meade 2006; Dunn et al. 2011)

CTMC

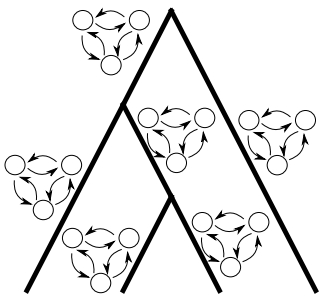
Markov process



Phylogeny



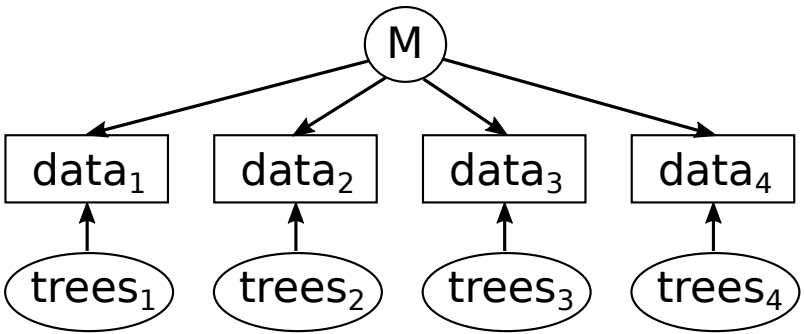
Branching process



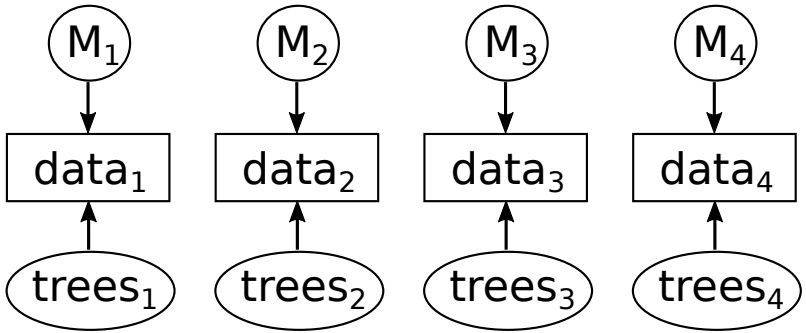
Lineage dependeny

two types of models

universal



lineage specific



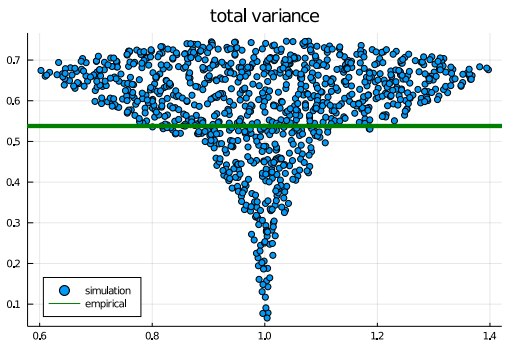
Prior predictive check

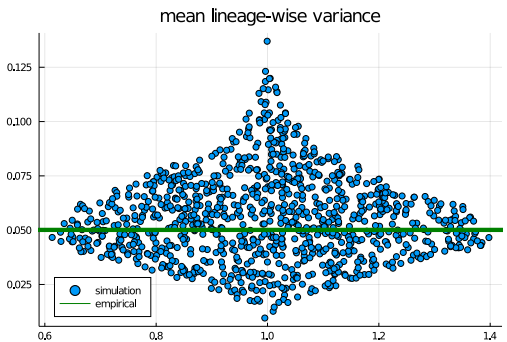
- all models use the same prior for rates:

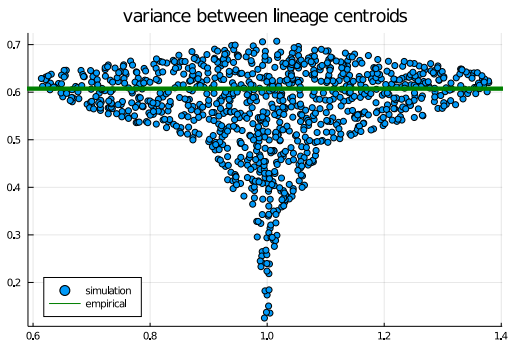
$$\text{rate}_i \sim \text{LogNormal}(0, 1)$$

- universal models: one set of rates across lineages
- lineage-dependent models: different set of rates for each lineage
- dependent features model: 8 rates per set
- independent features model: 4 rates per set

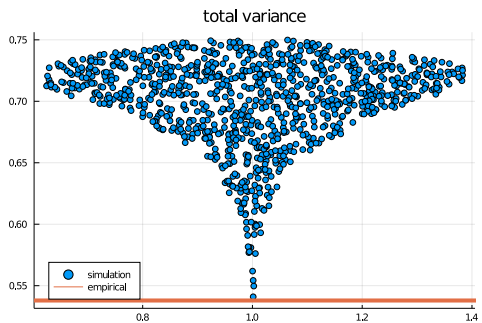
universal rates, dependent features

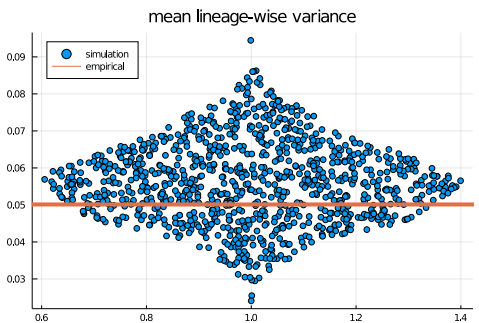


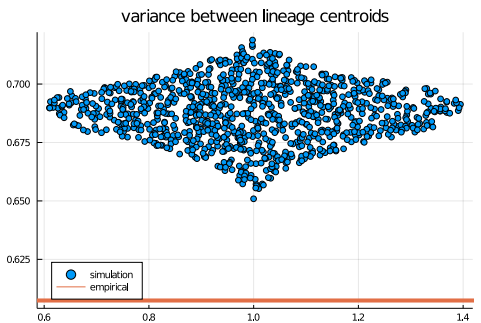




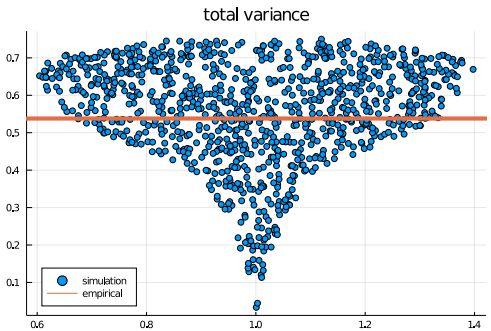
lineage-dependent rates, dependent features

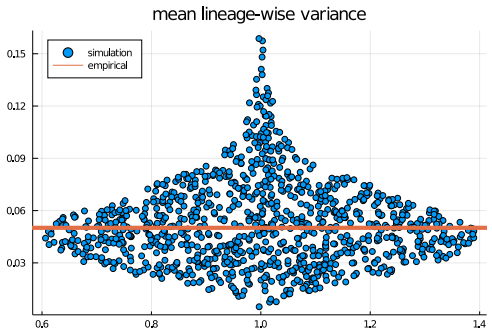


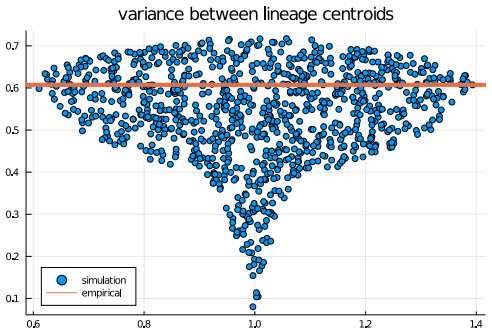




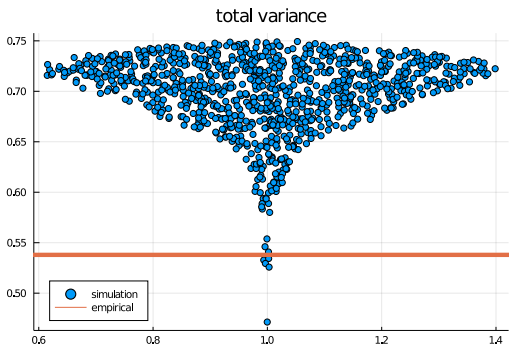
universal rates, independent features

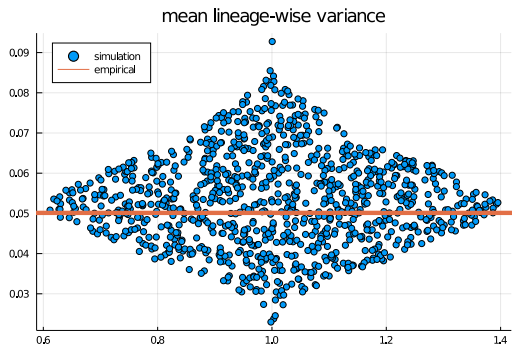


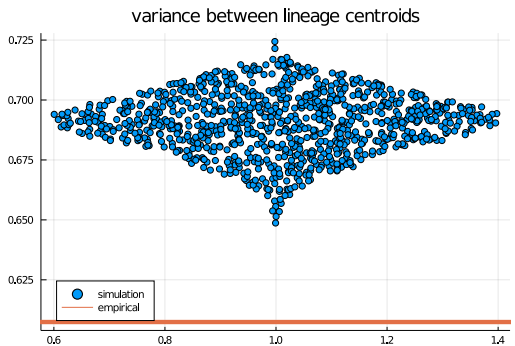




lineage-dependent rates, independent features





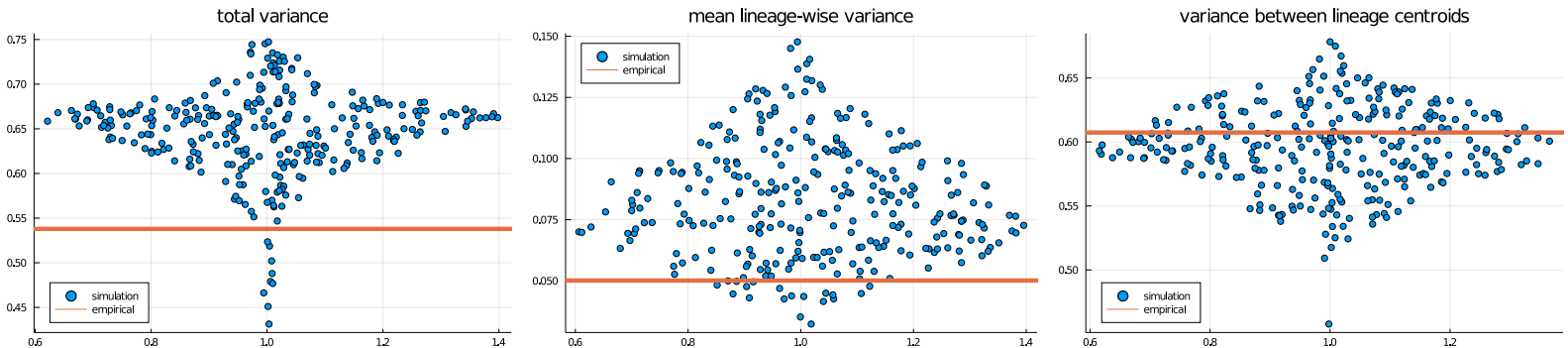


Run MCMC to infer posterior distribution

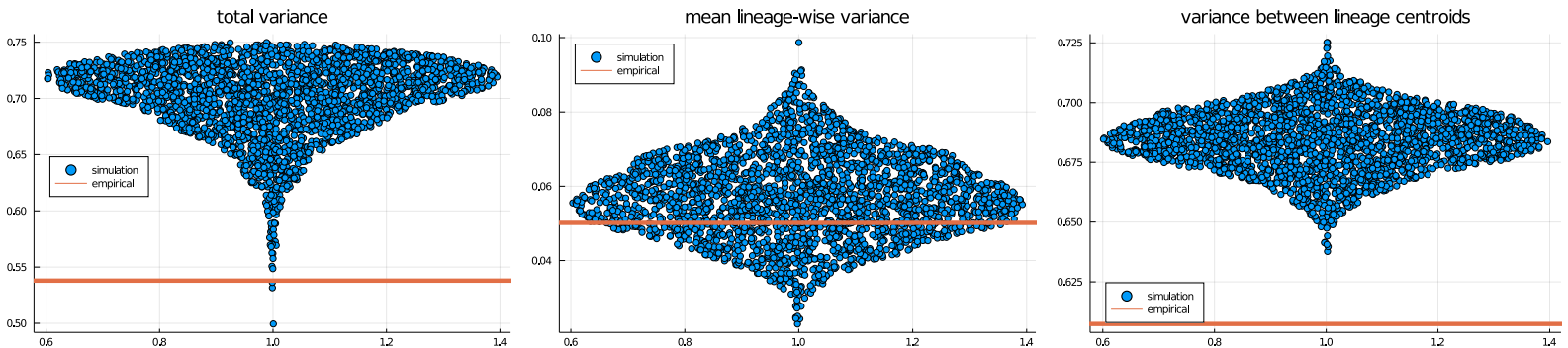
- here: done with Johannes Wahle's *Julia* package *Julia_Tree*
- currently under submission
- If you want to give it a try yourself, get in touch with Johannes

Posterior predictive check

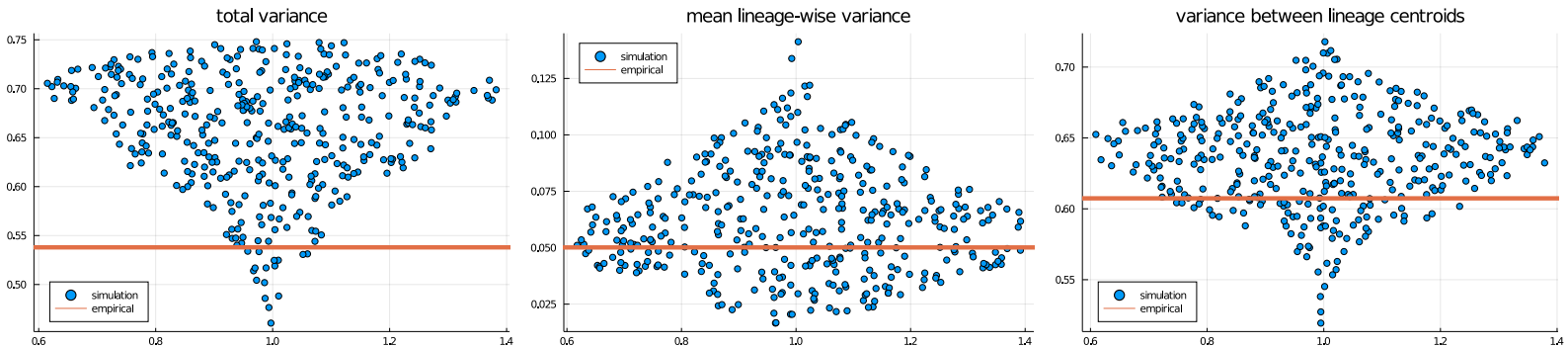
PPC: universal rates, dependent features



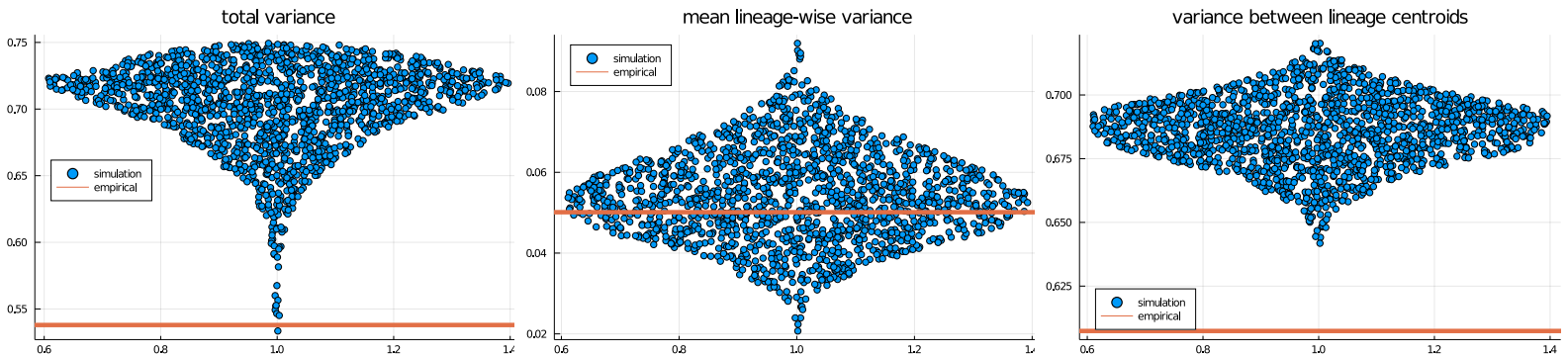
PPC: lineage-dependent rates, dependent features



PPC: universal rates, independent features



PPC: lineage-dependent rates, independent features



Model comparison

Bayes factor

(determined via bridge sampling)

rates	features	BF to best model
universal	dependent	0.0
universal	independent	-19.4
lineage-dependent	dependent	-24.3
lineage-dependent	independent	-31.7

Strong evidence for model with universal rates and dependent features.

Leave-one-out cross-validation

- computationally too expensive to carry out
- can be approximated via *Pareto-smoothe leave-one-out cross-validation* (Aki Vehtari, Andrew Gelman, and Jonah Gabry, 2016, “Efficient implementation of leave-one-out cross-validation and WAIC for evaluating fitted Bayesian models”, implemented in R-package *loo*)
- approximation depends on *conditional independence of observations*
- can be interpreted in two ways here
  - Each language is an observation. To achieve conditional indendence, we have to sample from posterior distribution of ancestral states. Can be done via simmap.
  - Each lineage (family or isolate) is an observation. Conditional independence for mcmc posterior sample.

LOO over languages

rates	features	Δ expected log-pointwise density
universal	dependent	102.4
universal	independent	0.0
lineage-dependent	dependent	202.8
lineage-dependent	independent	217.1

LOO over lineages

rates	features	Δ expected log-pointwise density
universal	dependent	0.0
universal	independent	54.7
lineage-dependent	dependent	75.5
lineage-dependent	independent	90.0

Reflections

- prior and posterior descriptive checks, as well as model comparison clearly favors universal rates over lineage-dependent ones
- to predict the feature values of a language from all other languages (including those in the same family), the independent model is the best
- to predict the distribution in an unknown family from the behavior of known families, dependent features do a better job.
- the latter question is of greater linguistic interest, so we can cautiously conclude that there is a correlation between verb-object order and noun-relative clause order

Reflections

- All these techniques assess the **predictive performance** of models
- A good predictive model may be a poor scientific model though.
- Good predictive performance is a necessary but not a sufficient condition for model evaluation.

Michael Dunn, Simon J. Greenhill, Stephen Levinson, and Russell D. Gray. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82, 2011.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, Boca Raton, 2014.

Mark Pagel and Andrew Meade. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *The American Naturalist*, 167(6):808–825, 2006. doi: 10.1086/503444.