Dyr – A Bayesian MCMC System for Inference of Language Phylogenies

Student Name: J.G. Byrne
Supervisor Name: Professsor M.J.R. Bordewich
Submitted as part of the degree of MEng Computer Science to the
Board of Examiners in the Department of Computer Sciences, Durham University

Abstract—Bayesian statistics provide a powerful framework for inferring the evolutionary history of language families from lexical data, a problem which is made computationally tractable by Markov Chain Monte Carlo (MCMC) algorithms. We present Dyr, a simple yet capable new system for Bayesian inference, and use it to replicate state-of-the-art results in the field of Indo-European linguistics.

Index Terms—Bayesian Inference, Markov Processes, Linguistics, Indo-European

✦

1 INTRODUCTION

LANGUAGES evolve over time. Some of these changes are phonetic or grammatical, but many are lexical (changes in the vocabulary). Words are forgotten and replaced, sometimes by other words that have changed meaning, and sometimes by borrowings from other languages.[1] Occasionally, some speakers of a language will come to speak so differently to the others that the groups can no longer understand each other, and the language splits in two.

Scholars have noticed similarities between languages for millenia. In the 18th century, the field of comparative linguistics was born, which sought to systematically reconstruct the historical relationships between languages. The early comparative linguists observed that while languages can evolve, sunder, and go extinct, they rarely merge, and are never created anew. Therefore, from the 19th century onwards it became *de rigeur* to describe language descent with evolutionary trees, or 'phylogenies', and though not without criticism, this model remains predominant in linguistics to this day.

Historically, constructing phylogenies has been a human labour, relying on the judgement of expert scholars with deep knowledge of languages ancient and modern. Although fruitful, this approach is naturally susceptible to human biases and oversights. It also provides no way to determine the age of unattested ancestor languages other than the (admittedly finely-honed) intuition of learned intellectuals.

The problem of ancestral dating is of particular relevance to Indo-European, the largest and most studied language family in the world. Encompassing nearly all of the languages of Europe and a great many in Western Asia and India, the challenge of reconstructing the history of this vast grouping has captivated comparative linguists since the advent of the discipline. And yet, one central question is yet to be conclusively answered: where and when was Proto-Indo-European (PIE; the ancestor of all Indo-European languages) originally spoken? Two main theories abound. The first, known as the 'Kurgan' hypothesis, postulates an *Urheimat* (original homeland) in the Pontic-Caspian steppe north of the Black Sea.[2] Conversely, the second proposes an origin in Anatolia.[3] The crucial difference between the two theories is that while the former ascribes an age of 6000 years to PIE, the latter hinges on it being considerably more ancient, at approximately 8500 years old. A reliable estimate for the age of PIE - that is, the root age of the Indo-European language tree - could therefore be potent evidence for one theory over the other.

Since the millenium, researchers have found a new method to analyse language families in general and Indo-European in particular – Bayesian inference. By re-appropriating software designed to analyse genetic data and construct biological evolutionary trees, they have successfully inferred plausible dated phylogenies from large vocabulary databases. Such research feels tantalisingly close to an objective solution to linguistic enigmas such as that of Indo-European provenance.

However, in reality, the human factor still plays a role. Bayesian inference requires the specification of prior distributions, the choice of which can radically affect the results of the inference. For research to be rigorous and reliable, priors should be chosen carefully and with good justification.

However, much study in the relatively small field of 'Bayesian Phylolinguistics' falls short of this mark. Priors are often chosen seemingly out of habit or convention. Such laxness poses particular danger when these conventions have their origins in bioinformatics; a model which is sensible in the context of biological evolution may not be so logical when applied to languages.

---

1. For example, the Old English word *dēor* was displaced by the French 'animal', and survives only in the narrower sense 'deer'. However, the equivalent word in Danish, *dyr*, whence this project name, retains the older meaning. Both words derive from Proto-Germanic *deuza*, ultimately from Proto-Indo-European *d^hwes, meaning 'breath'.

2. A *kurgan* is a tumulus or burial mound. This theory is associates the spread of Indo-European with warlike tumulus-building charioteers.

3. This rather more sedate theory suggests a gradual proliferation of Indo-European in tandem with the spread of agriculture

We suggest that one reason for this tendency is the reliance on large, featureful bioinformatics software packages like BEAST2 and MrBayes. Any dedicated support for linguistics in these programs tends to be something of an afterthought, and their intimidating size makes them hard to understand in their totality. Therefore, we submit that to step out of the shadow of the older, larger discipline, it is desirable for Bayesian Phylolinguistics to have its own dedicated software package.

We therefore present *Dyr*, a new program for Bayesian Inference of linguistic phylogenies. Small enough that its source code may be read and understood in a day, and offering only the features required for linguistic analysis, *Dyr* is nonetheless built to be flexible and extensible. Capable of replicating state-of-the-art results in Bayesian Phylolinguistics, our hope and intention is that *Dyr* can serve as a worthy platform for trialling novel methods with sound linguistic justification.

2 RELATED WORK

This section presents a survey of existing work on the problems that this project addresses. It should be be-tween 2 to 4 pages in length. The rest of this section shows the formats of subsections as well as some general formatting information for tables, figures, references and equations.

3 METHODS

The aim of phylogenetic inference is to learn which phylogenies (dated trees) are most likely given a particular evolutionary model and a set of known data. This is known as the 'posterior likelihood'. In a linguistics context, the known data is typically 'lexical trait data' – that is, information about the vocabulary of the languages in question. The evolutionary model, which in Bayesian terms provides our 'prior likelihood', encapsulates our beliefs about how likely languages are to diverge and how rapidly their vocabulary changes. To allow our inferred phylogeny to best fit the signal present in the data, we also infer some parameters of our evolutionary model, though these too are subject to their own prior likelihood distributions.

The space of possible paramaterisations is very large and it is not typically feasible to derive a closed-form expression for the posterior likelihood distribution. However, Bayes' theorem allows us to assess the posterior likelihood of any specific choice of paramaterisation given the data. We therefore use a stochastic process called Markov Chain Monte Carlo (MCMC) to iteratively step-through the space of possible paramaterisations, with a preference for augmentations to the paramaterisation that improve the posterior likelihood. It is provable that the long-run outcome of this process will be to simulate the desired posterior distribution.

3.1 Likelihood

Bayes' theorem is stated as follows:

$$Pr(A \mid B) = \frac{Pr(B \mid A) \cdot Pr(A)}{Pr(B)} \tag{1}$$

In the context of phylogenetic inference, we seek to infer the probability distribution of possible parameterisations given the observed data. Therefore $Pr(B)$ corresponds to $Pr(x)$, the probability of the trait data, which is by definition equal to 1 since it has been observed. Meanwhile, $Pr(A)$ corresponds to $Pr(\Gamma)$, the prior likelihood of a given parameterisation. We define Γ more thoroughly below:

$\Gamma = (\psi, \omega, \lambda)$

ψ = parameters describing a dated tree

ω = parameters of the prior model for dated trees

λ = parameters of the prior model for trait evolution

We thus derive equation 2. The posterior likelihood is proportional to the likelihood of the data given the paramaterisation multiplied by the prior likelihood of the paramaterisation. This is a proportionality because we do not require that our prior distributions sum to 1.

$$Pr(\psi, \omega, \lambda \mid x) \propto Pr(x \mid \psi, \omega, \lambda) \cdot f(\psi \mid \omega) \cdot f(\omega) \cdot f(\lambda) \tag{2}$$

We define $\psi = (\tau, \delta)$, where τ is the topology of the tree and δ is the branch lengths. We consider the node or nodes with the longest path from the root to be at $t = 0$ and therefore in the present day, while all other nodes are understood to be at some $t > 0$, corresponding to their distance from the present in years. Naturally, any interior node is required to be older than its children. The particular composition of the tuples ω and λ will be defined subsequently.

3.2 Felsenstein's Algorithm

3.3 Prior Distributions

3.4 Metropolis-Hastings

4 IMPLEMENTATION

5 RESULTS

6 EVALUATION

7 CONCLUSION