Master's thesis
Master's Programme in Data Science

Phylogenetic Machine Learning Methods and Application to Mammal Dental Traits and Bioclimatic Variables

Mikko Olavi Niemi

April 22, 2020

Supervisor(s): Professor Indrė Žliobaitė

Examiner(s): Professor Indrė Žliobaitė
Professor Kai Puolamäki

University of Helsinki
Faculty of Science

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Science	Master's Programme in Data Science

Tekijä — Författare — Author

Mikko Olavi Niemi

Työn nimi — Arbetets titel — Title

Phylogenetic Machine Learning Methods and Application to Mammal Dental Traits and Bioclimatic Variables

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Master's thesis	April 22, 2020	83

Tiivistelmä — Referat — Abstract

Standard machine learning procedures are based on assumption that training and testing data is sampled independently from identical distributions. Comparative data of traits in biological species breaks this assumption. Data instances are related by ancestry relationships, that is phylogeny. In this study, new machine learning procedures are presented that take into account phylogenetic information when fitting predictive models. Phylogenetic statistics for classification accuracy and error are proposed based on the concept of effective sample size. Versions of perceptron training and KNN classification are built on these metrics. Procedures for regularised PGLS regression, phylogenetic KNN regression, neural network regression and regression trees are presented. Properties of phylogenetic perceptron training and KNN regression are studied with synthetic data. Experiments demonstrate that phylogenetic perceptron training improves robustness when the phylogeny is unbalanced. Regularised PGLS and KNN regression are applied to mammal dental traits and environments to both test the algorithms and gain insights in the relationship of mammal teeth and the environment.

ACM Computing Classification System (CCS):
CCS → Computing methodologies → Machine learning → Machine learning algorithms CCS → Applied computing → Life and medical sciences → Computational biology

Avainsanat — Nyckelord — Keywords

machine learning, phylogenetic comparative methods, mammal, dental draits, bioclimatic variables

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1. Introduction

In supervised machine learning, it is usually assumed that data is sampled independently from one distribution. In other words, the feature values of instances in learning and test sets are assumed to be independent and identically distributed (i.i.d.). Theory of regression and classification is well developed for i.i.d. data [76]. The topic of this research is supervised machine learning on data which contains hierarchical dependencies. These kind of statistical structures emerge in biological comparative data which contains features of different species or other mutually related taxonomical units. We propose new ways to conduct regression and classification on hierarchically dependent comparative data, and apply them to mammal biology and ecology.

Biological species are related through their evolutionary history, so their features are correlated through evolution [18]. The hierarchical dependencies of traits can also be called phylogenetic dependencies. The ancestry relationships of some set of species can be described with phylogenetic trees. Various approaches to reconstructing ancestry relationships have been proposed since Darwin [60]. Phylogenies are constructed based on phenotypic or genetic data, using probabilistic or parsimony inference [54]. In maximum likelihood approach, one constructs trees that maximise the likelihood of some statistical model. Maximum parsimony refers to building trees that minimise the evolutionary changes in the branches of the tree [54]. Several phylogenetic trees can be combined to form a supertree of a large set of species, like supertree of 5020 mammals [6, 27]. As the trees are estimates of evolutionary history, they contain uncertainty about their topology and branch lengths [22].

Procedures that take into account phylogenetic dependencies are called phylogenetic comparative methods. These methods answer questions of evolutionary associations or strength of phylogenetic dependencies in traits of species. Compared to ordinary methods that ignore phylogeny, phylogenetic comparative methods reduce model variance, and thus produce more reliable models on phylogenetically dependent data [69]. Properties of different phylogenetic methods are often proved with simulation studies, because formal proofs of asymptotic model properties with phylogenetic data include additional complexity compared to i.i.d. data [4, 44]. Evolutionary models like Brownian motion [18] or Ornstein-Uhlenbeck models [49, 8] provide mappings from phylogenetic trees to instance-

instance covariances in features. Measures of phylogenetic signal, that is dependence, are often closely related to evolutionary models. These include branch length transformations on phylogenies [54, 25], Ornstein-Uhlenbeck parameters [8] and D-statistic [28]. Regression methods like independent contrasts [18], phylogenetic generalised least squares [34] and phylogenetic eigenvector regression [15] are often variants of linear regression. Other regression methods include phylogenetic logistic regression [44] and phylogenetically weighted regression [12]. A phylogenetic variant of principal component analysis can be used as part of regression models [62].

Westoby [78] criticises phylogenetic methods for attributing all trait similarity between related species to phylogeny, when some of that similarity could be attributed to shared ecology. Traits in different species are correlated with both phylogeny and ecology, as relative species tend to inhabit similar ecological niches. Principle of phylogenetic niche conservatism states that the traits of an ancestor make it fit for its specific habitat, and thus its descendants will mostly thrive in similar habitats. It is not plausible to assume that trait similarity between related species is due to inertia in evolutionary timescales of millions of years, because of continuous selection pressure affects the properties of species.

Most animals cut and chew food with their teeth, so teeth function as an interface for energy intake. Efficient nutrition acquisition is an integral part of survival and fitness, so selection pressure demands that teeth work efficiently. The plant-based diet of herbivorous mammals places special demands on the functionality and durability of teeth [48, 30]. Availability of plant material for the consumption of herbivores is partly determined by climate, which connects the dental traits of herbivorous animals to climatic features. The closeness of form and function in teeth and feeding being a direct interaction with environment makes dental traits ecomorphological features [17]. The functional connection of herbivore teeth with climate suggests that dental traits can be used as proxies for estimating climate conditions [17].

The relationships of animal community features and environmental conditions have been studied in ecometric literature [1, 17, 16, 23, 30, 48, 80]. Mammal dental traits are connected to the dietary environment of the species. In herbivorous mammals, high crowned teeth are an evolutionary response to diet that demands durability and tolerance for wear [16]. Average crown height in communities tends to increase with decreasing annual precipitation [16] or net primary productivity [48]. Cutting capacity is seen as a response to annual primary productivity, so the relationship of cutting features and bioclimatic variables would estimate primary productivity [23]. Of environmental features in general, dental traits should have the clearest signal to those that describe the diet of herbivorous mammals in some way.

In addition to teeth, environment can affect mammalian body size. Aava [1] proposed that mammalian body size should grow with increasing primary productivity, and

showed that this relationship holds in Australian communities. Hypothesis of the study was that energy availability limits the size of animals in a given environment. As primary productivity measures the plant matter that is available for herbivores to consume, the body mass of herbivorous mammals should increase with primary productivity. The means of body mass distributions were found to increase, and standard deviations to decrease with primary productivity in Australian communities.

Primary productivity tends to decrease with decreasing temperature, and climate in extreme latitudes is cold. Bergmann's law [10] is a biophysical hypothesis on the relationship of temperature and body size. Anatomy of an animal determines the ways of heat exchange with the environment. Because the ratio of body volume to surface area increases with increasing size, larger animals should be more efficient in conserving heat and surviving in cold climates. Due to anatomical differences, this relationship should be evident in anatomically similar animals, that is individuals of the same species, or closely related species. Clauss et al. [10] found that there is a positive correlation between mammalian body mass and mean or maximum latitude, when phylogeny is taken into account.

Ecometric approach has been described as taxon-free. It is based on the idea that communities in geographical locations form based on environmental conditions [17]. Animals change location and spread to new ones practically instantly, so communities are formed so fast that community structures do not show evolutionary inertia. In this analysis, the relationship of dental traits and environment is studied in mammal species rather than communities. Analysing this relationship on species level makes phylogeny relevant as both dental traits and environmental features can be phylogenetically dependent.

Freckleton [24] recommends against reporting results from both phylogenetic and phylogenetically independent models, because the two types of models are based on different assumptions about the data and model residuals. In this study we are interested in the properties of the models and effects of phylogeny, so we report both if there is an interesting difference between them. The nature of difference between phylogenetic and ordinary models can give insights into the structure of the data and provide a deeper understanding of the situation than observing only one model.

In this study, we develop evaluation procedures and machine learning methods that take into account phylogenetic information. For evaluation of phylogenetic regression models, error metrics exist, so we only consider cross-validation. For classification tasks, we propose a phylogenetic evaluation framework based on effective sample size. As new methods are concerned, we show how to reduce regularised phylogenetic generalised least squares to ordinary regularised linear regression. We touch briefly on principal component regression and regression trees. Phylogenetic versions of k-nearest neighbour regression and classification are proposed, as well as phylogenetic perceptron training and neural

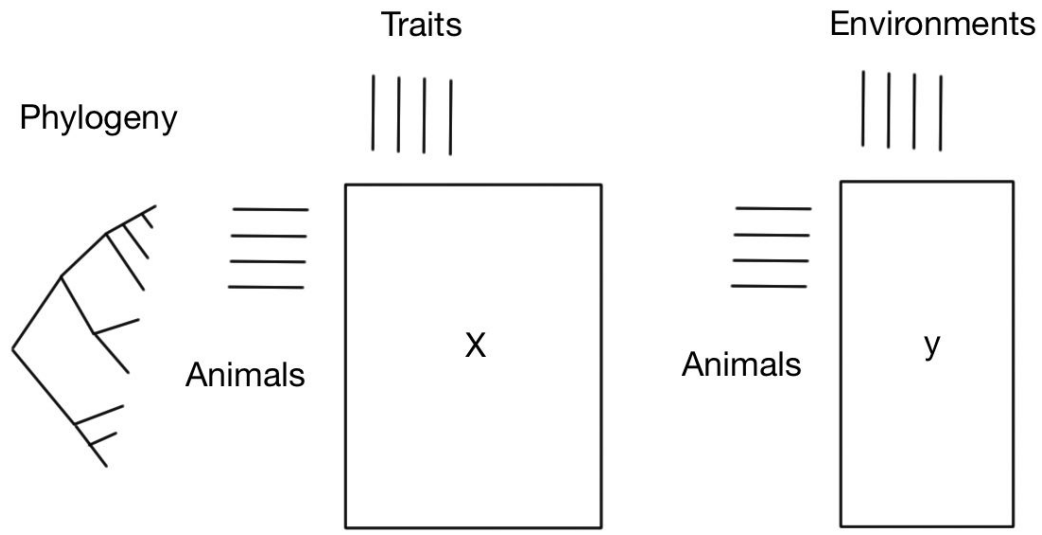

Figure 1.1: Scheme of the dataset. Traits **X** are input variables, and environments **y** output variables. Each row of input and output features are measurements or estimates of traits and environments of different animal species. As biological species are related through evolution, a phylogeny connects the species causing hierarchical statistical dependencies.

network regression. We apply some of these methods to real data on mammal dental traits and environments. General scheme of this comparative dataset is presented in Figure 1.1. We also test k-nearest neighbour regression and perceptron training on synthetic data.

2. Related Work

Assumptions of independence and identical distribution of instances can be violated in several ways. In this section we go through different settings of non-i.i.d. data in predictive modelling, and theory and methods associated with them. The key topics are population drift, domain adaptation, concept drift, temporal and spatial autocorrelation, transfer learning and phylogenetic comparative methods.

Hand [35] analysed what kind of effects method complexity and sampling of training data have on classifier performance. In many practical situations, if a model is trained with one data and applied to some future data, these two sets of data are drawn from different distributions. Population drift means that even if the training instances were drawn independently from identical distributions, the data that the model should predict is differently distributed. Changing environment can result to population drift.

A machine learning problem closely related to population drift is domain adaptation. If training and testing data are drawn from different distributions, the model needs to adapt for this change. A data environment which provides labelled training data is called source domain. Target domain is the environment which provides possibly unlabelled testing data [45]. Formally, a domain $\mathcal{D}$ is a pair of feature space $\mathcal{X}$ of all possible values of observation $\mathbf{x}$ and marginal probability distribution $P(\mathbf{x})$, $\mathcal{D} = \{\mathcal{X}, P(\mathbf{x})\}$ [56]. Training data can be combined from several separate source domains, and target domain can be modelled as a combination of source domains [41].

Concept drift refers to changes in conditional distribution $p(y|\mathbf{x})$ of the output variable y and input variables $\mathbf{x}$ [32]. In a concept drift situation, instances in a time series are not identically distributed. A change from one distribution to another can happen abruptly or incrementally. The change can also be gradual so that both the original and the new distribution are present in some time interval, or the original distribution can recur after a time.

Machine learning models that operate in changing environments need to detect concept drift and differentiate between concept drift and noise. They are also required to adapt to new distributions. These models are structured as online learning models, which produce predictions and make updates as new instances arrive. Concept drift sets demands for memory management, change detection, learning and loss estimation [32].

Online model operation in a changing environment requires receiving new, and forgetting old data. Data can function as a short term memory of the observed phenomenon, whereas model provides a long term memory. One way to manage data storage in a changing environment is to store a finite number of recent instances. The amount of stored instances can range from one to several, using a sliding window in time. Old data can be omitted from the memory, or weighted with a decreasing function of time [32].

A retraining method builds a new model if a change occurs. Another approach is incremental learning, in which model is updated in case of change. Approach to adaptation can be blind or informed. Blind adaptation methods do not monitor change, whereas informed adaptation methods include diagnostics about concept drift. A model can be replaced or updated globally or locally [32].

Data streams from environmental or other sensors often contain temporal dependence [79]. Time series data with temporal autocorrelation is not independent, and possibly not even identically distributed at different moments in time. Žliobaitė et al. [79] presented two predictive modelling approaches for time series data, temporal correction classifier and temporally augmented classifier. The former assumes first order temporal dependence, that is dependence between observation y_t and previous observation y_{t-1}. The latter is a heuristic approach to include higher order temporal dependency in predictive models. Temporally augmented classifier is a model which at time t predicts label y_t using input features $\mathbf{x}_t$ augmented with previous labels $y_{t-1}, y_{t-2}..., y_{t-l}$ for some l: $\hat{y}_t = f(\mathbf{x}_t, y_{t-1}, ..., y_{t-l})$.

For i.i.d. data, random classifier and majority class classifier are baseline methods for performance evaluation. Random classifier is a baseline for situation where information about data distribution is not available. The model predicts a class uniformly at random. If prior probabilities of the classes are available, it is possible to use majority classifier, which outputs the class with the maximum prior probability as prediction. If the data shows temporal autocorrelation, a temporally local persistent classifier performs better than the majority class classifier [79]. Persistent classifier outputs the previously observed class y_{t-1} as prediction for y_t. In general, properties of the baseline classifiers depend on the properties of the data stream.

Evaluation statistics developed for i.i.d. data can give misleading results if temporal autocorrelation is present. Kappa statistic for classification of imbalanced data can overestimate accuracy for temporally autocorrelated data streams. Žliobaitė et al. [79] formulated evaluation metrics suitable for predictive models on these data streams. In addition to evaluation statistics, temporal autocorrelation can also break assumptions of concept drift detection algorithms [79].

Like domain adaptation, transfer learning uses information from source domain $\mathcal{D}_S$ to model target domain $\mathcal{D}_T$. Need for transfer learning can arise in case if data distribution

changes so that training data and test data are not identically distributed, like in concept drift. However, transfer learning is its own subfield of machine learning, where learning of new tasks is augmented with information from previous tasks. Knowledge transfer can have positive or negative effect on learning of the target task, and naturally transfer with negative results should be avoided [56]. An example of transfer learning in humans is knowledge transform across different languages: knowing French makes learning Spanish easier, because knowledge of French grammar and vocabulary can be applied for Spanish with some modifications.

Learning task $\mathcal{T}$ is pair of output label space $\mathcal{Y}$ and conditional probability distribution $P(y|\mathbf{x})$, $\mathcal{T} = \{\mathcal{Y}, P(y|\mathbf{x})\}$. In transfer learning, source and target domains can be different, or learning tasks in the domains can be different: $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$ [56]. If some relationship connects source and target feature spaces $\mathcal{X}_S$ and $\mathcal{X}_T$, the domains $\mathcal{D}_S$ and $\mathcal{D}_T$ are related.

Transfer learning can be categorised to inductive, transductive and unsupervised learning. In inductive transfer learning, source and target tasks are different. In this case, some labelled data needs to be available in the target domain. In transductive transfer learning, source and target domains are different, but source and target tasks are the same. It is assumed that labelled data is available in the source domain, but not in target domain. Third type of transfer learning, that is unsupervised transfer learning, concerns situations where label spaces $\mathcal{Y}_S$ and $\mathcal{Y}_T$ are not observable [56]. Transferred knowledge can concern instances, feature representation, parameters or relational knowledge. Each of these types of transfer takes different forms in inductive, transductive and unsupervised transfer learning.

Geospatial data and spatial patterns are one area where assumption of independent samples is broken. Gradual variation in quantities (e.g. temperature) and geographical clustering of similar things (e.g. people with similar income level) causes spatial autocorrelation between samples [72]. The strength of this autocorrelation depends on spatial relation of the instances. The spatial relation of two geographic units can be their distance or some other connection, and this information is represented in an instance-instance contiguity matrix [72]. Contiguity matrices are analogous to phylogenies and phylogenetic covariance matrices. Main tasks in analysis of spatial data are outlier detection, co-location patterns, spatial regression and classification, spatial clustering and hotspot analysis [72].

So far we have considered machine learning in situations where i.i.d. assumptions break in some other way than with a shared phylogeny. Phylogenetic comparative methods can be divided into analysis of trait evolution and analysis of lineage diversification [59]. The subfield most relevant to this study is analysis of trait evolution. These methods can concern evolutionary correlations and models of trait evolution. Trait correlations

are analysed with different kinds of regression methods. Evolutionary models describe evolution of measured traits, and they can be used to map phylogenies to instance-instance covariances.

Phylogenetically independent contrasts (IC) [18] is a linear regression method for continuous traits. The method assumes a Brownian motion model of trait evolution, and it outputs a univariate linear regression slope. Phylogenetic generalised least squares (PGLS) [34] is a more general linear regression method, which uses the generalised least squares (GLS) framework. Model residuals are assumed to be normally distributed with phylogenetic instance-instance covariance. The difference to ordinary least squares (OLS) is that in OLS the residuals are independent. IC regression estimate is equivalent to PGLS slope with Brownian covariance, so independent contrasts are a special case of PGLS [7]. As with OLS, output feature of PGLS is assumed to be continuous, but input features can be continuous or discrete.

PGLS-models can in include other evolutionary models than Brownian motion. Branch length transformations like Pagel's λ [54, 25] are often applied evolutionary models. If the used mapping from phylogeny to covariance is a complex one, the resulting regression can look very different from the simplest PGLS-models. Ornstein-Uhlenbeck models are a family of evolutionary models which contain both random walk and adaptation components [36]. These evolutionary models and regressions built with them vary most on how adaptation is modelled. To give an example, regression with adaptation to randomly changing environment is presented in [37]. If parameters of evolutionary models are estimated with regression coefficients, closed form generalised least squares estimator does not provide the correct solution. Instead parameters are estimated with maximum multivariate Gaussian likelihood, like in [25].

Phylogenetically weighted regression (PWR) [12] is a data exploration framework that allows evolutionary correlations to vary in different parts of the studied phylogeny. Instead of a global regression model produced by PGLS, PWR builds a separate model for each studied species. The method is a weighted least squares regression, where the weights are phylogenetic distances from the studied species, or some other mappings of said distances.

Rolshausen et al. [70] explored temporal properties of trait diversification in a phylogeny with a quantity called trait space saturation. This analysis is based on trait distances and phylogenetic distances, standardised to interval $[0, 1]$ for all pairs of species in a phylogeny. If fraction a of the relative trait distances is filled in relative time interval θ_a, then time θ_a corresponds to trait space saturation level a. Brownian motion, Ornstein-Uhlenbeck processes and early burst processes all have their characteristic trait space saturation behaviour, so this analysis can measure phylogenetic signal and identify an evolutionary process behind a measured trait. Evolution of continuous traits can be

studied with this analysis without maximum likelihood or other statistical fitting methodology.

Frequentist methods do not usually incorporate uncertainty in phylogeny or other measurements and estimates. Different sources of uncertainty can be included with Bayesian statistics. A Bayesian formulation of PGLS is presented by de Villemereuil et al. [13]. The difference between the method of de Villemereuil et al. and Bayesian ordinary least squares regression is that variance-covariance matrix in the model likelihood is not proportional to identity matrix, but phylogenetic variance-covariance matrix. The prior distribution for the covariance matrix is defined using existing Bayesian methods for phylogeny construction. A more complex Bayesian phylogenetic regression with less restrictive assumptions is presented in [29].

Phylogenetic factor analysis combines phylogenetic information with Bayesian factor analysis to reduce data dimensionality [75]. The method of Tolkoff et al. [75] assumes that measured traits arise from a smaller number of factors, which are traits themselves. The factors develop as Brownian motion stochastic processes [18], but the method can be extended to other evolutionary models like Orntein-Uhlenbeck processes [8].

Variation in a continuous trait can be divided to phylogenetic and specific component with phylogenetic eigenvector regression (PVR) [15]. In this method, a phylogenetic distances of the studied species are collected to a distance matrix, and principal coordinate analysis is performed on this matrix. This results in phylogenetic eigenvectors and eigenvalues that serve as input features in a linear regression. The output feature of this regression is an observed continuous trait. Model predictions of PVR are the phylogenetic component of trait variation, and model residuals are the specific component.

Ives and Garland mention three different methods for phylogenetic regression on binary output features in [43]: logistic regression of the same authors, frequentist generalised linear mixed models and Bayesian generalised linear mixed models. Without input features, these methods serve as measurement procedures for phylogenetic signal [43]. Indications of phylogenetic signal are often connected to particular models of trait evolution explicitly or implicitly. Models for evolution of continuous traits are not appropriate for binary traits, so binary features require both separate models and phylogenetic signal measurement procedures. In addition to phylogenetic signal statistics mentioned by Ives and Garland, D-statistic is another measure of phylogenetic signal for binary traits [28].

Pagel [53] presented an evolutionary model for binary traits based on a continuous time Markov process. Based on this model, Pagel and Meade [55] studied correlated evolution of binary traits with Markov Chain Monte Carlo (MCMC) inference. In addition to Markov processes, discrete traits can be modelled with underlying continuous traits and thresholds. Felsenstein [19] presents a model of binary traits based on an underlying Brownian trait. If the continuous trait has a value greater than a threshold, the binary

feature is assigned value 1, and 0 otherwise. In this model, likelihood of observed data is based on the multivariate normal distribution, and it can be estimated with MCMC techniques. This method can estimate phylogenetically informed correlations between binary traits. In [20], Felsenstein expanded the method to include both discrete and continuous traits, and proposed a different MCMC sampling procedure than in the earlier article.

Ives and Garland [44] developed a phylogenetic logistic regression method and phylogenetic signal measure for binary traits. The method is based on Markov process model with two states and two transition probabilities. The covariance structure that this model produces for the binary trait resembles that of Ornstein-Uhlenbeck covariances for continuous traits [36]. When used with one or more input features, training phylogenetic logistic regression consists of estimating an overall state transition rate which measures phylogenetic signal, and regression coefficients for input features. The parameters are estimated by maximising (and penalising) a quasi-likelihood expression which resembles multivariate Gaussian likelihood [44].

Separate methods have been developed for modelling the evolution of multiple traits or multiple output features. In this area of study, covariation of traits and high dimensionality result in additional issues compared to univariate phylogenetic comparative methods. Adams and Collyer go through different approaches for multivariate phylogenetic comparative methods in [2]. These methods mainly concern fitting evolutionary models to multivariate data.

For estimation of evolutionary model parameters with multivariate data, one can compute model log-likelihoods and conduct tests with them like Revell and Harmon did with continuous traits in [65]. Different surrogate log-likelihoods have been proposed for situations in which direct estimation of model log-likelihood is not feasible. An example of these methods is SURFACE-algorithm, which fits univariate evolutionary models to different dimensions of the data, and summarises them to select the best fitting evolutionary model [42].

Phylogenetic principal component analysis (PCA) [62] can be used to reduce dimensionality of a phylogenetic dataset. The difference between this and ordinary principal component analysis is that the utilised version of feature-feature covariance matrix contains phylogenetic information. This matrix is called evolutionary variance-covariance matrix, and it is employed also by other methods like phylogenetic partial least squares [3]. Both phylogenetic PCA and partial least squares were presented using Brownian motion evolutionary model in [62] and [3], but they can be used with other evolutionary models.

Canonical correlation analysis is a dimensionality reduction method for multiple output regression. Its goal is to find pairs of transformed input and output features

with maximum pairwise correlations [26]. Phylogenetic canonical correlation analysis [66] is a phylogenetic version of this method. It was also introduced with Brownian motion evolutionary model, but is not limited to it. Combining Pagel's λ-transformation [54, 25] with canonical correlation analysis lead to a multivariate generalisation of λ-transformation [66].

Clavel et al. [11] presented a penalised likelihood framework for estimating parameters of multivariate evolutionary models. In univariate setting, parameter estimation and model comparisons of evolutionary models is traditionally done with Gaussian maximum likelihood estimation. If the number of traits approaches or exceeds the number of instances, maximum likelihood approach no longer works. Clavel et al. formulate regularised matrix normal objectives for evolutionary variance-covariance matrices with different kinds of ridge- and LASSO-regularisations. Additionally, they present a leave-one-out cross-validation procedure based on matrix normal likelihood [11].

In addition to the assumptions of statistical models, dependencies in data affect error estimation and model selection. Roberts et al. [67] considered cross-validation with non-i.i.d. data in ecological context. Standard validation procedures like k-fold cross-validation assume that the data is independent, which is not the case with phylogenetic data, spatially autocorrelated data or time series data. If validation data is dependent on training data, models appear more robust than they are. Roberts et al. established that strategic, non-random data splits of dependent data to training and validation sets resulted in validation errors that were closer to true error than with random splits. When data had spatial, temporal or phylogenetic dependence, random splits underestimated model error. Non-random splits were done by forming data blocks using the dependency structures, which are phylogenies in the case of comparative species data.

3. Problem Formulation

In this chapter we define supervised machine learning, formulate the main research question and explain the theoretical background of the work. We explain evolutionary models as tools to model statistical dependencies between species. A generalisation of squared error loss function for phylogenetic regression will be introduced.

3.1 Supervised Machine Learning

Supervised machine learning is a prediction problem, where the goal is to predict value of output variable y based on input variables $\mathbf{x} \in \mathbb{R}^p$. In regression tasks $y \in \mathbb{R}$ or some subset of $\mathbb{R}$. In classification tasks y is categorical. In this study, we mostly consider binary classification $y \in \{0, 1\}$. Output variable y is modelled as a function of $\mathbf{x}$ and parameters θ. Value of y is also affected by random error ε:

$$y = f(\mathbf{x}; \theta) + \varepsilon$$

Optimal parameter values $\hat{\theta}$ are estimated from the data. Prediction $\hat{y}$ is a function of $\mathbf{x}$ and parameter estimates $\hat{\theta}$:

$$\hat{y} = f(\mathbf{x}; \hat{\theta})$$

An instance is a pair of values $(\mathbf{x}, y)$. Dataset is a collection of n instances, and we notate a set of input variables as $\mathbf{X} \in \mathbb{R}^{n \times p}$. Often dataset is divided randomly into training, validation and testing data, but this division is based on i.i.d. assumption, so we regard all data as training data. In addition to a singe value, we can make a prediction for the whole dataset:

$$\hat{\mathbf{y}} = f(\mathbf{X}; \hat{\theta}) \in \mathbb{R}^n$$

In addition to predicted values, we are also interested how $f(\mathbf{x}; \hat{\theta})$ can increase our knowledge of the relationship between input and output variables, and thus develop understanding of the underlying phenomenon.

3.2 Research Question

How to formulate and validate supervised machine learning models using samples that are statistically dependent in a hierarchical way? Training set consists of input and output features, and phylogeny of all training instances. Using evolutionary models, the phylogeny can be mapped to estimates of statistical dependency between the training instances.

3.3 Models of Trait Evolution

Species that recently shared a common ancestor tend to resemble each other by appearance, behaviour and physiology. Phylogeny is a description of evolutionary ancestry relationships of a set of species or other taxonomical units. It is often presented in a tree format, like the phylogeny of 10 extant mammals in Figure 3.1. Each node in the tree represents a common ancestor of its child species.

When data is sampled from measurements on species or other hierarchically related units (e.g. families, individuals, languages, cultures), the samples are not drawn independently from the same distribution [18]. Most standard statistical and machine learning procedures assume that data is independent and identically distributed. Analysis of phylogenetic data requires thus methods that take into account phylogenetic correlation.

Traits are properties of animals (e.g. body mass), their behaviour (e.g. does the species display parental care) or their environment (e.g. temperature of the habitat). If phylogenetic correlation is present in a trait, we say that there is phylogenetic signal in the trait. Phylogenetic correlation (or signal) can also be present in a set of multiple traits or residuals of a statistical model.

Process that generates the data is a central concept in any statistical modelling. In the analysis of phylogenetic data, we want to formulate explicit models for these processes so that we can include assumptions on the statistical dependence between instances. Evolution of continuously valued traits can be modelled with a random walk, where change of a trait y in a species in time Δt is proportional to the time [25]:

$$\Delta y \sim N(0, \Delta t \sigma^2)$$

This simple model is called Brownian motion model of trait evolution. Here we assume that the rate of evolution σ^2 is same for all the species. We also assume that the phylogenetic tree is ultrametric, so all the leaves of the tree are at the same distance T from the root. Biological interpretation of this assumption is that the tree depicts ancestry relations of extant species. Without loss of generality, we can choose relative time units in which $T = 1$. Speciation is modelled as bifurcation where the paths of two species i

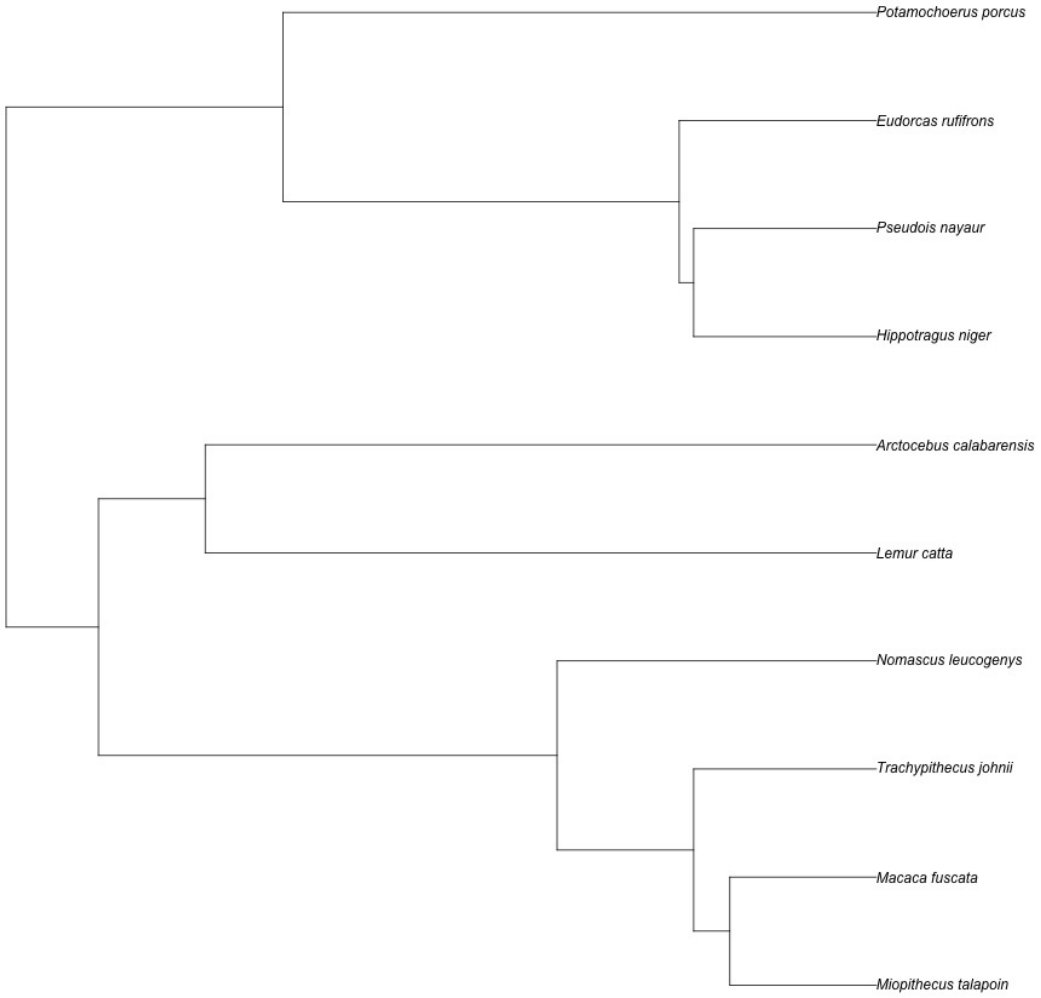

Figure 3.1: A phylogenetic tree of 10 extant mammals. The tree is constructed as a subtree of large mammalian phylogeny of 5020 species [27]. In this representation, the horizontal axis represents time from an inferred common ancestor on the left, to present day mammals on the right. Time units are arbitrary, and in this study we assume relative time from 0 to 1. The vertical axis contains no information about evolutionary history.

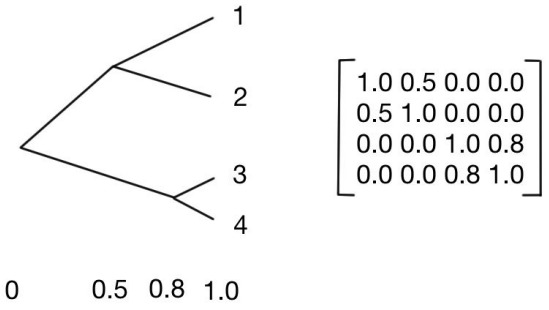

Figure 3.2: A phylogeny of four species with a corresponding Brownian covariance matrix. Time flows from zero to 1.0. $\mathrm{Cov}(i, j) = t_a\sigma^2$, where t_a is the last time when i and j shared a common ancestor, and parameter σ has been chosen $\sigma = 1$.

and j diverge. In this model, rate of trait change and speciation are independent of each other. Covariance of trait values of species i and j is [25]:

$$\mathrm{Cov}(y_i, y_j) = t_a\sigma^2$$

where t_a is the latest time when i and j shared a common ancestor. These covariances form the elements of Brownian species-species covariance matrix $\sigma^2\mathbf{C} \in \mathbb{R}^{n\times n}$, where the elements are $\mathbf{C}_{ij} = t_a$. An example phylogeny and corresponding Brownian covariance matrix are presented in Figure 3.2.

If trait y changes so quickly that the change seems instantaneous in the time scale of ancestry relations, then phylogenetic correlation is not present and $\mathrm{Cov}(y_i, y_j) = 0$ if $i \neq j$. In this case we say that y shows no phylogenetic signal. Strength of phylogenetic signal in continuous traits can be measured with Pagel's λ, which is a multiplier of non-diagonal elements of $\mathbf{C}$ [54, 25]. Usually it is assumed that $0 \leq \lambda \leq 1$.

Other models of trait evolution for continuous traits include Ornstein-Uhlenbeck [36, 8] and early burst-models [38]. In Ornstein-Uhlenbeck models, traits experience random variation in time like in Brownian motion model, but are simultaneously pulled towards optima. The trait optima can themselves vary as random walk, or remain stable. In early burst models, most trait variation in a phylogeny emerges early in diversification, and evolution slows down as time progresses [38].

Binary traits can be modelled as Markov processes [53, 55, 44], or with underlying continuous traits [19, 20]. Based on the latter approach, D-statistic [28] is a measure of phylogenetic signal in binary traits. Value $D = 1$ indicates no phylogenetic signal, and $D = 0$ indicates signal which is comparable to Brownian motion. Possible values of D-statistic are not constrained to this interval.

Ives et al. [44] presented a covariance model for binary traits with elements

$$\mathrm{Cov}(y_i, y_j) = e^{-\alpha d_{ij}}$$

where d_{ij} is phylogenetic distance between species i and j, and parameter α is estimated from data. In this model, $\alpha \approx 1$ corresponds to phylogenetic signal of similar strength with Brownian motion. Parameter α can measure phylogenetic signal in a binary trait, or residuals of a logistic regression model.

3.4 Phylogenetic Loss Function for Regression

Our sample contains n species or other taxonomic units. We have measurements of p covariates x_j, $j = 1...p$ and of target variable y. Vectors of covariates form a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. We can assume that the first column of $\mathbf{X}$ is a vector of ones $\mathbf{1} \in \{1\}^n$, so that we do not need to add intercept terms to our models. Model f determines how trait y depends on input $\mathbf{X}$ and parameters β:

$$\mathbf{y} = f(\mathbf{X}, \beta) + \varepsilon$$

Residuals ε are normally distributed with mean zero and covariance $\sigma^2\mathbf{C}$: $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{C})$. If $\mathbf{C}$ is identity matrix, the data is assumed to be mutually independent. In linear model $f(\mathbf{X}, \beta) = \mathbf{X}\beta$, $\beta \in \mathbb{R}^p$, the resulting regression is ordinary least squares (OLS). If $\mathbf{C}$ in a linear model is a phylogenetic covariance matrix, the regression becomes phylogenetic generalised least squares (PGLS) [34].

The probability distribution function for $\mathbf{y}$ is [33]:

$$p(\mathbf{y}; f(\mathbf{X}, \beta), \sigma^2\mathbf{C}) = \frac{1}{\sqrt{(2\sigma\pi)^n \det(\mathbf{C})}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-f(\mathbf{X},\beta))^T\mathbf{C}^{-1}(\mathbf{y}-f(\mathbf{X},\beta))}$$

The loss function for generalised least squares can be derived from cross-entropy [33] $H(P, N)$ of the process P that generates the observations $(\mathbf{X}, \mathbf{y})$ and the n-dimensional normal distribution N with probability distribution function p:

$$H(P, N) = -\log \mathbb{E}_{\mathbf{X},\mathbf{y}\sim P} p(\mathbf{y}; f(\mathbf{X}, \beta), \sigma^2\mathbf{C})$$

The resulting loss function equals negative logarithmic likelihood of the observations given the model:

$$L(\beta; \mathbf{X}, \mathbf{y}, \mathbf{C}) = \frac{n}{2}\log(2\sigma^2\pi) + \frac{1}{2}\log\det(\mathbf{C}) + \frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{X}, \beta))^T\mathbf{C}^{-1}(\mathbf{y} - f(\mathbf{X}, \beta))$$

In general, phylogenetic covariance $\mathbf{C}(\theta)$ can depend on some parameters θ, but now we assume that $\mathbf{C}$ is constant. If we seek an optimal estimator $\hat{\beta}$ for parameters β, it is sufficient to minimise the part that depends on β:

$$L(\beta; \mathbf{X}, \mathbf{y}, \mathbf{C}) = (\mathbf{y} - f(\mathbf{X}, \beta))^T\mathbf{C}^{-1}(\mathbf{y} - f(\mathbf{X}, \beta))$$

If $\mathbf{C}$ is identity matrix, $L(\beta)$ is equivalent to sum of square error loss (SSE).

In practical implementations we want to avoid computing $\mathbf{C}^{-1}$. The inverse covariance $\mathbf{C}^{-1}$ can be decomposed with root $\mathbf{C}^{-\frac{1}{2}}$ of $\mathbf{C}^{-1}$. The specific matrix root used in this study is Cholesky decomposition [50], $\mathbf{C} = \mathbf{L}\mathbf{L}^T$, $\mathbf{L} \in \mathbb{R}^{n \times n}$. The loss function becomes:

$$L(\beta) = (\mathbf{y} - f(\mathbf{X}, \beta))^T (\mathbf{L}\mathbf{L}^T)^{-1} (\mathbf{y} - f(\mathbf{X}, \beta)) = (\mathbf{y} - f(\mathbf{X}, \beta))^T \mathbf{L}^{-T} \mathbf{L}^{-1} (\mathbf{y} - f(\mathbf{X}, \beta))$$

$$= [\mathbf{L}^{-1}(\mathbf{y} - f(\mathbf{X}, \beta))]^T \mathbf{L}^{-1} (\mathbf{y} - f(\mathbf{X}, \beta)) = (\mathbf{L}^{-1}\mathbf{y} - \mathbf{L}^{-1} f(\mathbf{X}, \beta))^T (\mathbf{L}^{-1}\mathbf{y} - \mathbf{L}^{-1} f(\mathbf{X}, \beta))$$

In linear regression, this decomposition leads us to phylogenetic GLS-transformation [9]. The model is $f(\mathbf{X}, \beta) = \mathbf{X}\beta$, and the loss function gets the form:

$$L(\beta) = (\mathbf{L}^{-1}\mathbf{y} - \mathbf{L}^{-1}\mathbf{X}\beta)^T (\mathbf{L}^{-1}\mathbf{y} - \mathbf{L}^{-1}\mathbf{X}\beta)$$

We can write $\mathbf{U} = \mathbf{L}^{-1}\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Z} = \mathbf{L}^{-1}\mathbf{y} \in \mathbb{R}^n$. The covariance matrix that links $\mathbf{U}$ and $\mathbf{Z}\beta$ is identity:

$$L(\beta) = (\mathbf{Z} - \mathbf{U}\beta)^T (\mathbf{Z} - \mathbf{U}\beta) = (\mathbf{Z} - \mathbf{U}\beta)^T \mathbf{I}^{-1} (\mathbf{Z} - \mathbf{U}\beta)$$

This formulation of $L(\beta)$ shows that variables $\mathbf{U}$ and $\mathbf{Z}$ are independent, so they can be used in many standard procedures for independent data. It is good to note that a model built with $\mathbf{U}$ and $\mathbf{Z}$ does not contain an intercept. What was a homogenous column of ones in $\mathbf{X}$ is a non-homogenous column $\mathbf{L}^{-1}\mathbf{1}$ in $\mathbf{U}$.

In a model f, output for $\mathbf{U}$ is $f(\mathbf{U}, \beta) = f(\mathbf{L}^{-1}\mathbf{X}, \beta)$. The prediction terms in the loss function are in the form $\mathbf{L}^{-1} f(\mathbf{X}, \beta)$, so this gives a model-dependent condition for the GLS-transformation:

$$f(\mathbf{L}^{-1}\mathbf{X}, \beta) = \mathbf{L}^{-1} f(\mathbf{X}, \beta)$$

We can apply the phylogenetic GLS-transformation in linear models $f(\mathbf{X}, \beta) = \mathbf{X}\beta$.

A special case of (linear) regression is regression onto a constant β_0. In the i.i.d. case the solution is the arithmetic mean of $\mathbf{y}$: $\hat{\beta}_0 = \frac{1}{n} \sum_1^n y_i$. In the phylogenetic case, this regression takes a different form. We can write the loss function for this regression with $f(\mathbf{X}) = \beta_0 \mathbf{1}$:

$$L(\beta_0) = (\mathbf{y} - \beta_0 \mathbf{1})^T \mathbf{C}^{-1} (\mathbf{y} - \beta_0 \mathbf{1})$$

To solve the estimator $\hat{\beta}_0$, we take a partial derivative of L along β_0:

$$\frac{\partial}{\partial \beta_0} L(\beta_0) = 2(\mathbf{y} - \beta_0 \mathbf{1})^T \mathbf{C}^{-1} \cdot (-\mathbf{1})$$

When $\beta = \hat{\beta}_0$, the partial derivative is equal to zero:

$$2(\mathbf{y} - \hat{\beta}_0 \mathbf{1})^T \mathbf{C}^{-1} \cdot (-\mathbf{1}) = 0$$

Solving the equation, estimator $\hat{\beta}_0$ equals so called phylogenetic mean of $\mathbf{y}$ with covariance $\mathbf{C}$:

$$\hat{\beta}_0 = \frac{\mathbf{y}^T \mathbf{C}^{-1} \mathbf{1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}$$

An efficient computation procedure for phylogenetic mean is given in Algorithm 2. The quantity $\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}$ in the denominator was identified as effective sample size of the intercept by Ané [4]. Computation procedure for effective sample size is given in Algorithm 1. Note that $\mathbf{y}^T\mathbf{I}^{-1}\mathbf{1} = \sum_1^n y_i$, and $\mathbf{1}^T\mathbf{I}^{-1}\mathbf{1} = n$. This mean that if $\mathbf{C} = \mathbf{I}$, we get arithmetic mean as a special case of phylogenetic mean. Weighted mean is also a special case of phylogenetic mean. In the weighted case, $\mathbf{C}$ is a diagonal matrix with a non-uniform diagonal.

Effective sample size is a function of a covariance matrix $\mathbf{C}$. Covariance matrices describe dependencies in some feature or model residuals, so means of different variables have different effective sample sizes. In addition to evolutionary model, the value of effective sample size also depends on the subset of data where it is computed. We notate effective sample size of covariance $\mathbf{C}$ by $\text{Ess}(\mathbf{C})$, and effective sample size for a subset s of the data by $\text{Ess}(\mathbf{C}_s) = \text{Ess}\{s\}$.

Ané [4] showed that PGLS estimate for the intercept is not a consistent estimate, if the number of branches from the root of the phylogenetic tree is limited, and the length of the branches has a lower limit. These limitations apply to comparative datasets in practice, because all the species are sampled from branches that exist today. Slopes on input features, on the other hand, are estimated consistently in PGLS.

Algorithm 1: Effective sample size. Here $\mathbf{1}$ is an $n \times 1$ vector of ones and $\mathbf{z} \in \mathbb{R}^{n\times 1}$.

input : Covariance matrix $\mathbf{C} \in \mathbb{R}^{n\times n}$
output: Effective sample size of $\mathbf{C}$

$\mathbf{L} \leftarrow$ Lower triangular Cholesky decomposition of $\mathbf{C}$
$\mathbf{z} \leftarrow \mathbf{L}^{-1}\mathbf{1}$
return $\mathbf{z}^T\mathbf{z}$

Algorithm 2: Phylogenetic mean. Here $\mathbf{1}$ is an $n \times 1$ vector of ones and $\mathbf{z}, \mathbf{z}_1 \in \mathbb{R}^{n\times 1}$.

input : Data vector $\mathbf{x} \in \mathbb{R}^{n\times 1}$ and covariance matrix $\mathbf{C} \in \mathbb{R}^{n\times n}$
output: Phylogenetic mean of $\mathbf{x}$ given covariance $\mathbf{C}$

$\mathbf{L} \leftarrow$ Lower triangular Cholesky decomposition of $\mathbf{C}$
$\mathbf{z}_1 \leftarrow \mathbf{L}^{-1}\mathbf{1}$
$\mathbf{z} \leftarrow \mathbf{L}^{-1}\mathbf{x}$
return $(\mathbf{z}_1^T\mathbf{z}_1)^{-1}\mathbf{z}_1^T\mathbf{z}$

The loss function presented here is suitable for building global models that take into account phylogenetic correlation in the observed quantities. In this type of modelling,

we aim to make generalisations of the data generating process, and sacrifice local accuracy. Another approach would be to build locally accurate models using the information that phylogenetically close species resemble each other in trait values. Phylogenetically weighted regression of Davies et al. is an example of such a model [12]. This is analogous to using spatial locality information with geospatial data [72], or temporal locality information in data streams which contain temporal autocorrelation [79].

4. Validation procedures

Here we consider how to evaluate model performance taking into account hierarchical dependence of instances. We consider cross-validation of phylogenetic regression models based on generalised residual sum of squares. As classification is concerned, we focus on developing phylogenetic accuracy and error metrics. Using the concept of effective sample size, we propose phylogenetic generalisations of indicator loss, classification accuracy, error rate and majority voting.

4.1 Cross-validation of regression

Ordinary leave-one-out cross-validation objective for regression is the sum of squared prediction error $\sum_i (y_i - \hat{y}_{-i})^2$, where $\mathbf{x}_i$ has been left out of the data for the training of model f_{-i}, giving prediction $\hat{y}_{-i} = f_{-i}(\mathbf{x}_i)$. We recognise this as squared Euclidean distance of observation $\mathbf{y}$ and prediction vectors $\hat{\mathbf{y}}$:

$$\sum_i (y_i - \hat{y}_{-i})^2 = (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}})$$

The squared Euclidean distance is equivalent to squared Mahalanobis distance with identity covariance matrix. In cross-validation of phylogenetic models, we need to treat the prediction errors not as independent but dependent through the evolutionary process. When we replace the identity covariance with the phylogenetic covariance matrix $\mathbf{C}$, the generalised leave-one-out objective becomes

$$(\mathbf{y} - \hat{\mathbf{y}})^T \mathbf{C}^{-1} (\mathbf{y} - \hat{\mathbf{y}})$$

where $\hat{\mathbf{y}}_i$ is prediction for y_i from model f_{-i} trained with data $\mathbf{X}_{-i}$ that includes all other instances but $\mathbf{x}_i$. We assume that covariance $\mathbf{C}$ is constant in all n models. We call this cross-validation objective phylogenetic error. The objective is consistent with the assumption that prediction errors are phylogenetically dependent. Form of the objective is the same as generalised residual sum of squares $(\mathbf{y} - f(\mathbf{X}, \beta))^T \mathbf{C}^{-1} (\mathbf{y} - f(\mathbf{X}, \beta))$ presented by Martins and Hansen for an individual model f [49].

If more than one instance is left out in cross-validation steps, there is phylogenetic dependence in the training subset, validation subset and between subsets. Different subsets

contain varying phylogenetic dependence determined by their local phylogenetic structure. Due to this additional complexity, we only consider leave-one-out cross-validation in this study.

Validation procedures for phylogenetic models based on non-random sampling have been proposed in the literature. Roberts [67] suggested a blocking procedure that tries to select independent samples from the phylogenetic tree. Another approach for cross-validation of phylogenetic models is presented in Clavel et al. [11] in the context of multivariate evolutionary model fitting.

4.2 Indicator loss and classification accuracy

Here we propose phylogenetic generalisations of indicator loss, classification error and classification accuracy. We are not aware of any previously proposed classification accuracy metrics that would take into account phylogenetic information. In the method evaluation of phylogenetic logistic regression [44], Ives and Garland focus on statistical properties of logistic regression coefficients instead of accuracy metrics.

Phylogenetic loss function should depend on model prediction, target variable and phylogeny. In this study, covariance represents the phylogenetic dependence, but if the loss is based on a statistical distribution which does not have covariance, then other quantities are appropriate. Indicator loss I for i.i.d. models measures the number of wrongly classified instances [76]:

$$I(y, f(\mathbf{x})) = \begin{cases} 0 & \text{if } y = f(\mathbf{x}) \\ 1 & \text{otherwise} \end{cases}$$

Let us collect the indicator loss values of all n instances and predictions in vector $\mathbf{y}'$. This vector has elements $\mathbf{y}'_i = 0$ if $y_i = f(\mathbf{x}_i)$, and 1 otherwise for all $i = 1, ..., n$. With i.i.d. models, indicator loss for the whole data would be $I(\mathbf{y}, f(\mathbf{X})) = \mathbf{y}'^T\mathbf{y}' = \mathbf{y}'^T\mathbf{I}^{-1}\mathbf{y}'$. Let us assume that we have an appropriate covariance estimate $\mathbf{C}$ which describes dependence structure in these model errors. Now we can write a phylogenetic generalisation of indicator loss I_p with this matrix product by including phylogenetic covariance $\mathbf{C}$. Only the elements $\mathbf{y}'_i$ for which $y_i \neq \hat{y}_i$ are non-zero, so this is equivalent to the effective sample size of subset $\{y \neq \hat{y}\}$:

$$I_p(\mathbf{y}, \hat{\mathbf{y}}) = \mathbf{y}'^T\mathbf{C}^{-1}\mathbf{y}' = \text{Ess}\{y \neq \hat{y}\}$$

Because instances are mutually correlated, ratios of correctly or wrongly classified samples are not appropriate validation metrics for phylogenetic classifiers. A phylogenetic algorithm gives different importances for the data instances based on their phylogenetic distances in the training set. As minimising various loss functions for classification of i.i.d.

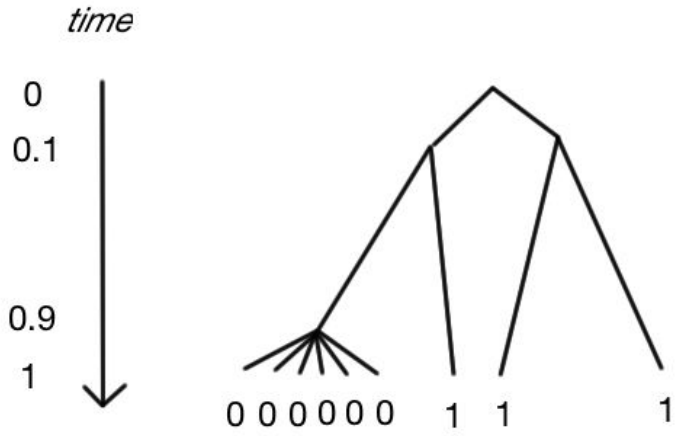

Figure 4.1: An example of classification problematics in phylogenetic data. In the phylogeny, each species has a binary trait value of either 0 or 1. Without any input variables and without the phylogeny, a baseline generalisation of this data of nine instances would be 0. With the phylogeny, intuitively better baseline generalisation is 1. Using effective sample size as a measure instead of instance counts gives 1 as the baseline generalisation: $\text{Ess}\{y = 0\} \approx 1.1$ and $\text{Ess}\{y = 1\} = 2.5$ with a Brownian covariance.

data indirectly minimise classification error, we suggest that phylogenetic classification loss functions like multivariate Gaussian loss or phylogenetic logistic regression loss [44] indirectly minimise wrongly classified effective sample size, and this is a desirable property.

To gain an intuitive understanding of this classification setup and evaluation procedure, consider phylogeny and binary trait $y \in \{0, 1\}$ distribution in Figure 4.1. Of the nine species, six has trait value $y = 0$, and three have trait value $y = 1$, so without phylogenetic information, baseline generalisation is 0. However, all the species with $y = 0$ are closely related, so this set of instances contains much less information than a set of six independent samples. Species in other parts of the tree have trait value $y = 1$, so 1 can be seen as better generalisation than 0. This reasoning can be quantified with effective sample size. The effective sample size of species with $y = 0$ is $\text{Ess}\{y = 0\} \approx 1.1$ using Brownian covariance matrix, and for species with $y = 1$ the effective sample size is $\text{Ess}\{y = 1\} = 2.5$. With this metric, better generalisation of the process is 1. In conclusion, using effective sample size as a metric, we can make make classification generalisations with phylogenetic information. Brownian covariance matrix is not statistically appropriate for binary traits [44], but it is very illustrative for this example.

In this framework, absolute number of correctly classified samples generalises into absolute correctly classified effective sample size:

$$\text{Ess}\{y = \hat{y}\}$$

Using effective sample size as a phylogenetic generalisation of number of instances, we

could define relative correctly classified effective sample size:

$$\frac{\text{Ess}\{y = \hat{y}\}}{\text{Ess}\{y\}}$$

This quantity measures how much information from a set is classified correctly. On the i.i.d. case the ratio becomes standard classifier accuracy:

$$\frac{\text{Ess}\{y = \hat{y}\}}{\text{Ess}\{y\}} \xrightarrow[\mathbf{C \to I}]{} \frac{\#\{y = \hat{y}\}}{\#\{y\}}$$

Similarly, we can define phylogenetic classification error rate with wrongly classified sample size:

$$\frac{\text{Ess}\{y \neq \hat{y}\}}{\text{Ess}\{y\}}$$

Again, on the i.i.d. limit this ratio becomes standard error rate:

$$\frac{\text{Ess}\{y \neq \hat{y}\}}{\text{Ess}\{y\}} \xrightarrow[\mathbf{C \to I}]{} \frac{\#\{y \neq \hat{y}\}}{\#\{y\}}$$

Effective sample size is a non-linear function, so unlike with standard accuracy and error rate, the sum of correctly and wrongly classified relative sample sizes is not necessarily equal to one. We can formulate an alternative quantity which sums up to one by replacing the denominator with $\text{Ess}\{y \neq \hat{y}\} + \text{Ess}\{y = \hat{y}\}$. We call this quantity *effective sample size accuracy* (EA):

$$\text{EA} = \frac{\text{Ess}\{y = \hat{y}\}}{\text{Ess}\{y = \hat{y}\} + \text{Ess}\{y \neq \hat{y}\}}$$

The same substitution of denominator can be done for the relative wrongly classified sample size. We call the resulting quantity *effective sample size error* (EE):

$$\text{EE} = \frac{\text{Ess}\{y \neq \hat{y}\}}{\text{Ess}\{y = \hat{y}\} + \text{Ess}\{y \neq \hat{y}\}}$$

These become standard classification accuracy and error rate on the i.i.d. limit.

Ordinary majority voting means taking majority output label c_{maj} from some set of instances: $c_{maj} = \arg\max_c \#\{y = c\}$. With the concepts presented above, we propose phylogenetic generalisation of majority vote as the label c_{max} that corresponds to maximum effective sample size:

$$c_{max} = \arg\max_c \text{Ess}\{y = c\}$$

This is a formal presentation of the reasoning that was applied to the data in Figure 4.1. Maximum effective sample size label for the data with given covariance was $c_{max} = 1$. In this example, we applied effective sample size-based classification framework as a baseline classifier without any input variables. In Chapter 5 we will present some predictive models in this framework that use both input and output variable information.

5. New Methods

Here we develop new computational procedures from the assumption that the data contains phylogenetic dependencies. The chosen methods represent varying types of supervised learning and operate with different principles: Regularised regression and principal component regression can be used for feature selection. Instance based regression makes local predictions using nearest neighbours. Regression trees are a greedy recursive method. Perceptron and general neural network regression represent neural computing.

5.1 Regularised regression

Predictive model f is a function of $\mathbf{X}$ and parameters β, $f = f(\mathbf{X}, \beta)$, and phylogenetic covariance is parametrised by θ, $\mathbf{C} = \mathbf{C}(\theta)$. A regularised loss function L based on Gaussian cross-entropy is:

$$L(\beta, \theta; \mathbf{X}, \mathbf{y}, \mathbf{C}) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\mathbf{C}(\theta))$$

$$+\frac{1}{2}(\mathbf{y} - f(\mathbf{X}, \beta))^T\mathbf{C}^{-1}(\theta)(\mathbf{y} - f(\mathbf{X}, \beta)) + \alpha_1\Omega_1(\beta) + \alpha_2\Omega_2(\theta)$$

where α_1 and α_2 are a regularisation parameters, and Ω_1 and Ω_2 functions of model and covariance parameters. We focus on regularised linear regression $f(\mathbf{X}, \beta) = \mathbf{X}\beta$. Here we do not consider simultaneous optimisation of β and θ, but limit the analysis on optimisation of model parameters. Regularisation procedures for phylogenetic covariance matrices $\mathbf{C}(\theta)$ are presented in [11]. Parameters θ have separate optimum values for different models f, but we approximate the resulting covariances with $\mathbf{C}(\theta)$ optimised for output variable $\mathbf{y}$. In other words, phylogenetic signal in regression residuals is approximated with phylogenetic signal in output variable. With these assumptions, our model is regularised PGLS, and loss function $L(\beta)$ takes the form:

$$L(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T\mathbf{C}^{-1}(\mathbf{y} - \mathbf{X}\beta) + \alpha\Omega(\beta)$$

Minimising $L(\beta)$ can be reduced to regularised OLS regression with a particular GLS-transformation.

Our procedure can be applied to any regularisation Ω, but we are most interested in LASSO regularisation because it produces sparse models. Ordinary LASSO regression is linear regression where the loss function $\sum_{i=0}^{n}(y_i - \beta_0 + \mathbf{x}_i\beta)^2$ is minimised on condition $\sum_{i=1}^{p}|\beta_i| \leq a$ for some $a \in \mathbb{R}$ [74]. This is practical for feature selection for OLS regression models. Current implementations of LASSO regression can allow an intercept that is not penalised [39]. Loss for ordinary LASSO can also be written in the same (Lagrangian) form as above, if we assume that there is no intercept in the model. Now regularisation function Ω is L_1-norm of regression coefficients, $\Omega(\beta) = ||\beta||_1$:

$$L(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \alpha||\beta||_1$$

Ordinary least squares regression in GLS-transformed variables $\mathbf{U}$ and $\mathbf{Z}$ does not contain intercept, because a possible intercept term vector $\mathbf{1}$ in $\mathbf{X}$ has been transformed with root of inverse into $\mathbf{L}^{-1}\mathbf{1}$ and is not uniform. In other words, regression in $\mathbf{U}$ and $\mathbf{Z}$ is regression through origin, so intercept must be omitted in phylogenetic regularised regression. There are two possible approaches for the slope of the transformed intercept variable $\mathbf{L}^{-1}\mathbf{1}$: either it must not be penalised to preserve the model correct, or variables must be centered so that the slope would be equal to zero. We follow the latter approach, which is more compatible with available implementations.

In the context of Brownian trait evolution, Rohlf [68] states that PGLS line goes through the point of ancestral state estimates or phylogenetic means [62] $\mu_x \in \mathbb{R}^p$ and $\mu_y \in \mathbb{R}$:

$$(\mu_x, \mu_y) = (\frac{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{X}}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}}, \frac{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{y}}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}})$$

This is true also in other models than Brownian motion, but then phylogenetic mean is not necessary interpretable as an ancestral state estimate. In order to modify a PGLS model so that the regression line goes through the origin, one needs to center the data with phylogenetic means μ_x and μ_y. Centering with phylogenetic mean is used also in evolutionary variance-covariance matrix (phylogenetic principal component analysis [62] and phylogenetic partial least squares [3]) and phylogenetic canonical correlation analysis [66].

If we center $\mathbf{X}$ and $\mathbf{y}$ with phylogenetic means and transform the centered variables with phylogenetic GLS transformation, the OLS line for the resulting independent variables has zero intercept coefficient $\hat{\beta}_0$ for the transformed dimension $\mathbf{L}^{-1}\mathbf{1}$. The centering GLS transformation is:

$$\mathbf{U}_0 = \mathbf{L}^{-1}(\mathbf{X} - \mathbf{1}\frac{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{X}}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}})$$

$$\mathbf{Z}_0 = \mathbf{L}^{-1}(\mathbf{y} - \mathbf{1}\frac{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{y}}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}})$$

Computing regularised PGLS regression with transformed variables $\mathbf{U}_0$ and $\mathbf{Z}_0$ is equivalent to regularised OLS with zero intercept. Computational procedure for centering GLS-transformation is presented in Algorithm 3, and regularised PGLS in Algorithm 4.

Algorithm 3: Centering GLS-transformation. Here $\mathbf{1}$ is an $n \times 1$ vector of ones, $\mathbf{X}_{mean} \in \mathbb{R}^{1\times p}$, $\mathbf{z}_1 \in \mathbb{R}^{n\times 1}$ and $\mathbf{z} \in \mathbb{R}^{n\times p}$.

input : Data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and covariance matrix $\mathbf{C} \in \mathbb{R}^{n\times n}$
output: Centered and GLS-transformed variable $\mathbf{U}_0 \in \mathbb{R}^{n\times p}$.

$\mathbf{L} \leftarrow$ Lower triangular Cholesky decomposition of $\mathbf{C}$
$\mathbf{z}_1 \leftarrow \mathbf{L}^{-1}\mathbf{1}$
$\mathbf{z} \leftarrow \mathbf{L}^{-1}\mathbf{X}$
$\mathbf{X}_{mean} \leftarrow (\mathbf{z}_1^T\mathbf{z}_1)^{-1}\mathbf{z}_1^T\mathbf{z}$
$\mathbf{U}_0 \leftarrow \mathbf{L}^{-1}(\mathbf{X} - \mathbf{1}\cdot\mathbf{X}_{mean})$
return $\mathbf{U}_0$

Algorithm 4: Regularised PGLS. Procedure for computing PGLS with LASSO, Ridge or other regularisation using available libraries for regularised OLS.

input : Data matrices $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{y} \in \mathbb{R}^{n\times 1}$ and covariance matrix $\mathbf{C} \in \mathbb{R}^{n\times n}$
output: Regularised PGLS regression

$\mathbf{U}_0 \leftarrow$ Center and GLS-tranform $\mathbf{X}$ given $\mathbf{C}$
$\mathbf{Z}_0 \leftarrow$ Center and GLS-tranform of $\mathbf{y}$ given $\mathbf{C}$
Call regularised OLS regression with input variables $\mathbf{U}_0$, output variable $\mathbf{Z}_0$ and no intercept

5.2 Principal component regression

Phylogeny can be taken into account in computing principal components of comparative data [62]. A related method is phylogenetic eigenvector regression [15, 14], which partitions trait variations into phylogenetic and taxon-specific components. Principal component analysis is often applied in data preprocessing to reduce input dimensionality for predictive models.

It is possible to choose principal components as input features for a regression model based on their ordering. However, there is no guarantee that the first principal components would be the ones with strongest signal to the output feature. That is why we recommend applying feature selection methods like LASSO regression on the principal components. Here we present a phylogenetic procedure for principal component regression:

1. Conduct phylogenetic principal component analysis for input variables
2. Compute data vectors from principal components
3. Perform centering GLS-transformation for the new data vectors and output variable
4. Compute regularised PGLS regression on transformed data to choose principal components for modelling
5. Build PGLS model on chosen principal components
6. Transform back to original input variables

5.3 Instance based methods

Ordinary k nearest neighbour (KNN) regression returns mean $\frac{1}{k}\sum_{i=1}^{k} y_i$ of output variable $\mathbf{y}$ in the set of k nearest neighbours of a point of query $\mathbf{x}' \in \mathbb{R}^p$ [26]. We assume that relation of $\mathbf{x}'$ with the phylogeny of the training set is unknown, $\mathbf{x}'$ is simply a point in the input feature space. In phylogenetic comparative data, the neighbours are not independent instances, so prediction with arithmetic mean can be biased by phylogeny. The set of nearest neighbours of point $\mathbf{x}'$ also contain varying amounts of information depending on their phylogeny. In the worst case, all k nearest instances in the training set are closely related and contain a lot less information than k independent instances would contain.

For phylogenetic version of this algorithm, we replace the output of arithmetic mean with phylogenetic mean, computed in some subset of the training data:

$$\hat{y}(\mathbf{x}') = \frac{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{y}}{\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}}$$

Here covariance $\mathbf{C}$ describes dependencies in the output variable $\mathbf{y}$. As mentioned before, the denominator was identified as effective sample size of intercept (or mean) [4]. To make sure that the instances selected for prediction form a sufficiently general subset of the data, we demand that the effective sample size of the nearest neighbours $\mathbf{1}^T\mathbf{C}^{-1}\mathbf{1} \geq k$ for some $k \geq 1$. If the phylogenetic covariance $\mathbf{C}$ is identity, $k \in \{1, 2, ..., n\}$ and we regain ordinary KNN regression. If some phylogenetic dependence is present, the effective sample size of the whole observation $\mathbf{y}$ is smaller than n, giving an upper limit for k. With this new meaning of hyperparameter k, we need to search a number of nearest neighbours that correspond to the given effective sample size threshold. Even if the evolution of input variables $\mathbf{X}$ and output variable $\mathbf{y}$ were described by different evolutionary models, we are interested in the information content of selected instances *in the output variable*, so it

is sufficient to use one covariance suitable for $\mathbf{y}$ in determining a sufficient neighbourhood and computing the output value.

Phylogenetic KNN regression is presented in Algorithm 5. An instance-based classification algorithm can be gained by replacing phylogenetic mean in the algorithm output with effective sample size voting, see Algorithm 6. In phylogenetic KNN classification, prediction is the label with maximum effective sample size in the set of nearest neighbours. We chose to use Euclidean distance in these algorithms, but alternative distance metrics can be applied as well. Specifically, use of phylogenetic information does not prevent application of any distance metrics in the input variable space.

Algorithm 5: Phylogenetic k nearest neighbour regression.

input : Input features $\mathbf{X} \in \mathbb{R}^{n \times p}$, output feature $\mathbf{y} \in \mathbb{R}^n$, covariance matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$, instance of query $\mathbf{x}' \in \mathbb{R}^p$ and hyperparameter k

output: Estimate $\hat{y}(\mathbf{x}')$

$\mathbf{X}_{sorted} \leftarrow$ sort $\mathbf{X}$ by increasing distance to $\mathbf{x}'$

$nn \leftarrow$ binary search number of neighbours in $\mathbf{X}_{sorted}$ such that $\text{Ess}\{\mathbf{X}_{sorted}[1:(nn-1),]\} < k$ and $\text{Ess}\{\mathbf{X}_{sorted}[1:nn,]\} \geq k$

$neighbours \leftarrow index(\mathbf{X}_{sorted}[1:nn,])$

return *Phylogenetic mean of* $\mathbf{y}_{neighbours}$ *given* $C_{neighbours,neighbours}$

Algorithm 6: Phylogenetic k nearest neighbour classification.

input : Input features $\mathbf{X} \in \mathbb{R}^{n \times p}$, output feature $\mathbf{y} \in \mathbb{R}^n$, covariance matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$, instance of query $\mathbf{x}' \in \mathbb{R}^p$ and hyperparameter k

output: Estimate $\hat{y}(\mathbf{x}')$

$\mathbf{X}_{sorted} \leftarrow$ sort $\mathbf{X}$ by increasing distance to $\mathbf{x}'$

$nn \leftarrow$ binary search number of neighbours in $\mathbf{X}_{sorted}$ such that $\text{Ess}\{\mathbf{X}_{sorted}[1:(nn-1),]\} < k$ and $\text{Ess}\{\mathbf{X}_{sorted}[1:nn,]\} \geq k$

$neighbours \leftarrow index(\mathbf{X}_{sorted}[1:nn,])$

return $\arg\max_c \text{Ess}\{\mathbf{y}_{neighbours} = c\}$

5.4 Regression trees

Regression tree divides input feature space into M regions $R_1, R_2, ..., R_M$ and models y as constant c_m in regions m [26]:

$$f(\mathbf{x}) = \sum_{m=1}^{M} c_m I(\mathbf{x} \in R_m)$$

I is an indicator function. In the i.i.d. case the tree is trained with a recursive greedy algorithm which minimises sum of square error loss in binary partition to regions R_1 and R_2 [26]:

$$\min_{R_1,R_2}[\min_{c_1}\sum_{x_i\in R_1}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2}(y_i-c_2)^2]$$

Solution for internal minimisations is arithmetic mean in each region.

We take into account phylogeny by replacing sums of square error with phylogenetic error. Binary partitioning is now determined by condition:

$$\min_{R_1,R_2}[\min_{\mu_1}(\mathbf{y}_1-\mu_1\mathbf{1})^T\mathbf{C}_1^{-1}(\mathbf{y}_1-\mu_1\mathbf{1})+\min_{\mu_2}(\mathbf{y}_2-\mu_2\mathbf{1})^T\mathbf{C}_2^{-1}(\mathbf{y}_2-\mu_2\mathbf{1})]$$

where $\mathbf{x}_j \in R_j$, and $\mathbf{C}_j$ are covariance matrices for instances in regions $j = 1, 2$. In this case, solution for internal minimisations is phylogenetic mean in each region:

$$\hat{\mu}_j = (\mathbf{1}^T\mathbf{C}_j^{-1}\mathbf{1})^{-1}\mathbf{1}^T\mathbf{C}_j^{-1}\mathbf{y}_j$$

Constructing a generalising tree, we should avoid branches that contain only closely related instances. That is, the branches should contain enough information and large enough effective sample size. This requirement leads to effective sample size regularisation $\mathbf{1}^T\mathbf{C}_m^{-1}\mathbf{1} \geq \alpha$ for region R_m and some real number α, which is a generalisation of leaf size regularisation.

5.5 Perceptron

Perceptron is a linear classification algorithm for a binary output feature [26], which we assume to be $y \in \{0, 1\}$. Prediction is based on input features $\mathbf{x} \in \mathbb{R}^p$ and weights $\mathbf{w} \in \mathbb{R}^p$. We assume that the first element of $\mathbf{w}$ represents the bias, and the first element of $\mathbf{x}$ equals 1. The output function of perceptron is:

$$\hat{y} = f(\mathbf{x}, \mathbf{w}) = \begin{cases} 1 & \text{if } \mathbf{w}\cdot\mathbf{x} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Ordinary perceptron learning algorithm updates the weights $\mathbf{w}$ with misclassified instances $(\mathbf{x}_i, y_i)$ multiplied with learning rate: $\mathbf{w}_{j+1} \leftarrow \mathbf{w}_j + \eta(y_i - \hat{y}_i)\mathbf{x}_i$. This learning algorithm handles misclassified instances in a local and greedy manner, and it is suitable for linearly separable data. If phylogenetic data is linearly separable, no changes are required for the perceptron learning algorithm. It is therefore more interesting to consider learning algorithms for non-separable cases, where we need to weigh different non-reducible errors in light of phylogenetic information.

A modification of perceptron learning for non-separable data is the pocket algorithm, which stores the weights $\mathbf{w}.s$ with the most consecutive correct predictions in the learning

process [31]. We propose two phylogenetic modifications of this pocket perceptron learning algorithm. In the case of classification of phylogenetic data, we do not want our algorithm to be greedy purely about misclassified instances, but misclassified effective sample size. Algorithm 7 is a pocket perceptron algorithm, where we iterate the instances, update the weights w and store a suitable set of weights $\mathbf{w}.s$. In this algorithm, the stored weights $\mathbf{w}.s$ are chosen by largest consecutively correctly classified effective sample size (instead of number of instances). If the phylogenetic covariance matrix $\mathbf{C}$ is the identity $\mathbf{I}$, we regain the ordinary pocket perceptron.

In Algorithm 7, the learning itself is not guided by phylogenetic information. To find the linear decision surface that classifies correctly as much effective sample size as possible, we can guide the learning algorithm by the wrongly classified effective sample size and minimise it. Additionally, we can store the weights with largest global effective sample size accuracy. This is the principle of Algorithm 8. In the case of identity covariance, Algorithm 8 reduces to a learning procedure with weight storing by the global classification accuracy. The advantage of the locally determined weight storing of the pocket perceptron compared to globally determined storing are lighter memory and time requirements. The same applies to this phylogenetic algorithm, namely the training process is heavy with frequent effective sample size calculations, but on the other hand training should require fewer epochs because of the directed nature of the training. Algorithm 8 should be considered as a brute-force approach to finding decision surface with maximum effective sample size accuracy.

5.6 Neural network regression

Let function f represent a neural network, and $\hat{y} = f(\mathbf{x}, \mathbf{w})$ is the prediction of the network with input $\mathbf{x}$ and weights $\mathbf{w} \in \mathbb{R}^p$. We assume that the phylogenetic covariance matrix $\mathbf{C}$ is constant. Phylogenetic Gaussian loss function for the network is:

$$L(\mathbf{w}; \mathbf{X}, \mathbf{y}) = (\mathbf{y} - f(\mathbf{X}, \mathbf{w}))^T \mathbf{C}^{-1} (\mathbf{y} - f(\mathbf{X}, \mathbf{w}))$$

The inverse covariance can again be decomposed with a matrix root (Cholesky decomposition), which gives:

$$L(\mathbf{w}) = [\mathbf{L}^{-1}(\mathbf{y} - f(\mathbf{X}, \mathbf{w}))]^T \mathbf{L}^{-1} (\mathbf{y} - f(\mathbf{X}, \mathbf{w}))$$

If $f(\mathbf{X}, \mathbf{w}) = \mathbf{X}\mathbf{w}$, then the network is linear and $\mathbf{L}^{-1} f(\mathbf{X}, \mathbf{w}) = f(\mathbf{L}^{-1}\mathbf{X}, \mathbf{w})$. In this case the loss function $L(\mathbf{w})$ is equivalent to:

$$(\mathbf{L}^{-1}\mathbf{y} - \mathbf{L}^{-1}\mathbf{X}\mathbf{w})^T (\mathbf{L}^{-1}\mathbf{y} - \mathbf{L}^{-1}\mathbf{X}\mathbf{w})$$

Algorithm 7: Perceptron with effective sample size -controlled weight storing.

input : Input features $\mathbf{X} \in \mathbb{R}^{n \times p}$, output feature $\mathbf{y} \in \{0, 1\}^n$, covariance matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$, number of epochs and learning rate η

output: Weight vector $\mathbf{w} \in \mathbb{R}^p$

```
randomly initialise weights w
w.s ← w
h.s ← 0
h ← 0
h.index ← Boolean vector with value False for all instances
for i ← 1 to epochs do
    Shuffle training data
    for j ← 1 to n do
        y.hat ← f(x_j, w)
        if y.hat = y_j then
            h.index_j ← True
            h ← Ess(C[h.index, h.index])
            if h > h.s then
                h.s ← h
                w.s ← w
        else
            set h.index False for all instances
            h ← 0
            w ← w + η(y_j − y.hat)x_j
return w, w.s
```

Algorithm 8: Perceptron with effective sample size -controlled learning. Instances $(\mathbf{x}, y)$ are chosen for weight update based on misclassified effective sample size on each side of the decision surface.

input : Input features $\mathbf{X} \in \mathbb{R}^{n \times p}$, output feature $\mathbf{y} \in 0, 1^n$, covariance matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$, number of epochs and learning rate η

output: Weight vector $\mathbf{w} \in \mathbb{R}^p$

randomly initialise weights $\mathbf{w}$
$\mathbf{w}.s \leftarrow \mathbf{w}$
$h.s \leftarrow 0$
for $i \leftarrow 1$ **to** *epochs* **do**
  for $k \leftarrow 1$ **to** n **do**
    $y.hat \leftarrow$ predictions $f(x, w)$ for all data
    $error.plus \leftarrow$ indices for which $y.hat > y$
    $error.minus \leftarrow$ indices for which $y.hat < y$
    $ess.plus \leftarrow \text{Ess}(C[error.plus, error.plus])$
    $ess.minus \leftarrow \text{Ess}(C[error.minus, error.minus])$
    $ess.ac \leftarrow$ effective sample size accuracy of $\mathbf{w}$
    if $ess.ac > h.s$ **then**
      $h.s \leftarrow ess.ac$
      $\mathbf{w}.s \leftarrow \mathbf{w}$
    if $ess.plus > 0$ *and* $ess.minus > 0$ **then**
      $u \leftarrow$ sample uniform distribution from 0 to 1
      if $u < ess.minus/(ess.minus + ess.plus)$ **then**
        choose j from $error.minus$
      else
        choose j from $error.plus$
    else if $ess.plus > 0$ **then**
      choose j from $error.plus$
    else if $ess.minus > 0$ **then**
      choose j from $error.minus$
    else
      return $\mathbf{w}$, $\mathbf{w}.s$
    $\mathbf{w} \leftarrow \mathbf{w} + \eta(y_j - y.hat_j)\mathbf{x}_j$
return $\mathbf{w}$, $\mathbf{w}.s$

With phylogenetic GLS transformation, weights $\mathbf{w}$ can be optimised with functions for i.i.d. data. In the general case of non-linear network the equality $\mathbf{L}^{-1}f(\mathbf{X},\mathbf{w}) = f(\mathbf{L}^{-1}\mathbf{X},\mathbf{w})$ does not hold.

To minimise $L(\mathbf{w})$ in the general non-linear case, we need to compute $\hat{\mathbf{w}}$ with some optimisation algorithm like gradient descent. Next we will derive the gradient of the loss function. Partial derivative of L along $\mathbf{w}$ is:

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}[(\mathbf{y} - f(\mathbf{X},\mathbf{w}))^T\mathbf{C}^{-1}(\mathbf{y} - f(\mathbf{X},\mathbf{w}))]$$

Matrix $\mathbf{C}^{-1}$ is symmetric and does not depend on $\mathbf{w}$, so by proposition 14 in [5], the gradient takes the form:

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = 2(\mathbf{y} - f(\mathbf{X},\mathbf{w}))^T\mathbf{C}^{-1}\frac{\partial(\mathbf{y} - f(\mathbf{X},\mathbf{w}))}{\partial \mathbf{w}}$$

When we write this with Cholesky decomposition, the gradient becomes:

$$\begin{aligned}\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} &= 2[\mathbf{L}^{-1}(\mathbf{y} - f(\mathbf{X},\mathbf{w}))]^T\mathbf{L}^{-1}\frac{\partial(\mathbf{y} - f(\mathbf{X},\mathbf{w}))}{\partial \mathbf{w}} \\ &= -2[\mathbf{L}^{-1}(\mathbf{y} - f(\mathbf{X},\mathbf{w}))]^T\mathbf{L}^{-1}\frac{\partial f(\mathbf{X},\mathbf{w})}{\partial \mathbf{w}}\end{aligned}$$

Here $\frac{\partial f(\mathbf{X},\mathbf{w})}{\partial \mathbf{w}}$ is a Jacobian matrix of $n \times p$ elements, and the gradient is a row vector of p elements.

6. Description of Data

The goal of the experimental analysis is twofold: to assess the performance of new methods and to assess the necessity and impact of phylogenetic correction in dental trait - environmental analysis. Species and variables in the analysed comparative data are described here, as well as employed computational tools.

6.1 Overview of Datasets and Tools

The computations were done with R [61], using libraries Ape [57], Caper [52], Kohonen [77], Lars [39], Madness [58], Phylolm [40], Phytools [64] and Rpart [73].

The sampled species in this study are large mammals from all over the world. The data on mammal dental traits is from New and Old Worlds database [48]. Bioclimatic variables for mammal species were extracted from WorldClim [21].

Mammal species were observed on different geographical sites that are spread as a grid around the world [51]. Values of bioclimatic variables were extracted for each site, and for each mammal species, a distribution of bioclimatic values were constructed. From this distribution, either mean, maximum, minimum, or standard deviation was computed. If a species was observed only on one site, the standard deviation was assigned value zero.

Phylogeny of the studied mammals was constructed as a subtree of externally given mammalian supertree with 5020 species (Figure 6.1). The tree was originally compiled by Bininda-Edmonds et al. [6], and later updated by Fritz et al. [27]. The branch lengths in the phylogeny represent dating estimates of the development of the mammal species, and Fritz et al. provided trees with earliest, best and latest estimates. The phylogeny with best time estimates was selected for the analysis.

Estimates for mammalian body masses were compiled from PanTHERIA [46] and MammalBase [47] by Kari Lintulaakso. I received pre-processed body mass values for 489 mammals. For the remaining 11 species, body mass value was either imputed as the average of existing body mass estimates, or the species dropped from the analysis.

Data on dental traits was received pre-processed from Galbrun et al. [30], and it contained information on several dental traits of large herbivorous mammals. The dataset also included family and order of each of the mammal species.

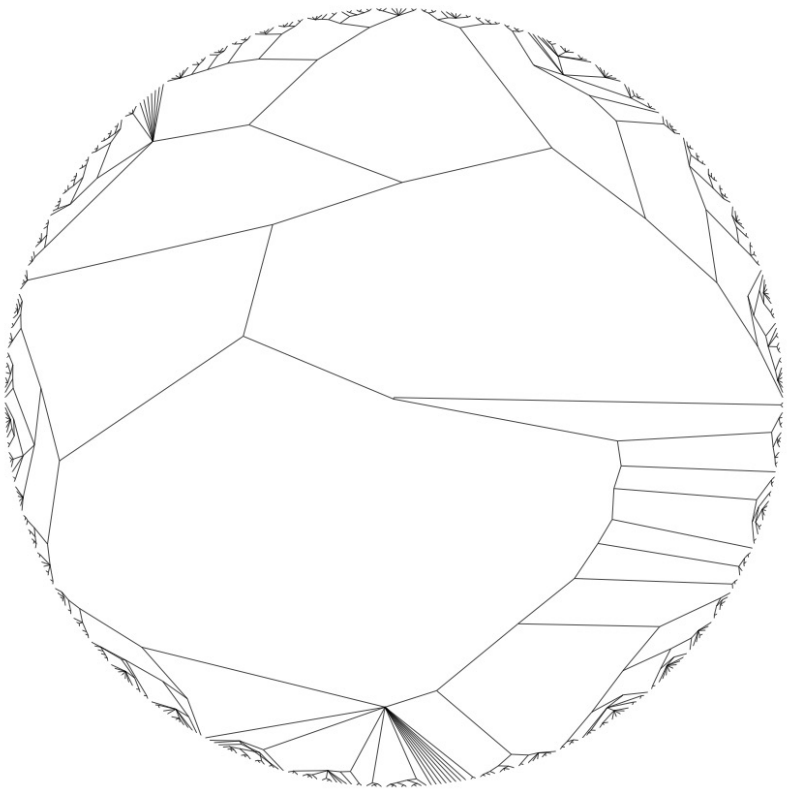

Figure 6.1: Mammalian supertree [27] in a radial format. The tree contains the ancestry relationships of 5020 mammals. The inferred ancestor of all mammals is the node in the centre of phylogeny. In this representation, time flows from the centre to the outer ring, which contains the leaves that represent 5020 extant mammals. The ancestor of all the mammal species was estimated to live 166.2 Ma ago, but in further analysis we use relative time from 0 to 1. Species names are omitted to keep the figure readable. Most diversification was estimated to happen towards the end of mammal evolution.

The combined dataset of mammal dental traits and per species values for bioclimatic variables consists of 490 species. The mammalian orders appearing in the dataset are presented in Table 6.1, and families in Table 6.2. The phylogeny of the dataset is presented in Figure 6.2.

6.2 Dental Traits

Mammalian dental features in the analysis are based on functional crown type scoring scheme [80], which describes functional features in the teeth of herbivorous mammals. The dental features are ordinal.

Hypsodonty (HYP) is a measure of relative crown height. Low teeth are called

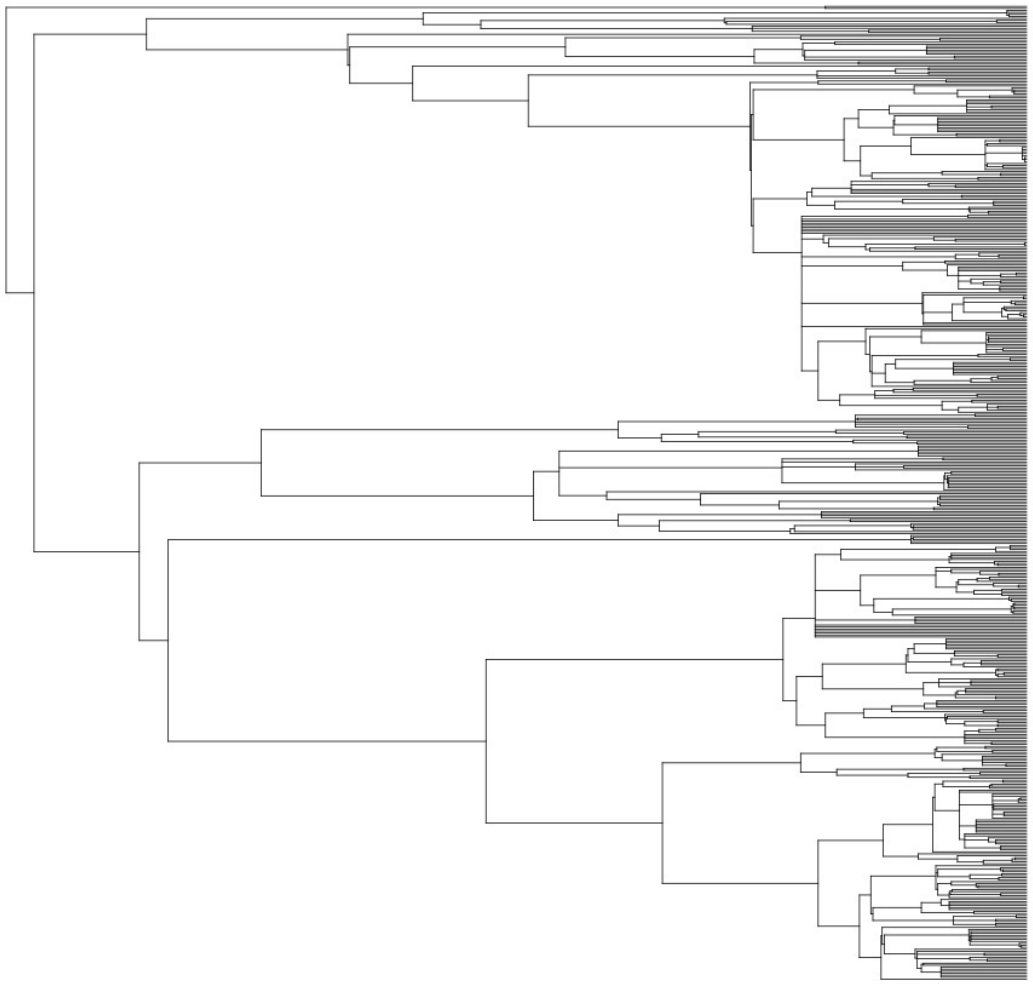

Figure 6.2: Phylogeny of the dataset. The data contains 490 large mammals. This tree is constructed as a subset of the mammalian phylogeny of 5020 species [27]. In this representation, horizontal axis from left to right represents time. The common ancestor of the studied species was estimated to live 98.9 Ma ago. The leaves of the tree in the right side represent the 490 extant mammals. In further analysis the absolute values or units of time are not important, so we assume relative time from 0 to 1. The vertical axis has no biological meaning. Like in the tree 5020 mammals (Figure 6.1), most diversification here happens late in the evolution. Names of the species were omitted to keep the figure readable.

brachyodont (1), equally high and long teeth are mesodont (2), and high teeth are hypsodont (3). Average ordinated hypsodonty by site tends to have negative relationship with net primary productivity and mean annual precipitation, that is high hypsodonty value is associated with low net primary productivity and low precipitation [48].

Two dental features, acute lophs (AL) and obtuse lophs (OL), indicate the presence of different kinds of lophs. A loph is a continuous structure that consists of enamel crests or closely spaced cusps, and that is at least half the size of the tooth. If the enamel crest forms a sharp edge, or supports a planar wear facet, then the lophs are acute. The other possibility is that the loph is obtuse or basin-like [80].

Longitudial loph count (LOP) measures the number of longitudial cutting edges per tooth from back of the mouth to front. Longitudial lophs count can be interpreted as cutting capacity of teeth [48]. Like HYP, LOP has a negative relationship with net primary productivity and precipitation. LOP is also negatively connected to mean annual temperature [48].

Cusps are elevated points of pointed or rounded shape on the tooth crown. Structural fortification of cusps (SF) indicates whether fortification of cusps is present (1) or not (0). The fortification amplifies the protrusion of the cusp [80].

Occlusal topography (OT) describes the topography of teeth when they are in contact. The contact of teeth can involve raised cusps of lophs (0), or happen in one surface with cusps and lophs embedded in cementum (1). It is observed in direction perpendicular to occlusal movement. Because occlusal topography is primarily determined by dental structure as opposed to diet, and measured in the direction that exhibits little connection with diet [80], it is expected to have little environmental signal.

Coronal cementum (CM) describes the presence of cementum on the crown and roots of the tooth. Absent or very thin cementum is indicated with value 0, and thick sementum with value 1 [80]. Enamel thickness (ETH) is a visual estimate of thickness of the enamel layer, and it can be thin (0) or thick (1).

Anisodonty (ADI) measures the width ratio of upper and lower teeth. Value 1 corresponds to ratios smaller than one. Maximum value on anisodonty is 5, and it corresponds to upper and lower teeth ratio of approximately 2. Like ETH, ADI is also based on subjective visual inspection.

6.3 Bioclimatic variables

Environment of the studied mammal species was modelled with bioclimatic variables, which are derived from monthly temperature and precipitation values on studied sites [21]. Net primary productivity (NPP) of a geographical location is an estimate of annual production of plant matter measured as grams of carbon per square meter per year. It

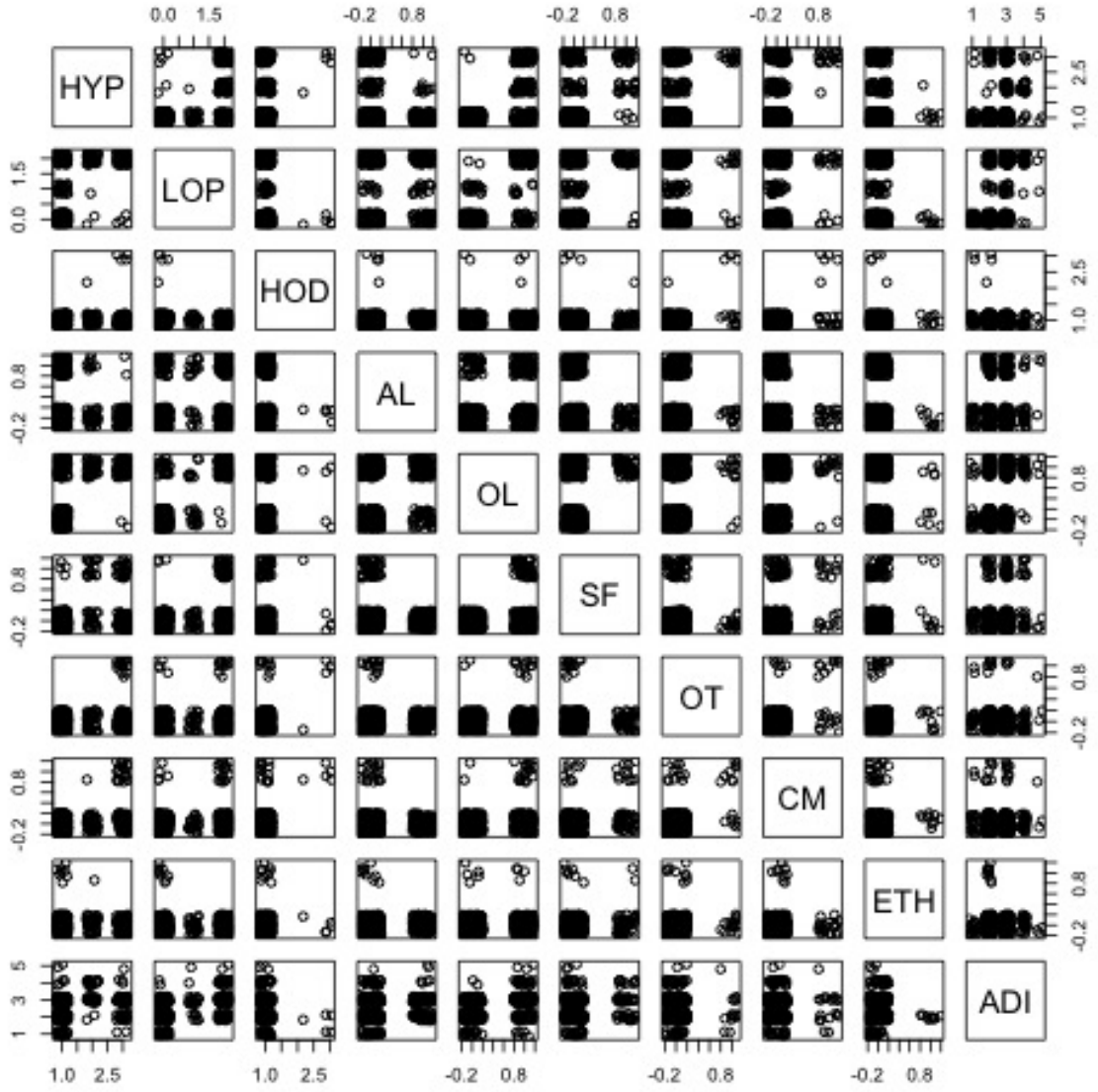

Figure 6.3: Scatter plots of dental traits. Some jitter was added to the data to make distributions visible on the ordinal dental trait data.

Table 6.1: Mammalian orders in the dataset.

Order	Artiodactyla	Perissodactyla	Primates	Proboscidea
Species	191	12	285	2

is minimum of productivities allowed by temperature and precipitation, so it is capped by either temperature or precipitation. As described in [48], temperature productivity is $\text{NPPt} = 3000/(1 + \exp(1.315 - 0.119\text{MAT}))$, and precipitation productivity is $\text{NPPp} = 3000 \cdot (1 - \exp(-0.000664\text{MAP}))$. Net primary productivity is:

$$\text{NPP} = \min(\text{NPPt}, \text{NPPp})$$

A binary version of NPP was constructed to serve as a target variable for binary classification. If the mean NPP for an animal was larger than median of all mean NPP values, binary NPP was assigned value 1, and 0 otherwise.

Mean annual temperature, mean annual precipitation and net primary productivity were studied with dental traits in [48]. Other studies about the relationship of dental traits and bioclimatic variables include [17], [16], [23] and [80].

Sometimes a biological law is evident in an extreme value of a feature rather than mean (e.g. Bergmann's law on body mass and maximum latitude [10]). Minimum and maximum of NPP were investigated in relation to dental traits. Temperature was studied through average mean annual temperature, maximum and minimum mean annual temperature, mean and maximum of warmest quarter temperature, mean and minimum of coldest quarter temperature. Variability of temperature was described with mean temperature seasonality and maximum temperature seasonality. Precipitation was described with mean annual precipitation, mean precipitation of driest quarter and mean precipitation of wettest quarter. Latitude was studied with mean and maximum absolute latitude.

Table 6.2: Mammalian families in the dataset.

Family	Antilocapridae	Bovidae	Callitrichidae	Camelidae
Species	1	117	23	1
Family	Cebidae	Cercopithecidae	Cervidae	Cheirogaleidae
Species	68	101	41	12
Family	Daubentoniidae	Elephantidae	Equidae	Galagidae
Species	1	2	4	8
Family	Giraffidae	Hippopotamidae	Hominidae	Hylobatidae
Species	2	1	6	11
Family	Indridae	Lemuridae	Lorisidae	Megaladapidae
Species	9	17	8	6
Family	Moschidae	Pitheciidae	Rhinocerotidae	Suidae
Species	7	10	4	12
Family	Tapiridae	Tarsiidae	Tayassuidae	Tragulidae
Species	4	5	3	6

7. Case Study

The goals of the empirical study on dental traits and environmental features are to test the new methods on real data, and to gain knowledge about the relationship of mammal dental traits and environments. Of the new methods presented in this study, instance based methods and perceptron training are not directly based on the established loss function presented in Section 3.4, so their behaviour was studied with synthetic data. Phylogenetic signal for both dental traits and bioclimatic variables was computed to demonstrate that the use of phylogenetic methods is necessary for this data. Some unsupervised analysis was performed for the dental traits to gain an overview of their distributions. Regularised PGLS, instance-based regression and logistic regression were applied on mammal dental traits and bioclimatic variables.

7.1 Instance-based regression with synthetic data

Nearest neighbours regression was studied with synthetic data, where input and output features contained phylogenetic signal. Performance of ordinary and phylogenetic k nearest neighbour regression was compared in modelling a non-linear target function. The datasets were three-dimensional with two continuous input features and one continuous output feature.

Input features x_1 and x_2 were sampled from Brownian processes $\mathbf{x}_1, \mathbf{x}_2 \sim \mathrm{N}(\mathbf{0}, \mathbf{C})$, where $\mathbf{C}$ is Brownian covariance of a random phylogeny (function `rcoal` of the library Ape). Output feature $\mathbf{y}$ was sampled with a non-linear function $\mathbf{y}_i = (\mathbf{x}_{1i}\mathbf{x}_{2i})^3 + \varepsilon_i$, where random variable ε is sampled from a Brownian process $\varepsilon \sim \mathrm{N}(\mathbf{0}, \mathbf{C})$. First, a validation set of 100 instances was sampled from this distribution. Then training sets of 200 instances were simulated 100 times. In each simulation, the output variable values of the validation set were predicted with ordinary and phylogenetic k nearest neighbour regression. To choose the values of hyperparameters k for ordinary (number of neighbours) and phylogenetic (effective sample size of neighbours) instance based regression, the y-values of the validation set were predicted with the training set, using different values of k. For ordinary KNN, the prediction was done with ten values from 1 to 200. In the case of phylogenetic KNN, ten values were tried evenly from the interval $[1, \mathrm{Ess}(C)]$. Values of k

that predicted the validation set y with smallest squared error were chosen for prediction. Output feature value $y(0,0)$, with theoretically optimal value 0, was predicted with both algorithms.

Mean number of neighbours in ordinary KNN prediction $\hat{y}(0,0)$ was 48, with average prediction -0.11 and variance 0.63. The average effective sample size in phylogenetic KNN prediction was 1.4, resulting in average of 33 neighbours used in prediction. The prediction average in phylogenetic KNN was -0.087 with variance 0.65. Both algorithms predicted values close to the true value 0 with similar prediction variance. We also experimented with selecting the hyperparameter k of phylogenetic KNN with phylogenetic error objective, but this approach did not lead to variance reduction.

The numbers of neighbours used in the predictions ranged from 1 to dataset size 200 in both algorithms. In 100 repeated experiments with phylogenetic KNN, the prediction $\hat{y}(0,0)$ was made with one neighbour 54 times. Another frequent value was 200 neighbours, which occurred in 10 experiments. Intermediate values were employed in the remaining 36 experiments. Histogram of the number of neighbours is presented in Figure 7.1. Ordinary KNN behaved similarly, prediction was done with one neighbour 51 times, and with 200 neighbours 14 times, leaving intermediate values for 35 experiments. Of these extreme values, dataset size 200 is problematic because in that case prediction $\hat{y}(0,0)$ is simply mean or phylogenetic mean of the whole training set. In these cases the validation set likely was not informative of the predictive power of the training set, possibly because training data ended up to different area of (x_1, x_2)-plane than the validation set. Data that is sampled with a phylogenetic covariance has a tendency to form clusters determined by branches of the phylogeny.

7.2 Perceptron with synthetic data

Perceptron learning algorithms were studied with synthetic data where both input and output features contained phylogenetic signal. The learning algorithms were ordinary perceptron learning with weight storing (pocket algorithm), perceptron learning with weight storing by effective sample size (Algorithm 7) and directed learning algorithm (Algorithm 8). Experiments were done first with random, and then with unbalanced phylogenies. For weight storing algorithms, the stored weights were chosen as outputs in all experiments.

In the experiment with random phylogenies, phylogenetic datasets were simulated 100 times. Each dataset with two continuous input features and a binary output feature was based on a simulated phylogeny with 200 tips, that is 200 species (Ape function `rcoal`). An output feature, a binary trait $y \in \{0,1\}$ was simulated along these trees with transition rates $\mathbf{Q}_{i,i} = 0.7$ and $\mathbf{Q}_{i,j} = 0.3$ for $i,j = 0,1$, $i \neq j$ (Phytools function

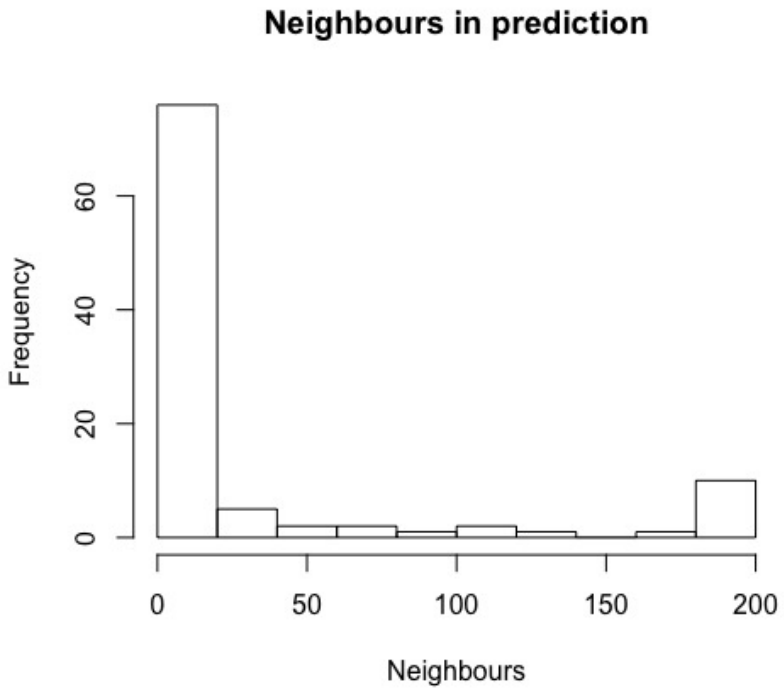

Figure 7.1: Number of neighbours in prediction with phylogenetic KNN in synthetic data experiments. Output value in $(0, 0)$ was predicted based on output values of nearest neighbours of the origin. Two continuous input features and an output feature were created with a Brownian process along a simulated phylogeny. Optimal value of hyperparameter k was determined by predicting output feature values in a validation set. Then prediction $\hat{y}(0, 0)$ was made with phylogenetic KNN. This figure shows the distribution of number of neighbours used in this prediction. The distribution is bimodal with maxima in one neighbour (54 cases) and 200 neighbours (10 cases).

sim.Mk). Data which contained less than 5% of either label was discarded and resampled. Input features x_1 and x_2 were created as continuous Brownian traits for each dataset, separately for species $y = 0$ and $y = 1$. Values for x_1 had the same mean for both classes, $\mathbf{x}_1 \sim N(\mathbf{0}, \mathbf{C}_i)$, $i \in \{0, 1\}$. For x_2, classes had separate means: $\mathbf{x}_2 \sim N(\mu_i, \mathbf{C}_i)$, $\mu_0 = -0.2$ and $\mu_1 = 0.2$. Note that these means are homogenous n-vectors.

Phylogenetic covariances were normed so that $\mathbf{C}_{i,i} = 1$ for all species i. With this design, simulated data was not necessarily linearly separable by label y. Theoretical optimal decision surface for classification of y was the x_1-axis. For perceptron weight vectors $\mathbf{w}$ with an added bias w_0 the theoretical optimal decision vectors were $[0, 0, a]$, $a > 0$. Perceptrons were trained for each of the 100 synthetic datasets using the binary trait covariance of Ives et al. [44] where applicable. The covariance parameter α was estimated from the data for each dataset, with average 1.5 for the 100 experiments.

The average weights are presented in Table 7.1. Effective sample size pocket algorithm produced weights closest to the theoretical values for bias and x_1, as well as correct sign for x_2-weight. Weight variances are presented in Table 7.2. None of the algorithms had clearly smaller weight variance than the rest.

Training accuracies and effective sample size accuracies were computed for all trained classifiers, the latter with the same binary trait covariances as the before. In the 100 experiments, ordinary perceptron reached an average accuracy of 0.68, and effective sample size accuracy of 0.54. The directed algorithm reached average accuracy of 0.92 and effective sample size accuracy of 0.79. The Ess-pocket algorithm had an average accuracy of 0.72, and its effective sample size accuracy was 0.57, which was slightly higher than for the ordinary algorithm. Comparing to this to the results of the directed algorithm, ordinary perceptron updates employed by Ess-pocket algorithm do not seem to be the best strategy to maximise effective sample size accuracy.

Additionally, the accuracy of the theoretical decision surface was studied in this experimental setup. The decision surface determined by vector $\mathbf{w}_t = [0, 0, 1]$ had an average accuracy of 0.59, with variance of 0.079. The decision surface given by $\mathbf{w}_t$ had an average effective sample size accuracy of 0.55, with variance of 0.017, computed with the fitted binary trait covariance. The absolute values of accuracy and effective sample size accuracy are not directly comparable, because they are different quantities. However, the smaller variance of effective sample size accuracy supports the claim that effective sample size-based classification decreases variance when analysing phylogenetic data.

The experiment with unbalanced phylogenies was identical to the previous one apart from the phylogeny. Ape functions `stree` and `compute.brlen` were used to construct an unbalanced ultrametric tree of 200 species and assign its branch lengths. The structure of this systematic tree is shown in Figure 7.2 in smaller size. Data was then simulated on this unbalanced phylogeny, and classifiers were trained with the same procedure as above.

algorithm	bias	x_1	x_2
ordinary	0.0589	0.0206	-0.0252
ordinary, ess	0.00554	-0.0131	0.0125
directed	-0.0173	-0.0266	0.199

Table 7.1: Weight vector means in a repeated simulation study with random phylogenies where the theoretical decision surface was given by vector $[0, 0, a]$, $a > 0$. The algorithms were ordinary perceptron with weight storing, perceptron with effective sample size -guided weight storing and perceptron with directed learning.

algorithm	bias	x_1	x_2
ordinary	0.193	0.103	0.134
ordinary, ess	0.130	0.130	0.157
directed	0.144	0.342	0.239

Table 7.2: Weight vector variances in a repeated simulation study with random phylogenies where the theoretical decision surface was given by vector $[0, 0, a]$, $a > 0$. The algorithms were ordinary perceptron with weight storing, perceptron with effective sample size -guided weight storing and perceptron with directed learning.

Weight vector means in this experiment are presented in Table 7.3, and variances in Table 7.4. In this experiment, the directed algorithm was the only one with average positive weight for x_2. Ordinary pocket perceptron shows larger variance of all weights than either of the phylogenetic algorithms. The ordinary algorithm reached an average accuracy of 0.82 and effective sample size accuracy of 0.66. Phylogenetic pocket algorithm reached accuracy of 0.81, and effective sample size accuracy of 0.67. The directed algorithm reached a comparable accuracy of 0.83, but higher effective sample size accuracy of 0.72. The theoretical decision surface had approximately same properties as before, the average accuracy was 0.57 with variance of 0.045, and effective sample size accuracy of 0.56 with variance 0.011.

algorithm	bias	x_1	x_2
ordinary	-0.00127	-0.0256	-0.00919
ordinary, ess	-0.0267	0.00723	-0.00759
directed	-0.00842	0.00606	0.0279

Table 7.3: Weight vector means in a repeated simulation study with unbalanced phylogenies where the theoretical decision surface was given by vector $[0, 0, a]$, $a > 0$. The algorithms were ordinary perceptron with weight storing, perceptron with effective sample size -guided weight storing and perceptron with directed learning.

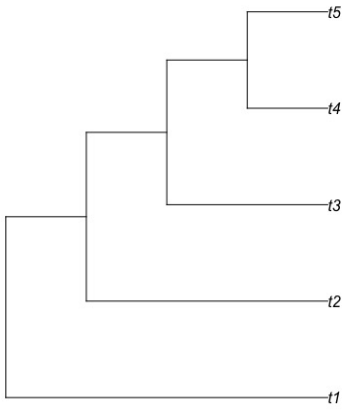

Figure 7.2: Unbalanced phylogenetic tree. Binary and continuous data was simulated on a tree with this structure and $n = 200$ tips.

algorithm	bias	x_1	x_2
ordinary	0.0980	0.0475	0.0271
ordinary, ess	0.0852	0.00560	0.0149
directed	0.0278	0.00450	0.0172

Table 7.4: Weight vector variances in a repeated simulation study with unbalanced phylogenies where the theoretical decision surface was given by vector $[0, 0, a]$, $a > 0$. The algorithms were ordinary perceptron with weight storing, perceptron with effective sample size -guided weight storing and perceptron with directed learning.

7.3 Unsupervised analysis of dental traits

A self organising map [77] was trained on the dental traits. The goal of this analysis was to present 10-dimensional dental data meaningfully in two dimensions. Hexagonal grid was chosen as the topology of the map. In experiments with different sized maps, the species were found to map only on a limited number of units. The reason for this tendency to clustering is that the ordinal dental data contains less variation than continuous data would. Therefore, a map of 36 units with 6×6 organisation was sufficient for the mammal data set. For the layout of the map, see Figure 7.3 or 7.4. Each unit on the self organising map was a 10-dimensional vector, matching 10-dimensional dental trait vectors x_i of species i. Map units m_k were initialised randomly, and then updated with the self organising map update procedure in 100 epochs. In an epoch, instances were processed one at a time so that the closest unit m_k to instance x_i in Euclidean distance was selected. The closest unit and its neighbours were updated with the rule $m_k \leftarrow m_k + \alpha(x_i - m_k)$, where α is a decreasing learning rate [26].

For a hexagonal map with 36 units, mammal species were mapped into 23 units, while 13 units were left empty. The code vectors of the trained map are presented in Figure 7.3. On the left side of the map, the elements of unit vectors have generally small values. Units on the right have larger element values. Mappings of the instances are presented in Figure 7.4. Species with dental traits close to zero were mapped on the left side of the map, and species with larger dental trait values were mapped to the units on the right. This means that species with simpler dentition were mapped mostly on two very populated units on the left, and species with more complex dentition were spread out on the right. One can see both highly populated and completely empty units on the map, and the reason for this is the ordinal nature of dental data. The different traits have two to five possible values, so this ordinal data contains less variability than continuous data would.

As the dental trait values of the species are products of an evolutionary process, one can expect to find a pattern between phylogenetic distance and distance on the self organising map. Unit distance on the self organising map was plotted as function of phylogenetic distance in Figure 7.5. Closely related species were often mapped to the same unit, but never to opposite sides of the map. Pairs of species that were mapped to units with maximum distance, that is six, had a phylogenetic distance that was equal or greater than 0.7 times maximum phylogenetic distance in the data. Based on Figure 7.5, mapping on the self organising map contains some phylogenetic signal. The upper left corner of the plot contains no pairs, which means that no closely related pair was different enough to be mapped on different sides of the self organising map. Pairs with large phylogenetic distance could be mapped to distant units, which can be seen on

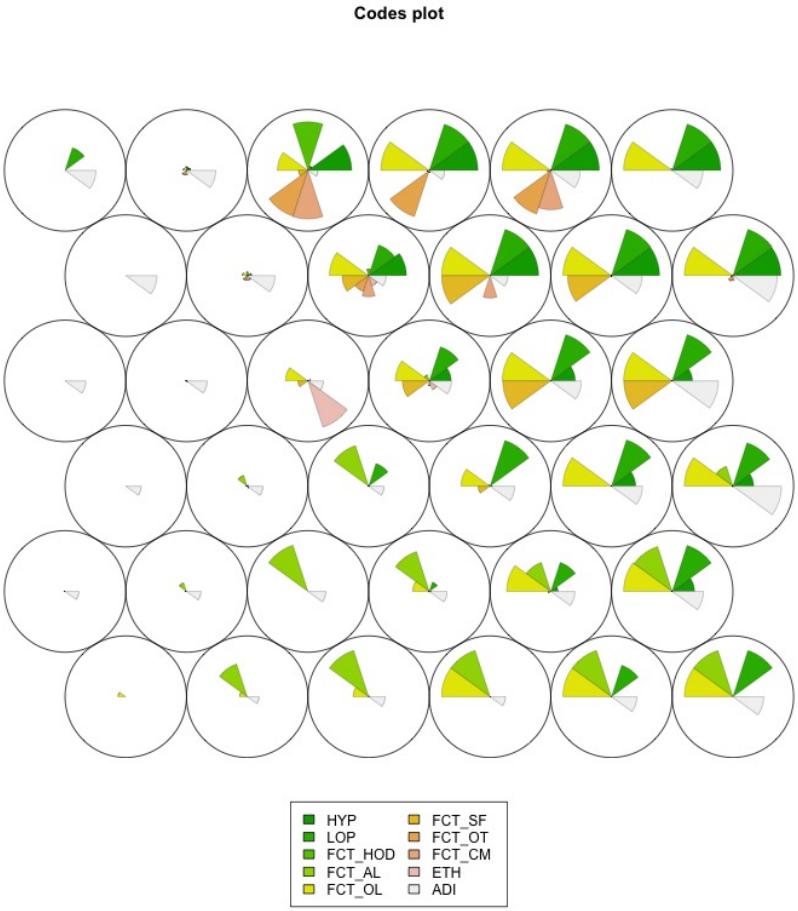

Figure 7.3: Codes plot of self organizing map on dental traits. The values of dental traits for different units in the converged map are shown in this figure. Even though dental traits are ordinal, the values plotted here are continuous parameters corresponding to different dental traits. In the training phase, the parameters were updated based on the dental data.

the upper right corner of the figure. This behaviour is expected, because dental traits have strong phylogenetic signal (see Table 7.8). This figure is an experimental study of phylogenetic signal, inspired by [70].

Ordinary and phylogenetic principal component analysis was performed on a matrix of (unscaled) dental traits. Variance distribution of phylogenetic principal components is shown in Figure 7.6. Loadings for first four principal components are presented in Table 7.5 for ordinary PCA, and Table 7.6 for phylogenetic PCA. For ordinary PCA, an ordinary LASSO procedure was conducted to select principal components for regression models with NPP. For phylogenetic PCA, phylogenetic LASSO was conducted to select components for PGLS.

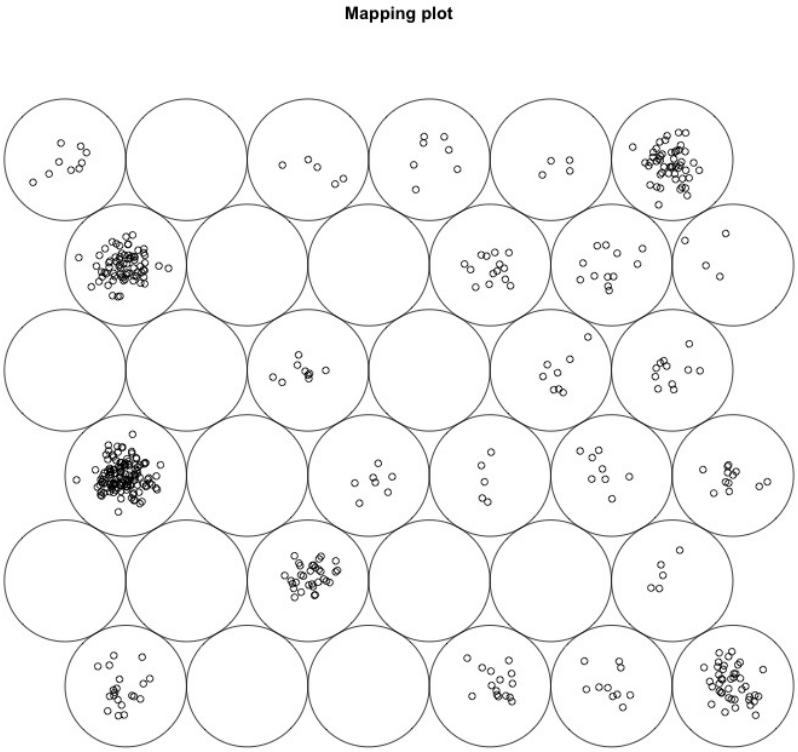

Figure 7.4: Instance mappings in self organizing map on dental traits. The trait vectors were mapped on a 6×6 hexagonal self-organising map. Positions of data points inside the units are arbitrary, being spread around the unit for readability. The values for most traits are zero on the left side of the map, and larger on the right side.

Variable	PC1	PC2	PC3	PC4
HYP	0.509	0.628	0.246	0.403
LOP	0.709	-0.140	-0.347	-0.422
HOD	0	0	0	0.256
AL	0	-0.356	-0.562	0.583
OL	0.336	0	-0.279	0.129
SF	0	0	0	-0.363
OT	0	0	0	0.204
CM	0	0.112	0	0.164
ETH	0	0	0	0
ADI	0.337	-0.647	0.651	0.193

Table 7.5: Loadings of dental traits in first four ordinary principal components. Zero values are not precisely zero, but they have small loading values.

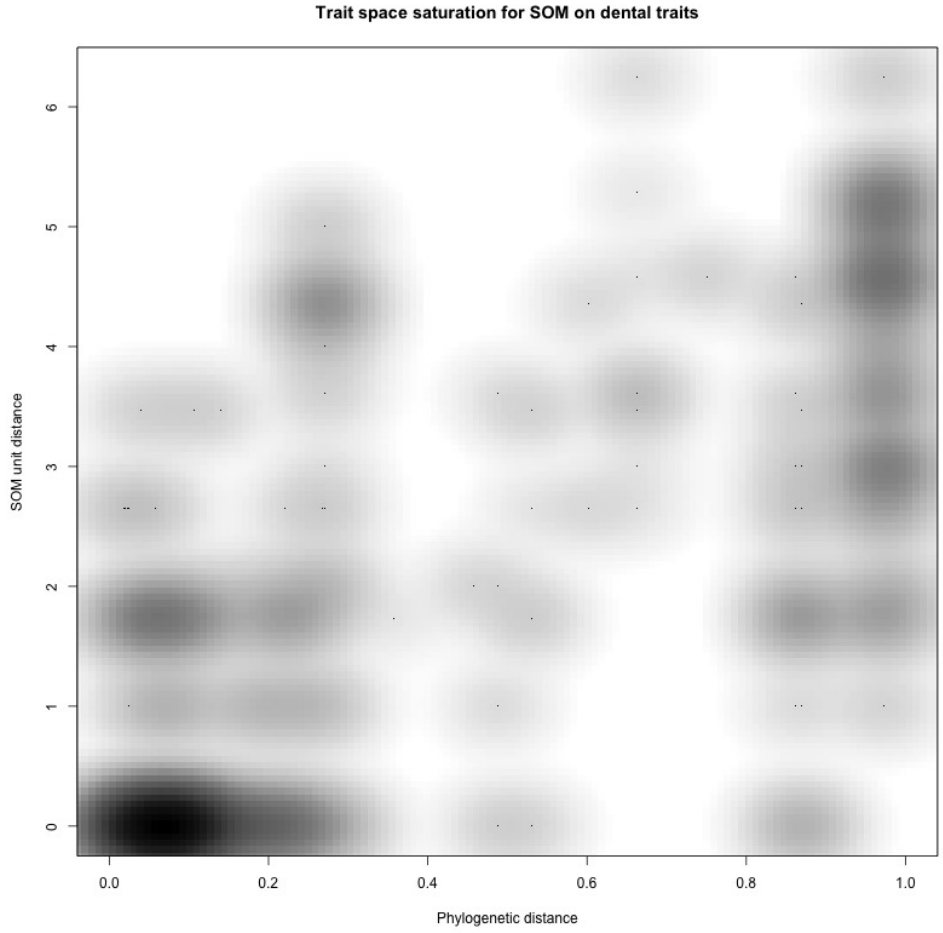

Figure 7.5: Unit distance on self organizing map as a function of phylogenetic distance. The data points in this figure are pairs of species. The horizontal axis is phylogenetic distance mapped between 0 and 1, and the vertical axis is unit distance on the converged 6×6 hexagonal self organising map. Closely related pairs of species are mapped on the left side of the figure, and we see that closely related species were mapped on either the same or nearby units.

Variable	PC1	PC2	PC3	PC4
HYP	-0.469	0.409	0.0165	-0.0116
LOP	-0.121	0.175	0.603	0.499
HOD	-0.786	0.0811	0.0228	-0.177
AL	0.101	-0.451	0.540	-0.469
OL	0.313	0.497	0.560	-0.291
SF	0.123	0.723	-0.320	0.0747
OT	-0.722	0.152	0.224	-0.156
CM	-0.299	0.231	0.108	-0.0547
ETH	0.629	0.419	0.108	-0.252
ADI	0.110	-0.0690	0.312	0.678

Table 7.6: Loadings of dental traits in first four phylogenetic principal components. The principal components were computed using a Brownian covariance matrix.

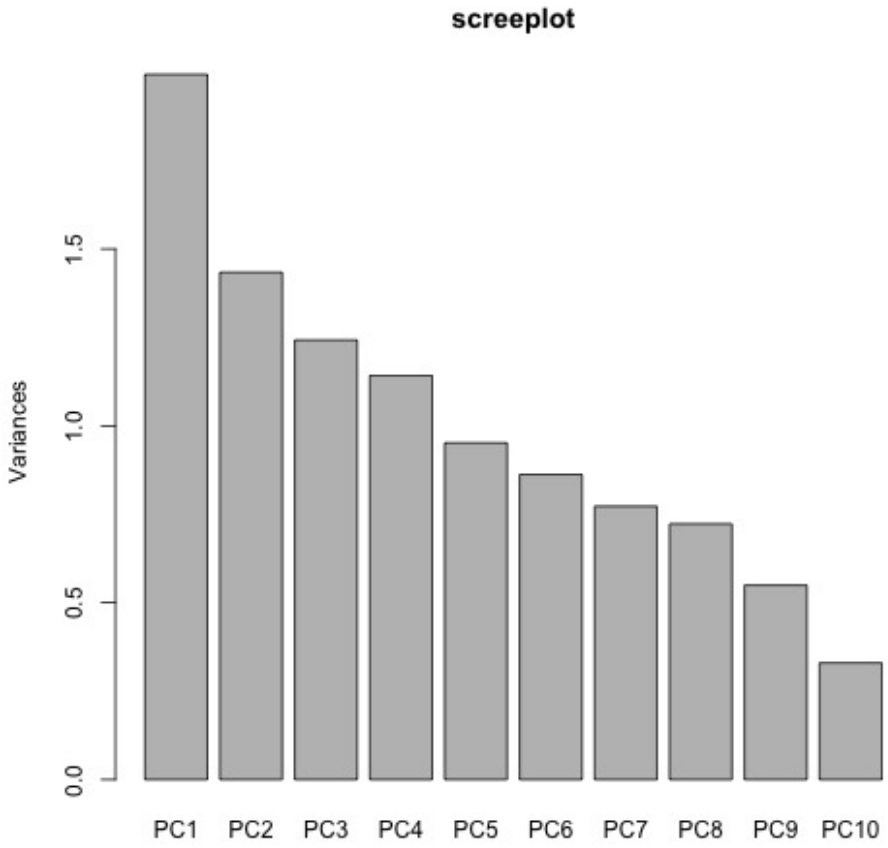

Figure 7.6: Relative variance of phylogenetic principal components of dental traits. Variance is distributed quite evenly among the ten principal components. Loadings of the first four principal components are presented in Table 7.6.

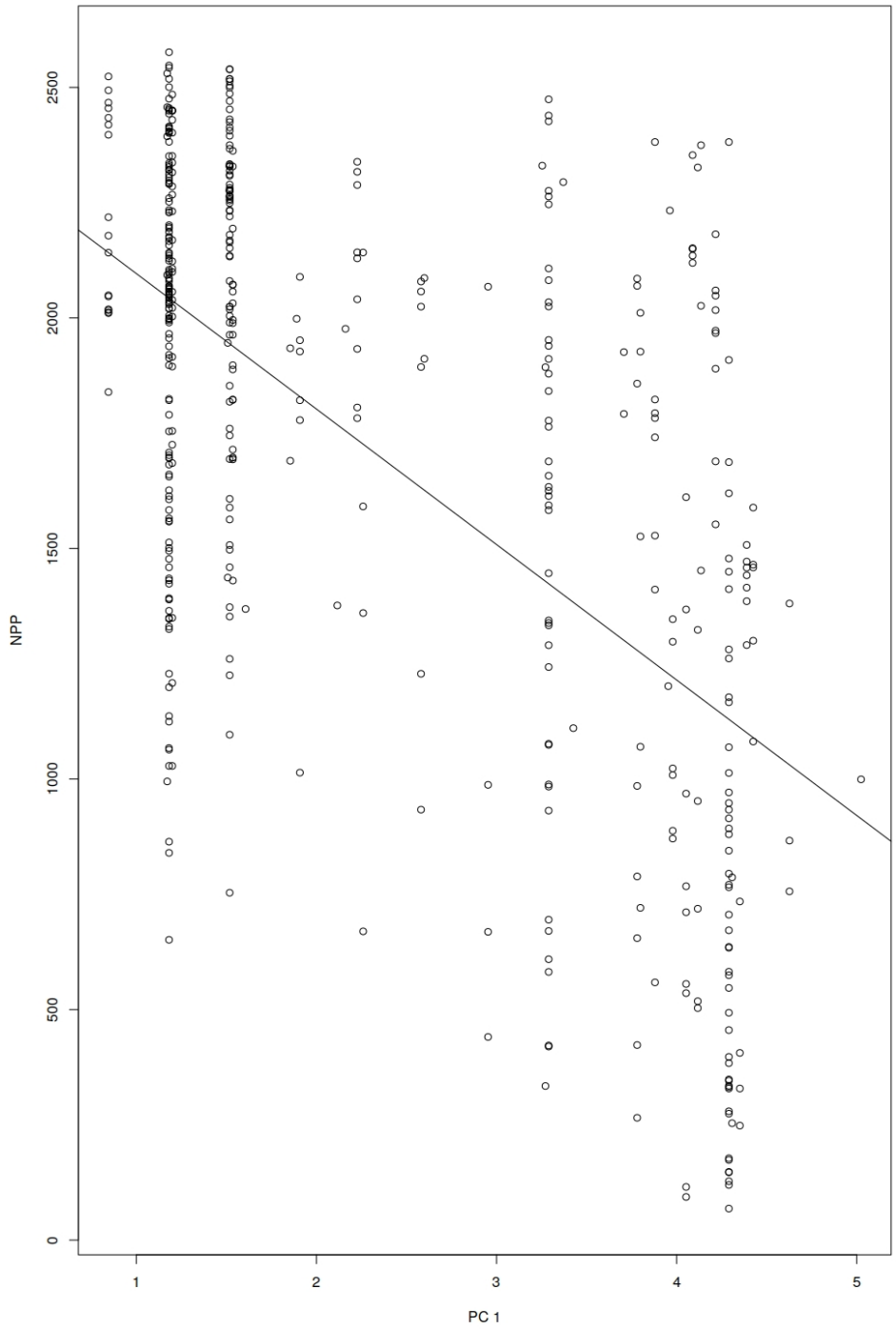

Figure 7.7: Principal component regression of ordinary principal component 1 on NPP. The OLS regression model drawn in black line is NPP = 2390 − 294PC1. Transforming the regression equation back to dental traits results in model that has large negative slopes for HYP, LOP and ADI. The first principal component was also the first of the components selected by ordinary LASSO procedure.

7.4 Phylogenetic signal

Pagel's λ was applied as a measure of phylogenetic signal for bioclimatic variables (Table 7.7). Means, minima and maxima of primary productivity, temperature and precipitation showed stronger phylogenetic signal than their standard deviations. Of the studied bioclimatic variables, NPP showed the strongest phylogenetic dependence with $\lambda = 0.952$.

Scatter plots of phylogenetic and NPP distances between species are displayed in Figures 7.8 and 7.9. The lower left corner in Figure 7.8 appears to be the most populated area in the graph, but actually the largest point density is in the lower right corner. This can be seen from the smoothed version of the graph (Figure 7.9). Very closely related species have similar average NPP, as seen from the pairs with phylogenetic distance close to zero. When phylogenetic distance increases, NPP distances get larger values. Small phylogenetic distances seem continuous, while large distances are discrete. The reason for this is the shape of the phylogenetic tree that is presented in Figure 6.2. Within closely related species, there is a varied distribution of phylogenetic distances. On the other hand, if the path from species i to j goes through earlier ancestors, number of possible path lengths decreases. If the path goes through the root of the tree, phylogenetic distance is always 1.

After missing body mass values were imputed, phylogenetic signal in logarithmic body mass was computed. Logarithmic body mass showed a strong phylogenetic signal with $\lambda = 0.996$.

Dental traits analysed in the study are binary or almost binary ordinal features. HYP, LOP, HOD and ADI were converted to binary variables, and D-statistic [28] estimates were computed on the dental traits. Strong phylogenetic signal was found in all of the dental traits. The results are displayed in Table 7.8. Scatter plots of phylogenetic distances and Euclidean distances of dental trait values for different species are displayed in Figures 7.10 and 7.11. No pairs of species are closely related but with very different dental trait values. Additionally, a smoothed scatter plot of phylogenetic distances and unit distances on self organising map is shown in Figure 7.5.

7.5 Net Primary Productivity

Both ordinary least squares and phylogenetic generalised least squares models were built for dental traits and NPP. LASSO regularisation was computed both for OLS and PGLS with lambda-transformed Brownian covariance. Sequences of LASSO moves for dental traits and NPP are presented in Table 7.9.

CM and SF coefficients for NPP are positive in OLS models and negative in PGLS models, so they were studied separately in univariate regressions. OLS model for coronal

Table 7.7: Pagel's λ [54, 25] for bioclimatic variables. The values are maximum likelihood λ estimates for the environmental variables in the phylogeny of the dataset (Figure 6.2). These environmental features are continuous, so we can estimate their phylogenetic signal with maximum likelihood λ-estimates. Here sd stands for standard deviation, NPP for net primary productivity and lat for absolute latitude. Lambda takes values $\lambda \in [0, 1]$, higher values corresponding to stronger phylogenetic signal.

mean NPP	0.952
sd NPP	0.392
min NPP	0.866
max NPP	0.794
mean lat	0.866
max lat	0.861
sd lat	0.561
mean mat	0.848
max mat	0.700
min mat	0.773
sd mat	0.503
mean warmest quarter temperature	0.753
max warmest quarter temperature	0.534
mean coldest quarter temperature	0.872
min coldest quarter temperature	0.832
mean temperature seasonality	0.752
max temperature seasonality	0.798
mean annual precipitation	0.853
sd annual precipitation	0.490
mean driest quarter precipitation	0.631
mean wettest quarter precipitation	0.687

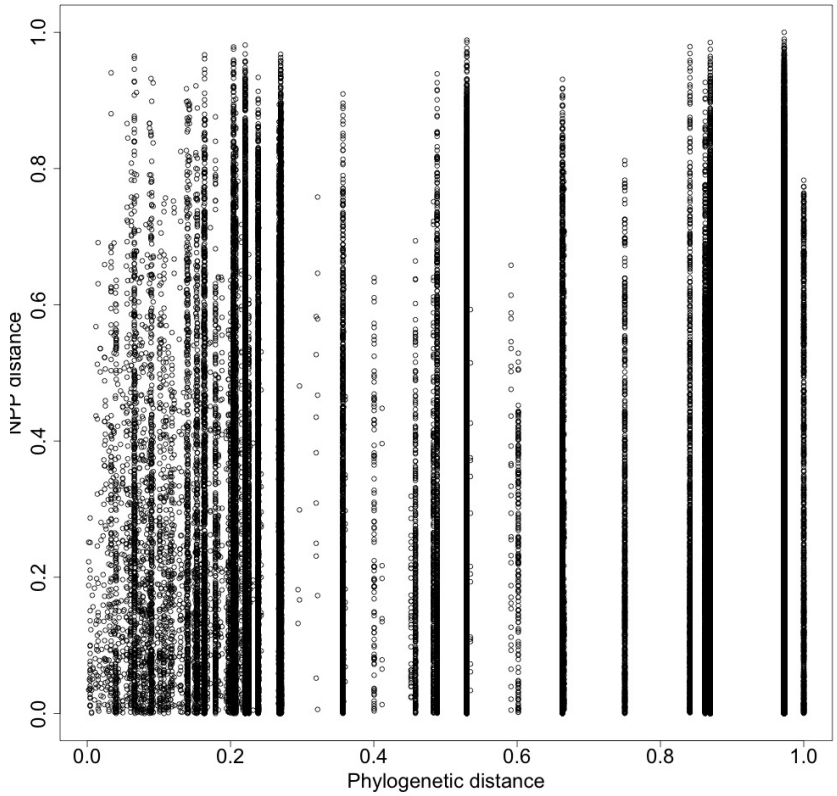

Figure 7.8: Normalised NPP distances as function of phylogenetic distances. Each pair of species i, j is represented here by a point. Proximity of points i, j and k, l in this graph does not imply that species i would be the same as k or l. Phylogenetic distances between species i and j were computed from the phylogenetic tree, and NPP distances as $|\text{NPP}_i - \text{NPP}_j|$. Both absolute NPP differences and phylogenetic distances were mapped between 0 and 1.

Table 7.8: D-statistic [28] for dental traits. This is a measure of phylogenetic signal for binary traits. Estimates were computed on binary versions of the variables. Negative values indicate that the traits contain strong phylogenetic signals. For this analysis, the non-binary traits HYP, LOP, HOD and ADI were converted to binary variables, which produces some bias. We can, however, conclude that phylogenetic signal is strong in dental traits, especially in HOD, OT and ETH.

HYP	-0.484
LOP	-0.494
HOD	-1.16
AL	-0.118
OL	-0.459
SF	-0.364
OT	-0.683
CM	-0.207
ETH	-0.970
ADI	-0.00941

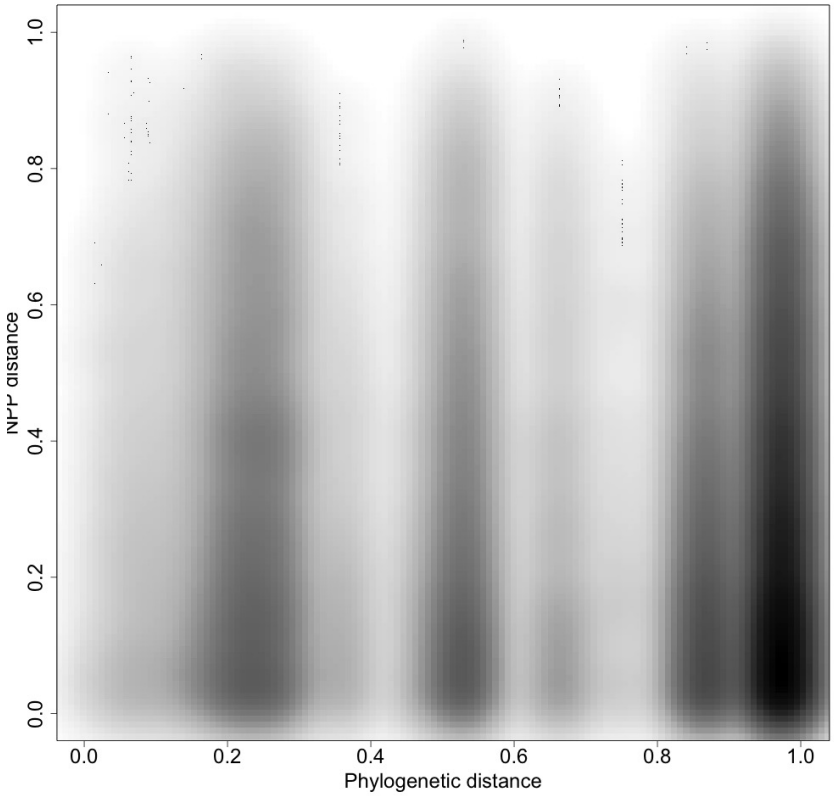

Figure 7.9: Smoothed scatterplot of normalised NPP distances and phylogenetic distances. The data in this figure is the same as in Figure 7.8. Areas of higher and lower density are clearly visible in this figure. Relatively few pairs of species are closely related, but have large NPP distances. This can be seen from the low density in the upper left corner.

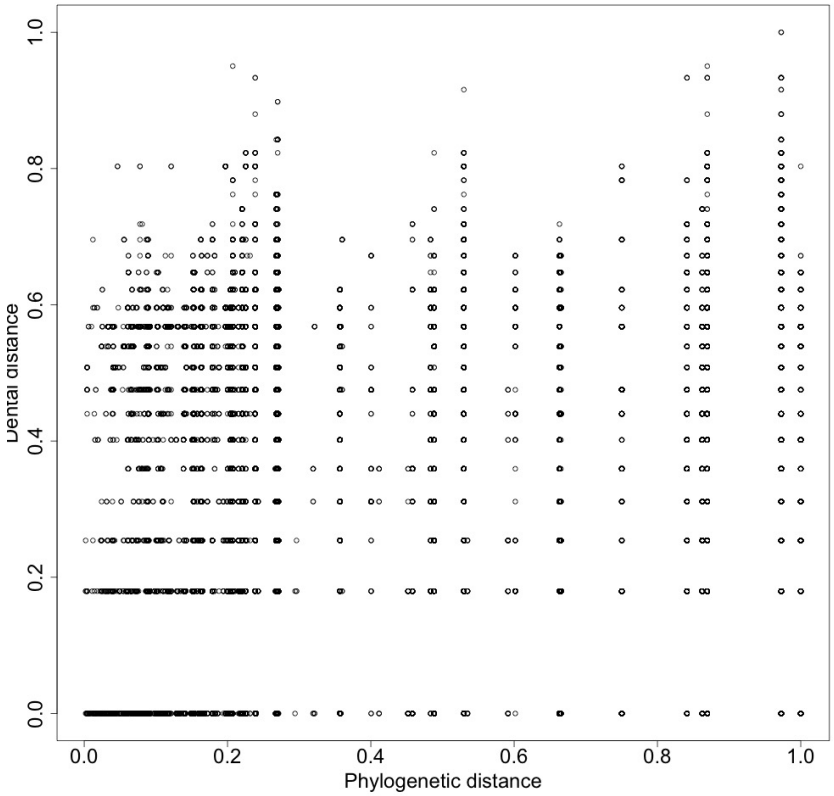

Figure 7.10: Normalised Euclidean distances of dental traits and phylogenetic distances. Both dental trait distances and phylogenetic distances are normalised to interval $[0, 1]$. All pairs of species i, j in the data are presented by a point in this figure. Proximity of points does not imply that the points would share a species.

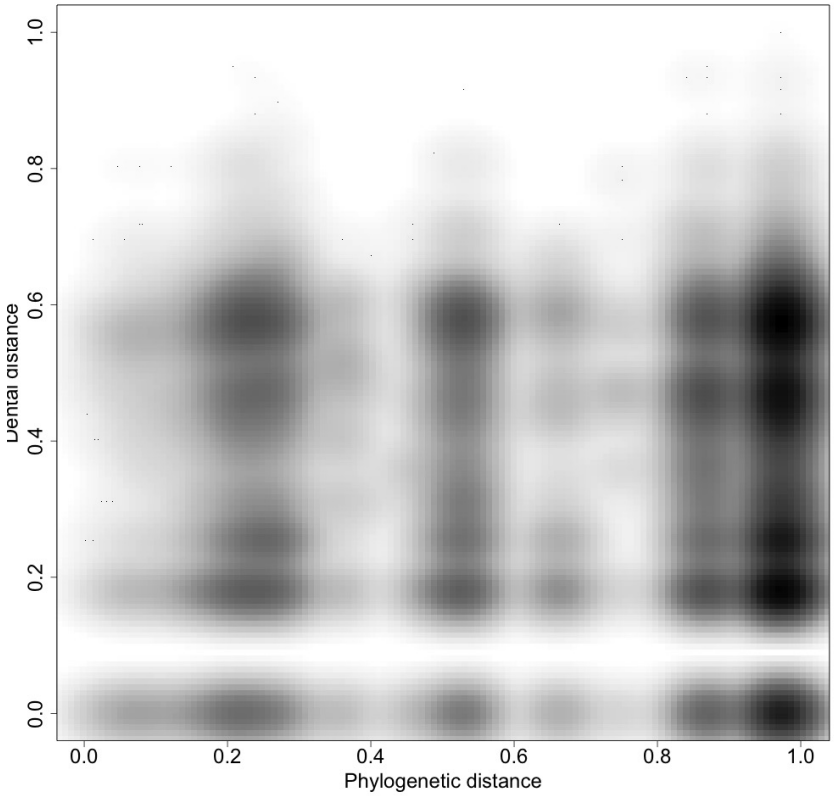

Figure 7.11: Smoothed scatterplot of normalised Euclidean distances of dental traits and phylogenetic distances. The data in this figure is the same as in Figure 7.10. Both dental trait distances and phylogenetic distances are normalised to interval $[0, 1]$. Point densities in different parts of relative dental trait space are clearly visible in this version.

cementum is NPP = 1719 − 306CM, with slope p-value 0.0338. PGLS model is NPP = 1697 + 259CM with maximum likelihood lambda value $\lambda = 0.954$, p-value of the slope is 0.0757.

Univariate OLS regression model for SF is NPP = 1724 − 181SF with slope p-value 0.0572. Corresponding PGLS model is NPP = 1743 + 261SF. The p-value of the slope is 0.01194, and maximum likelihood value for λ is 0.948.

Based on the LASSO moves, the HYP, SF and CM were chosen for a multivariate PGLS regression model, giving NPP = 2197−334 HYP+297 SF+372 CM. Corresponding multivariate OLS model for three first features from LASSO moves is NPP = 2333 − 312 HYP − 124 LOP − 156 OL. Comparison of Akaike information criterion, Bayesian information criterion and logarithmic likelihood values for some regression models on dental traits and NPP is presented in Table 7.10.

Ordinary LASSO regularisation was computed for NPP and ordinary principal components of dental traits. The first selected principal components were the first, fifth and second components (for the loadings of dental traits in components 1 - 4, see Table 7.5). Using the first principal component as input feature and NPP as output feature gave OLS regression model NPP = 2390 − 294PC1, where PC1 in the first principal component. Transforming back from the principal component to dental traits resulted in the following regression model:

$$\text{NPP} = 2390 - 150\text{HYP} - 208\text{LOP} + 0.256\text{HOD} - 5.26\text{AL}$$

$$-98.7\text{OL} - 29.1\text{SF} - 7.31\text{OT} - 10.6\text{CM} + 2.99\text{ETH} - 98.9\text{ADI}$$

NPP is plotted as a function of the first ordinary principal component in Figure 7.7, with the above regression model. Phylogenetic principal components were also used as input features in the phylogenetic LASSO procedure. The first selected components were the fourth, eight and third principal component (for the loadings in components 1 - 4, see Table 7.6).

Linear classifiers were trained and validated with classifier accuracy and effective sample size. Logistic regression models were trained with dental traits using binary NPP as the output feature and Gaussian loss function. Classifiers were trained both with phylogenetic and ordinary loss functions, using 10, 4 (HYP, LOP, SF, CM) and 2 (HYP, LOP) input features. Classification accuracies and effective sample size accuracies were computed for each classifier, the results are displayed in Table 7.11. Phylogenetic models had higher effective sample size accuracies when at least four input features were employed.

Table 7.9: Sequence of LASSO moves for dental traits and NPP. The first row shows the optimal sequence of dental traits as input features in L_1-penalised OLS regression mode for NPP as regularisation factor decreases and allows the inclusion of more input features. The second row shows the sequence of dental traits in L_1-penalised PGLS model. To compute this sequence, LASSO-regression was called with centered GLS-transformed variables. The sequences are different for OLS and PGLS regressions. Especially LOP and OL are less optimal input features for PGLS than OLS. HYP, SF and CM were among the most optimal input features for PGLS-models for most of the studied environmental variables.

OLS	HYP	LOP	OL	SF	CM	ADI	ETH	AL	OT	HOD
PGLS	HYP	SF	CM	HOD	LOP	AL	OL	ADI	ETH	OT
Step	1	2	3	4	5	6	7	8	9	10

Table 7.10: Akaike information criterion, Bayesian information criterion and logarithmic likelihood values for some predictive models for NPP. Variables for the OLS model are the first four variables from ordinary LASSO-regularised regression.

Type	Variables	AIC	BIC	LogL
OLS	HYP, LOP, OL, SF	7421	7446	-3704
PGLS	HYP, SF, CM, HOD	7291	7312	-3641
PGLS	HYP, SF, CM	7291	7308	-3642
PGLS	HYP, SF	7295	7308	-3645

Table 7.11: Classifier accuracy metrics for dental traits and binary NPP. Classification accuracy (AC) and effective sample size accuracy (EA) are shown for phylogenetic and ordinary classifiers. Binary NPP is an output variable that was constructed from NPP by assigning 1 for species with larger than median NPP, and zero otherwise. Input features for the classifiers in the middle rows were HYP, LOP, SF and CM, and HYP and LOP for the bottom rows.

Number of features	Type	AC	EA
10	Phylogenetic	0.698	0.536
10	Ordinary	0.735	0.525
4	Phylogenetic	0.690	0.530
4	Ordinary	0.731	0.522
2	Phylogenetic	0.502	0.493
2	Ordinary	0.527	0.517

7.6 Latitude

Predictive power of latitude is interesting, because it is connected to Bergmann's law [10], and can be linked to energy economy of animals. Relationship of dental traits and latitude was studied with mean and maximum absolute latitude. Maximum likelihood λ was estimated for both mean and maximum latitude, and Brownian correlation matrix was scaled with the λ-estimates. LASSO-moves for regularised PGLS are presented in Table 7.12. Based on the LASSO-moves, a PGLS model was constructed for maximum latitude: $\max \text{LAT} = -12.3 - 7.89\text{SF} + 6.91\text{HYP} - 11.1\text{CM}$, with maximum likelihood $\lambda = 0.86$.

PGLS and OLS models for logarithmic body mass and maximum latitude are shown in Figure 7.18. These models show that a positive correlation between logarithmic body mass and maximum latitude is present in this data. Bergmann's law [10] states that because of energy economy, larger species inhabit higher latitudes among closely related species. These regressions show that among large herbivorous mammals, larger species tend to inhabit higher latitudes, compared to close and distant relatives alike. The correlation found here is likely not related to energy economy, but instead caused by availability of habitats suitable for large herbivores as function of latitude.

Maximum latitude was predicted with ordinary and phylogenetic k nearest neighbour regression. Latitude was chosen for this analysis, because effective sample sizes associated with it are larger than ones associated with NPP which has stronger phylogenetic signal than latitude. Ordinary KNN was cross-validated with leave-one-out sum of square error objective. Minimum error was reached with 49 neighbours. Errors with different values of k are shown in Figure 7.14.

For phylogenetic KNN, Brownian covariance matrix was transformed with $\lambda = 0.8614453$. The effective sample size of the dataset was then 8.151891. Most diversification in the dataset phylogeny occurs late in time, which results in large covariances between many species, and further in low effective sample size values.

The algorithm was cross-validated with leave-one-out sum of square error and phylogenetic error (of the form $(\mathbf{y} - \hat{\mathbf{y}})^T \mathbf{C}^{-1} (\mathbf{y} - \hat{\mathbf{y}})$) objectives. While the algorithm uses phylogenetic information for instance selection and output value computation, the cross-validation objective SSE does not. Regression residuals from this model can have phylogenetic dependence, so the algorithm was cross-validated both with a phylogenetic and ordinary error objectives. Minimum square error was reached with effective sample size 5, and phylogenetic error with effective sample size 4. Errors with ordinary sum of square error objective are presented in Figure 7.12, and errors with phylogenetic error objective in Figure 7.13.

Maximum latitude was predicted with ordinary KNN with $k = 49$. The predictions

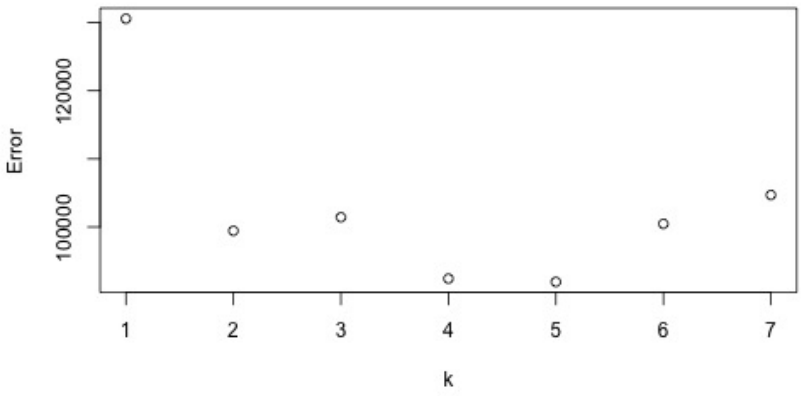

Figure 7.12: Leave-one-out cross-validation for maximum latitude with phylogenetic k nearest neighbour regression. The objective is sum of square error. Input features were dental traits, and distance metric was Euclidean distance. Effective sample size can take non-integer values, but integers were chosen here for convenience. Relatively low values of effective sample size compared to dataset size $n = 490$ result from the dependency model that was λ-transformed covariance.

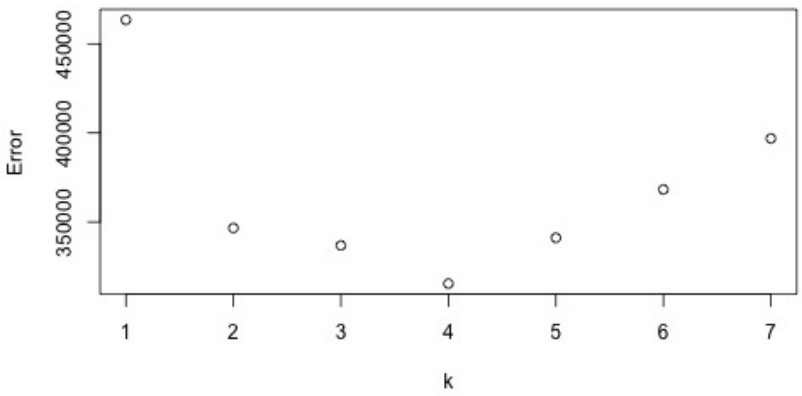

Figure 7.13: Leave-one-out cross-validation for maximum latitude with phylogenetic k nearest neighbour regression. The objective is phylogenetic error, and parameter k is effective sample size. Input features were dental traits, and distance metric was Euclidean distance. Phylogenetic k nearest neighbour regression uses phylogenetic information in instance selection and output value computation.

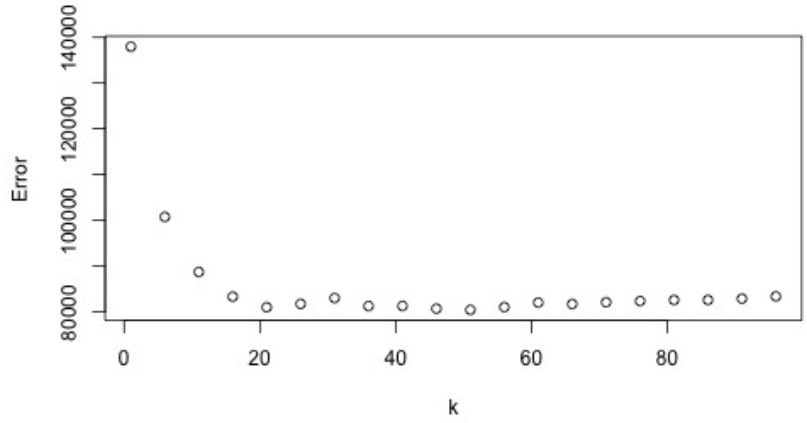

Figure 7.14: Leave-one-out cross-validation for maximum latitude with ordinary k nearest neighbour regression. The cross-validation objective is sum of square error. Input features were dental traits, and distance metric was Euclidean distance. Minimum error was reached with 49 neighbours used for latitude prediction. Here both the prediction algorithm and cross-validation objective assumed independent data.

and observations are shown in Figure 7.16. Maximum likelihood λ was estimated for the prediction residuals with value $\lambda = 0.84$. A prediction was also made with phylogenetic KNN using effective sample size $k = 4$. The predicted maximum latitude values are shown in Figure 7.17 with the observations. The residuals of phylogenetic KNN predictions have maximum likelihood λ value $\lambda = 0.86$, supporting the use of phylogenetic error validation objective. Dental data is ordinal, so it contains less variation than continuous variables. Because of this, predictions are relatively uniform. Predictions from the phylogenetic version of the algorithm vary less than predictions from ordinary KNN regression.

Phylogenetic KNN considers varying numbers of neighbours for predictions, depending on the phylogenetic structure of the local instances in feature space. Histogram of the number of neighbours with is shown in Figure 7.15. The average number was 88, median 83, and standard deviation 55. Minimum of neighbours in the predictions was 7, and maximum 185. Average number of neighbours divided by total number of instances was 0.18. Proportion of effective sample size used in the prediction was $\frac{k}{\mathrm{Ess}\{y\}} = 0.49$.

To demonstrate a neural network regression equivalent to a PGLS model, logarithmic body mass was chosen as input feature, and maximum latitude as output feature. PGLS model for latitude and logarithmic body mass with Brownian covariance is max LAT $= -19.4 + 3.98 \log M$. An equivalent neural network regression is presented in Figure 7.19. The one-layer neural network model was trained with GLS-transformed logarithmic body mass and intercept column (that is $L^{-1}\mathbf{1}$). Internal bias term was excluded in order to attribute the bias to the transformed intercept. Activation function was identity, and loss function was based on squared error. This is equivalent to an OLS model with two

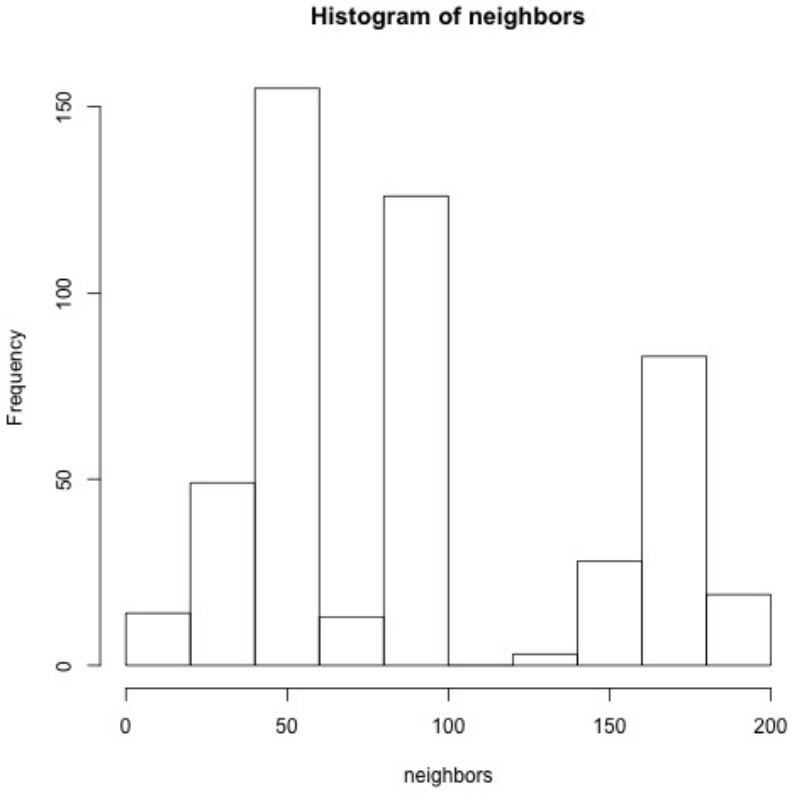

Figure 7.15: Numbers of neighbours used in phylogenetic k nearest neighbour regression for latitude with $k = 4$. In this algorithm, hyperparameter k is effective sample size, so in different parts of the predictor space, the number of neighbours that corresponds to effective sample size k varies based on local phylogenetic structure. Here the number of neighbours used for maximum latitude prediction varied between 7 and 185.

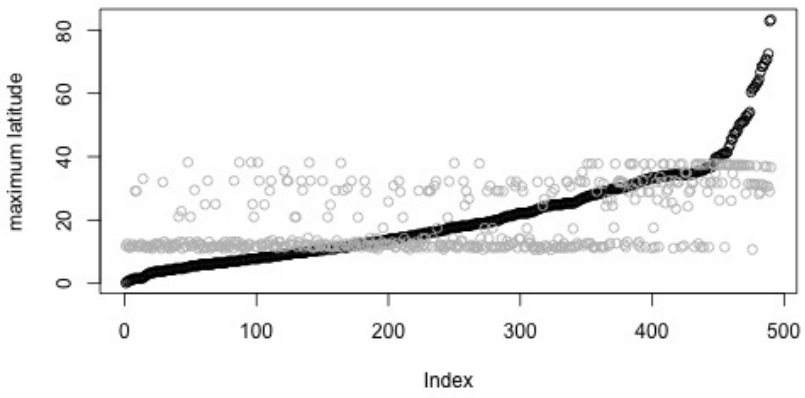

Figure 7.16: Absolute maximum latitude (black) and ordinary KNN prediction with $k = 49$ (grey). The prediction was done with the optimal value of hyperparameter k. The latitude observations are sorted in ascending order, so predictions are unsorted.

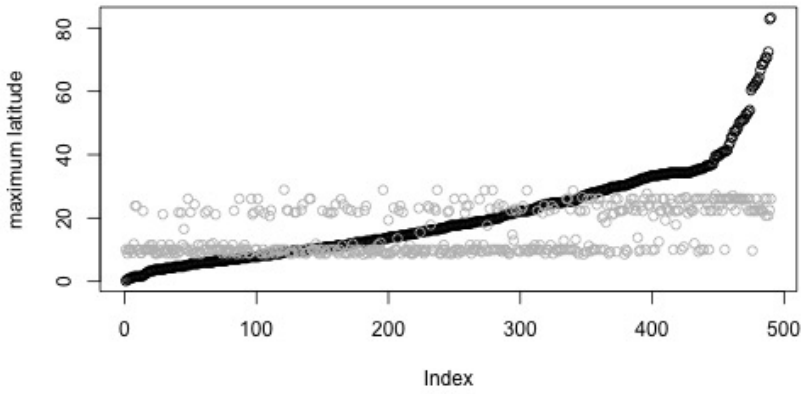

Figure 7.17: Absolute maximum latitude (black) and phylogenetic KNN predicition with effective sample size $k = 4$. The observation are sorted, and predictions unsorted. The prediction was done with hyperparameter value chosen by cross-validation with phyogenetic error objective, which is a generalisation of sum of square error.

Table 7.12: Sequence of LASSO moves for dental traits and latitude. For both regularised regression analyses, Brownian covariance matrix was transformed with maximum likelihood λ for the output variables mean latitude and maximum latitude. Regularised regression analyses was then conducted with centered GLS-transformed data. The LASSO moves show the optimal input variables in L_1-penalised PGLS regression. New variables are added as the regularisation constant decreases, until it ceases to limit the number of variables.

Variable	λ	1	2	3	4	5	6	7	8	9	10
mean latitude	0.87	OT	CM	ADI	HOD	LOP	AL	ETH	SF	HYP	OL
max latitude	0.86	SF	HYP	CM	AL	ETH	OL	LOP	ADI	HOD	OT
Step		1	2	3	4	5	6	7	8	9	10

input features and zero intercept, except here the computation was done with originally non-independent GLS-transformed variables. Linear models can be trained with these transformed variables, but non-linear models require taking the phylogeny into account in the loss function.

7.7 Temperature

Temperature was studied through mean, maximum and minimum of mean annual temperature, coldest quarter mean, coldest quarter minimum, warmest quarter mean and warmest quarter maximum temperature. Additionally, mean and maximum temperature seasonality were studied. For each temperature variable, maximum likelihood λ was estimated, and Brownian correlation matrix was scaled with the estimate. Centering

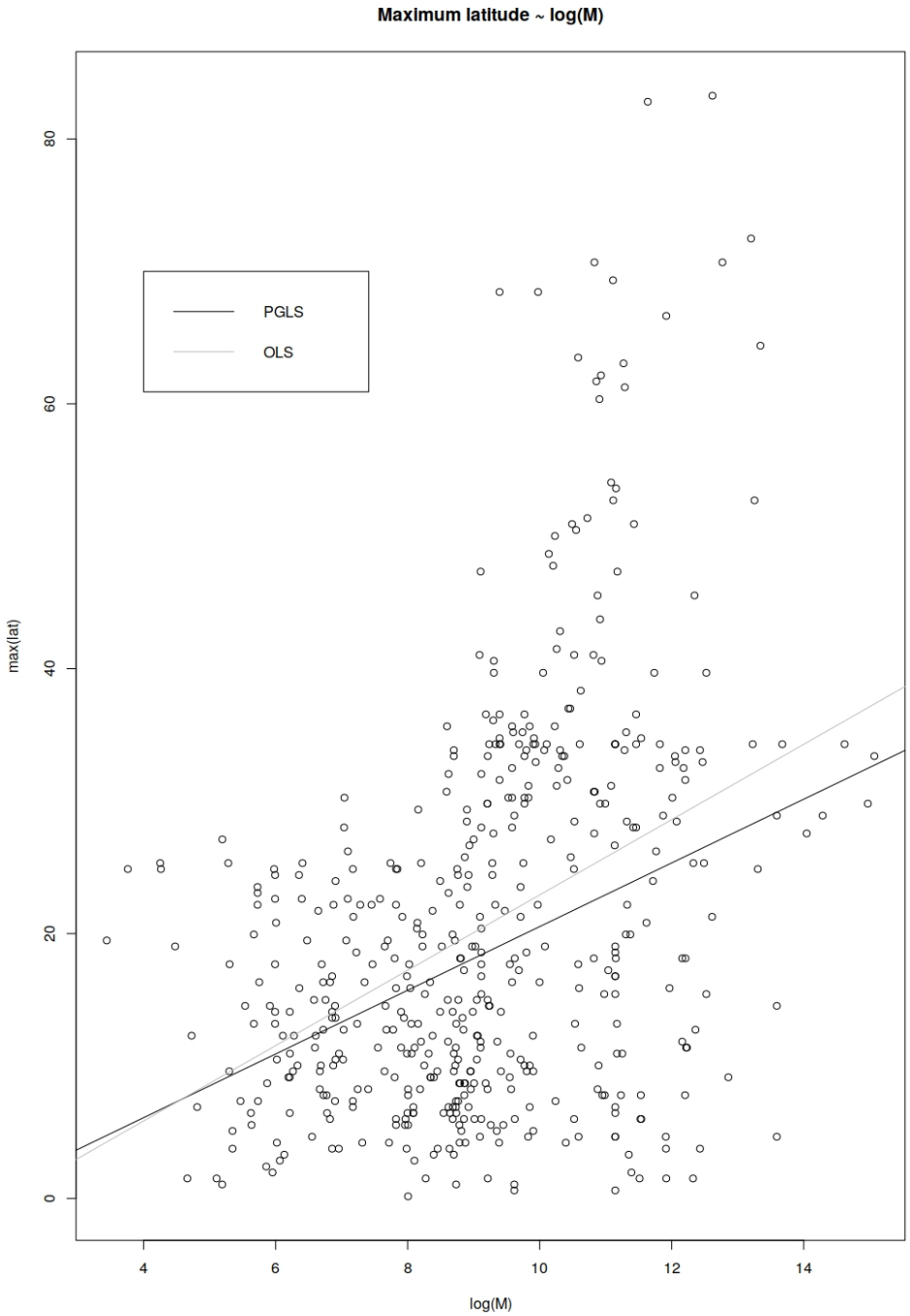

Figure 7.18: Regression of logarithmic body mass on maximum latitude. Logarithmic body mass is on the horizontal axis, and absolute maximum latitude is on the vertical axis. PGLS-model is drawn with the black line, and OLS-model with the grey line.

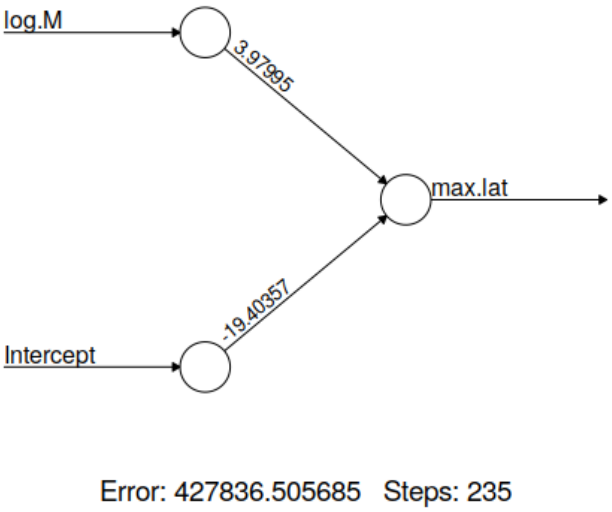

Figure 7.19: A neural network regression equivalent to PGLS model max LAT $= -19,4 + 3,98 \log M$. The training variables logarithmic body mass, intercept (vector of ones) and maximum latitude were GLS-transformed, and bias of the neural unit was set to zero.

Table 7.13: Sequence of LASSO moves for dental traits and temperature. The sequence shows the optimal order of dental traits as they are added as input variables to an L_1-penalised PGLS model when regularisation constant decreases. For each regularised regression, Brownian covariance matrix was transformed with maximum likelihood λ for the temperature quantity in question.

Variable	λ	1	2	3	4	5	6	7	8	9	10
mean MAT	0.85	SF	CM	HYP	OL	ETH	LOP	OT	AL	HOD	ADI
max MAT	0.70	SF	CM	OT	ADI	LOP	OL	HYP	AL	HOD	ETH
min MAT	0.77	SF	HYP	OL	ETH	CM	OT	HOD	AL	ADI	LOP
coldest quarter mean	0.87	SF	CM	HYP	AL	LOP	OL	ETH	OT	ADI	HOD
coldest quarter min	0.83	SF	HYP	CM	OL	ETH	OT	ADI	HOD	LOP	AL
warmest quarter mean	0.75	SF	ETH	AL	OL	CM	OT	LOP	HYP	ADI	HOD
warmest quarter max	0.53	HOD	SF	AL	HYP	ADI	OT	ETH	CM	LOP	OL
mean seasonality	0.75	HYP	CM	SF	AL	LOP	OL	ADI	OT	HOD	ETH
max seasonality	0.80	HYP	SF	AL	CM	LOP	HOD	OL	ADI	OT	ETH

GLS-transformation was performed for dental traits and temperature variables with corresponding covariance matrices. Moves for LASSO-regularised PGLS are presented in Table 7.13.

Based on the LASSO moves, SF, CM and HYP were selected for a regression model on mean annual temperature (in units 10 degrees C): $\text{MAT} = 239 + 40.2\ \text{SF} + 41.0\ \text{CM} - 17.5\ \text{HYP}$. Generally SF, CM and HYP seem to be the optimal input variables for temperature quantities.

7.8 Precipitation

The studied precipitation variables were mean annual precipitation, driest quarter mean, and wettest quarter mean precipitation. Maximum likelihood λ was computed for different precipitation measures, and Brownian covariance matrix was scaled with each λ-estimate. Centering GLS-transformation was performed on dental traits and the precipitation variables, and LASSO regularised regression was computed on the dental traits using the precipitation variables as targets. The results are displayed in Table 7.14. Generally HYP, SF and CM dominate as explanatory variables for precipitation output variables.

First LASSO moves for mean annual precipitation are the same as for NPP: HYP, SF and CM. However, the features were ordered differently in relation to driest quarter precipitation. PGLS model for driest quarter precipitation is $\text{DQP} = 221 - 33.5\ \text{HYP} + 41.3\ \text{AL} - 68.2\ \text{OL}$.

Table 7.14: Sequence of LASSO moves for dental traits and precipitation. The sequence shows the optimal order of dental traits as they are added as input variables to an L_1-penalised PGLS model. For each precipitation variable, Brownian covariance matrix was transformed with maximum likelihood λ. Mean annual precipitation shows the strongest phylogenetic signal of these three variables, while driest quarter mean precipitation shows the weakest.

Variable	λ	1	2	3	4	5	6	7	8	9	10
mean	0.85	HYP	SF	CM	OT	HOD	AL	OL	ADI	LOP	ETH
driest quarter mean	0.63	HYP	AL	OL	HOD	LOP	OT	CM	SF	ADI	ETH
wettest quarter mean	0.69	SF	HYP	CM	OT	AL	ADI	HOD	OL	ETH	LOP

8. Discussion

In synthetic data experiments, the data was created with a phylogenetic process, so the assumptions of phylogenetic signal in input and output variables were true. Based on this, it would have been expected that phylogenetic KNN produced better predictions than the ordinary KNN algorithm. However, the variance of predictions of ordinary and phylogenetic KNN was similar, even if taking phylogeny into account should decrease variance [69]. The reason for lack of variance reduction in the phylogenetic algorithm can be the changing number of neighbours in predictions, determined by local effective sample size. Another reason for this might be non-consistency of phylogenetic mean as an estimator [4]. Experimentation with variants of the algorithm and different validation objectives like phylogenetic error objective are left for future work, as well as varying evolutionary models in data generation and modelling. Experimenting with unbalanced phylogenies and different output functions is also possible.

Experiments with perceptron training in the case of unbalanced phylogeny support the proposition of the effective sample size based classification framework. When the data was sampled on random phylogenies, Ess-pocket algorithm was on average slightly more accurate in retrieving the theoretical decision surface than the ordinary perceptron, but variances were similar. However, when the phylogeny was unbalanced, the phylogenetic algorithms produced less weight variance. The directed perceptron training algorithm is expensive with many matrix inversions, but it was effective in retrieving the ground truth of synthetic data on unbalanced phylogeny. With regard to the mammalian data, i.i.d. classifiers produced higher training accuracy, and phylogenetic classifiers produced higher effective sample size accuracy, as long as there were at least four input features. With two input features, ordinary classifier was more accurate in both metrics, but the accuracies were not better than what a random classifier would produce.

All the studied dental traits show strong phylogenetic signal. Measured with D-statistic, the strongest phylogenetic signal was found in HOD, ETH and OT. D-statistic values were smaller than zero for all the dental traits, indicating phylogenetic dependency that is at least as strong as in a binary trait with an underlying continuous Brownian trait [28]. This level of phylogenetic signal in dental traits suggests phylogenetic patterns which Fritz and Purvis called clumped or extremely clumped [28].

Based on D-statistic values, one could have expected a slower saturation of trait space as a function of phylogenetic distance than what is observed in Figure 7.10. Rolshausen et al. [70] present a nearly linear characteristic trait space saturation function for Brownian motion, as opposed to exponential-like steep rise seen on the left side of Figure 7.10, which would suggest a stable optimum Ornstein-Uhlenbeck process. However, the smoothed version of the same data (Figure 7.11) shows that it is relatively rare for a pair of species to be closely related but have very different dental trait values. Also, one difference with this analysis and simulations of Rolshausen et al. is that Rolshausen discovered the characteristic saturation curves in the context of a univariate trait, but here trait distances are Euclidean distances of 10-dimensional dental trait vectors.

Some ecological features of species show a strong phylogenetic signal. Of the studied bioclimatic variables, NPP has the strongest phylogenetic signal with $\lambda = 0.952$, followed by average coldest quarter temperature with $\lambda = 0.872$. High λ-value of NPP indicates that NPP is distributed in the phylogeny of the dataset almost like a Brownian trait. Based on the phylogenetic signal in NPP, species tend to live in similar environments as their ancestors as productivity is concerned. Standard deviations of NPP, latitude, mean annual temperature and annual precipitation have weaker phylogenetic signal than the corresponding means, maxima and minima. Weakest phylogenetic signal was found in standard deviation of NPP with $\lambda = 0.392$. Weak phylogenetic signal in standard deviations implies that scale of variation in possible environments for herbivorous mammals is not strongly passed to descendant species in evolution.

Of the bioclimatic variables, NPP has the clearest connection to dental traits. Considering that NPP also had the strongest phylogenetic signal of the studied environmental features, NPP seems to affect mammal species the most of the studied variables. This is not surprising because NPP is an estimate of annual food production for herbivorous mammals.

Body mass is more correlated with latitude and temperature than NPP. In this data, body size showed only a weak connection with NPP, contrary to results of Aava [1]. The connection of latitude and body mass has been studied in the context of Bergmann's law [10], but considering that the dataset in study contained only large herbivorous mammals, the causes for the connection here might be different than in [10]. Area of land, desert, forest and grassland varies as a function of latitude, so the positive correlation between body mass and maximum latitude in OLS and PGLS regressions might be caused by availability of habitats.

Without phylogeny the data implies that structural fortification has a negative effect on NPP in univariate regression, but with phylogeny the effect is positive. The same applies for coronal cementum. This may imply that while the overall trend for NPP in SF and CM is decreasing, some clades of mammals show an increasing trend. These

kinds of patterns for differing OLS and PGLS slopes are presented in supplement of [10]. The results show that the relationship of NPP with SF and CM is more complex than previously thought.

Phylogenetic linear regression models had better values of information criteria (AIC, BIC) and likelihood than comparative ordinary linear regression models on the mammal data. Comparison of phylogenetic and i.i.d. models is best to do with likelihood or AIC, if possible. BIC contains an i.i.d. assumption [4], as obviously do ordinary squared error and accuracy statistics. If one uses i.i.d. validation statistics, phylogenetic models likely fare worse than i.i.d. models, even if their assumptions matched the data better than i.i.d. assumptions.

Study of phylogenetic signal in input and output features revealed that both dental traits and environmental features of mammal occurrence have strong phylogenetic dependencies. For linear regression models, phylogenetic signal in output features is more consequential than one in input features, although it is not equivalent to phylogenetic signal in residuals [63]. Covariance structures in NPP, latitude, temperature and precipitation models were based on maximum likelihood λ-transformations, so they were optimised for the training data. Phylogenetic and ordinary (linear) regression models are based on different assumptions of the training data, so is not surprising that a phylogenetic model fares better in a model comparison when its assumptions match the data better, and covariance is optimised in relation to the data.

9. Conclusions

We have proposed new machine learning methods and validation procedures for phylogenetic data and models. (1) Regularised PGLS as presented here offers a simple and easily interpretable way to build regression models. It can be used for feature selection or to avoid overfitting. Sequence of input variables in a LASSO-regularised PGLS model can be different from sequence of the same variables in a LASSO-regularised OLS model.

For feature selection, LASSO-regularised PGLS offers two benefits compared to phylogenetic PCA. First, the selection of features in regularised PGLS is made directly in relation to the output variable. In PCA on the other hand, the principal components do not necessarily have any connection to the target variable. Second, selected individual features are easier to interpret than principal components that are linear combinations of all features. A way to combine the two approaches is to choose principal components in phylogenetic principal component regression using LASSO-regularisation.

(2) Phylogenetic instance based regression algorithm was presented, designed to take into account local phylogenetic information in instance selection and output function. Prediction correctness and decrease of prediction variance are desirable properties of phylogenetic regression algorithms. However, decrease of prediction variance was not observed with phylogenetic KNN regression compared to the ordinary KNN algorithm. Despite these results, experiments with phylogenetic instance based methods are interesting because generative and instance based prediction is unexplored territory in phylogenetic comparative methods.

(3) Neural network regression is another family of methods that is not explored with phylogenetic data. We have presented a phylogenetic loss function for neural network regression in a general form, and shown how to recreate PGLS regression with neural networks. Non-linear feed-forward networks with phylogenetic loss can be built on these principles, increasing the available flexibility in phylogenetic discriminative regression models.

(4) A classification framework based on effective sample size was presented, including a generalisation of indicator loss, classification error and classification accuracy. Effective sample size accuracy of a ground truth decision surface was found to have less variance in resampling than ordinary classification accuracy, so the concept of effective sample size

seems to offer good basis for a phylogenetic classification framework. Effective sample size -based variants of the perceptron learning algorithm were able to learn weights closer to the ground truth with less variance than ordinary pocket perceptron, when the phylogeny was unbalanced. The quantity of effective sample size was introduced in the context of phylogenetic mean, so it relates to this estimator, but experiments with synthetic and real data indicate that it also has predictive power outside of context of phylogenetic mean.

Compared to i.i.d. models, phylogenetic models have additional complexity. To begin with, one needs to choose an evolutionary model to map phylogenies to statistical dependencies. Interpretation of phylogenetic predictive models can be more difficult than i.i.d. models because of additional evolutionary model component, some data variation attributed to phylogeny (for discussion, see Westoby [78]) and patterns on different hierarchical levels (see supplement of Clauss et al. [10]).

Model comparison can be done with likelihood and information criteria, if model likelihoods are available. Especially regarding regression models based on Gaussian cross-entropy loss, likelihood is a concept that applies regardless of specific covariance assumptions. However, phylogenetic and ordinary models are based on different assumptions, so model comparisons between them are comparisons of assumptions rather than direct modelling success. The assumptions about data should influence evaluation procedures.

The analysis revealed diverse connections between mammal dental traits and their environment. Previous research has concentrated mostly on the connection of community HYP, LOP and primary productivity. This study diversifies the use of dental traits in ecometric modelling, with the difference that here modelling is based on individual species instead of communities.

NPP seems to capture herbivore diet the best of all studied environmental features. Dental traits show only weak signal to environmental variability of geographical distributions of mammal species. Average and extreme values of bioclimatic variables generally have both clearer signal on teeth and stronger phylogenetic signal. Personal hypsodonty is a good predictor of NPP both in ordinary and phylogenetic models. HYP has consistently a negative effect on NPP in all models, unlike SF and CM which can have positive or negative effect on NPP depending if phylogeny is accounted for or not. In ordinary univariate regression SF and CM have a negative effect on NPP. In univariate PGLS or multivariate models the effect on NPP is positive. Personal LOP is a good predictor for NPP in ordinary but not phylogenetic models.

Even if two species share similar ecological conditions, they can have different evolutionary strategies. Because of phylogenetic niche conservatism, samples of related species give information about traits in animals that share similar ecology and survival strategies. Even if similarity of traits in relatives were not result of phylogenetic inertia but selection pressure in similar ecologies [78], the related strategies form a correlation between the

sample species.

Whether phylogenetic comparative methods assign too much of the trait variation on phylogeny and too little on environment is out of scope for this study. In light of the phylogenetic signals in the bioclimatic variables, species tend to live in similar environments as their relatives. Does it matter if phylogenetic signal arises from inertia or similarities in environment? The methods do not make assumptions about the mechanics of the phylogenetic signal. In causality modelling [71] phylogenetic inertia vs. shared environment might be different situations, while statistical models do not make this distinction.

Most procedures in machine learning contain some explicit or implicit i.i.d. assumptions. These include methods, training algorithms and validation procedures. We have presented some procedures which assume that the data comes from a phylogenetic process. Be it i.i.d. or phylogenetic methods, in predictive modelling we aim to find structures that stay constant in resampling. In phylogenetic comparative methods, this is counter-intuitive because resampling would mean rerunning the evolutionary process. For the studied herbivorous mammals, development starting from the common ancestor has taken 166 million years, and it cannot be redone. Current phylogenetic comparative methods have their critics, but in mainstream comparative biology they are considered to give better generalisations of phylogenetic data than i.i.d. models.

For models based on phylogenetic data, one should look beyond individual instances and think about information in general, for example with the concept of effective sample size. Then one is able to build generalisations and conduct statistical learning on the processes that generated the data.

Bibliography

[1] B. Aava. Primary productivity can affect mammalian body size frequency distributions. *Oikos*, 93(2):205–212, 2001.

[2] D. C. Adams and M. L. Collyer. Multivariate phylogenetic comparative methods: evaluations, comparisons, and recommendations. *Systematic Biology*, 67(1):14–31, 2017.

[3] D. C. Adams and R. N. Felice. Assessing trait covariation and morphological integration on phylogenies using evolutionary covariance matrices. *PloS one*, 9(4):e94335, 2014.

[4] C. Ané et al. Analysis of comparative data with hierarchical autocorrelation. *The Annals of Applied Statistics*, 2(3):1078–1102, 2008.

[5] R. J. Barnes. Matrix differentiation, 2006. https://atmos.washington.edu/~dennis/MatrixCalculus.pdf.

[6] O. R. Bininda-Emonds, M. Cardillo, K. E. Jones, R. D. MacPhee, R. M. Beck, R. Grenyer, S. A. Price, R. A. Vos, J. L. Gittleman, and A. Purvis. The delayed rise of present-day mammals. *Nature*, 446(7135):507, 2007.

[7] S. P. Blomberg, J. G. Lefevre, J. A. Wells, and M. Waterhouse. Independent contrasts and pgls regression estimators are equivalent. *Systematic Biology*, 61(3):382–391, 2012.

[8] M. A. Butler and A. A. King. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist*, 164(6):683–695, 2004.

[9] M. A. Butler, T. W. Schoener, and J. B. Losos. The relationship between sexual size dimorphism and habitat use in greater antillean anolis lizards. *Evolution*, 54(1):259–272, 2000.

[10] M. Clauss, M. T. Dittmann, D. W. Müller, C. Meloro, and D. Codron. Bergmann' s rule in mammals: a cross-species interspecific pattern. *Oikos*, 122(10):1465–1472, 2013.

[11] J. Clavel, L. Aristide, and H. Morlon. A penalized likelihood framework for high-dimensional phylogenetic comparative methods and an application to new-world monkeys brain evolution. *Systematic biology*, 68(1):93–116, 2018.

[12] T. J. Davies, J. Regetz, E. M. Wolkovich, and B. J. McGill. Phylogenetically weighted regression: A method for modelling non-stationarity on evolutionary trees. *Global ecology and biogeography*, 28(2):275–285, 2019.

[13] P. de Villemereuil, J. A. Wells, R. D. Edwards, and S. P. Blomberg. Bayesian models for comparative analysis integrating phylogenetic uncertainty. *BMC evolutionary biology*, 12(1):102, 2012.

[14] J. A. F. Diniz-Filho, L. M. Bini, T. F. Rangel, I. Morales-Castilla, M. Á. Olalla-Tárraga, M. Á. Rodríguez, and B. A. Hawkins. On the selection of phylogenetic eigenvectors for ecological analyses. *Ecography*, 35(3):239–249, 2012.

[15] J. A. F. Diniz-Filho, C. E. R. de Sant'Ana, and L. M. Bini. An eigenvector method for estimating phylogenetic inertia. *Evolution*, 52(5):1247–1262, 1998.

[16] J. Eronen, K. Puolamäki, L. Liu, K. Lintulaakso, J. Damuth, C. Janis, and M. Fortelius. Precipitation and large herbivorous mammals i: estimates from present-day communities. *Evolutionary Ecology Research*, 12(2):217–233, 2010.

[17] J. T. Eronen, P. D. Polly, M. Fred, J. Damuth, D. C. Frank, V. Mosbrugger, C. Scheidegger, N. C. Stenseth, and M. Fortelius. Ecometrics: the traits that bind the past and present together. *Integrative Zoology*, 5(2):88–101, 2010.

[18] J. Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15, 1985.

[19] J. Felsenstein. Using the quantitative genetic threshold model for inferences between and within species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1427–1434, 2005.

[20] J. Felsenstein. A comparative method for both discrete and continuous characters using the threshold model. *The American Naturalist*, 179(2):145–156, 2012.

[21] S. E. Fick and R. J. Hijmans. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, 37(12):4302–4315, 2017.

[22] N. M. Foley, M. S. Springer, and E. C. Teeling. Mammal madness: is the mammal tree of life not yet resolved? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1699):20150140, 2016.

[23] M. Fortelius, I. Žliobaitė, F. Kaya, F. Bibi, R. Bobe, L. Leakey, M. Leakey, D. Patterson, J. Rannikko, and L. Werdelin. An ecometric analysis of the fossil mammal record of the turkana basin. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1698):20150232, 2016.

[24] R. Freckleton. The seven deadly sins of comparative analysis. *Journal of evolutionary biology*, 22(7):1367–1375, 2009.

[25] R. P. Freckleton, P. H. Harvey, and M. Pagel. Phylogenetic analysis and comparative data: a test and review of evidence. *The American Naturalist*, 160(6):712–726, 2002.

[26] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[27] S. A. Fritz, O. R. Bininda-Emonds, and A. Purvis. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecology letters*, 12(6):538–549, 2009.

[28] S. A. Fritz and A. Purvis. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conservation Biology*, 24(4):1042–1051, 2010.

[29] J. A. Fuentes-G, P. D. Polly, and E. P. Martins. A bayesian extension of phylogenetic generalized least squares (pgls): incorporating uncertainty in the comparative study of trait relationships and evolutionary rates. *Evolution*, 2019.

[30] E. Galbrun, H. Tang, M. Fortelius, and I. Žliobaitė. Computational biomes: The ecometrics of large mammal teeth. *Palaeontologia Electronica*, 21(21.1.3A):1–31, 2018.

[31] S. I. Gallant. Perceptron-based learning algorithms. *IEEE Transactions on neural networks*, 50(2):179, 1990.

[32] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):44, 2014.

[33] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[34] A. Grafen. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 326(1233):119–157, 1989.

[35] D. J. Hand. Classifier technology and the illusion of progress. *Statistical science*, pages 1–14, 2006.

[36] T. F. Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351, 1997.

[37] T. F. Hansen, J. Pienaar, and S. H. Orzack. A comparative method for studying adaptation to a randomly evolving environment. *Evolution: International Journal of Organic Evolution*, 62(8):1965–1977, 2008.

[38] L. J. Harmon, J. B. Losos, T. Jonathan Davies, R. G. Gillespie, J. L. Gittleman, W. Bryan Jennings, K. H. Kozak, M. A. McPeek, F. Moreno-Roark, T. J. Near, et al. Early bursts of body size and shape evolution are rare in comparative data. *Evolution: International Journal of Organic Evolution*, 64(8):2385–2396, 2010.

[39] T. Hastie and B. Efron. *lars: Least Angle Regression, Lasso and Forward Stagewise*, 2013. R package version 1.2.

[40] L. S. T. Ho and C. Ane. A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Systematic Biology*, 63:397–408, 2014.

[41] J. Hoffman, M. Mohri, and N. Zhang. Algorithms and theory for multiple-source adaptation. In *Advances in Neural Information Processing Systems*, pages 8246–8256, 2018.

[42] T. Ingram and D. L. Mahler. Surface: detecting convergent evolution from comparative data by fitting ornstein-uhlenbeck models with stepwise akaike information criterion. *Methods in Ecology and Evolution*, 4(5):416–425, 2013.

[43] A. R. Ives and T. Garland. Phylogenetic regression for binary dependent variables. In *Modern phylogenetic comparative methods and their application in evolutionary biology*, pages 231–261. Springer, 2014.

[44] A. R. Ives and T. Garland Jr. Phylogenetic logistic regression for binary dependent variables. *Systematic biology*, 59(1):9–26, 2009.

[45] J. Jiang. A literature survey on domain adaptation of statistical classifiers. *URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey*, 3:1–12, 2008.

[46] K. E. Jones, J. Bielby, M. Cardillo, S. A. Fritz, J. O'Dell, C. D. L. Orme, K. Safi, W. Sechrest, E. H. Boakes, C. Carbone, et al. Pantheria: a species-level database of life history, ecology, and geography of extant and recently extinct mammals: Ecological archives e090-184. *Ecology*, 90(9):2648–2648, 2009.

[47] K. Lintulaakso. Mammalbase – database of recent mammals, 2013. http://www.mammalbase.net.

[48] L. Liu, K. Puolamäki, J. T. Eronen, M. M. Ataabadi, E. Hernesniemi, and M. Fortelius. Dental functional traits of mammals resolve productivity in terrestrial ecosystems past and present. *Proceedings of the Royal Society B: Biological Sciences*, 279(1739):2793–2799, 2012.

[49] E. P. Martins and T. F. Hansen. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149(4):646–667, 1997.

[50] J. F. Monahan. *Numerical methods of statistics*. Cambridge University Press, 2011.

[51] O. Oksanen, I. Žliobaitė, J. Saarinen, A. M. Lawing, and M. Fortelius. A humboldtian approach to life and climate of the geological past: estimating palaeotemperature from dental traits of mammalian communities. *Journal of Biogeography*, 46(8):1760–1776, 2019.

[52] D. Orme, R. Freckleton, G. Thomas, T. Petzoldt, S. Fritz, N. Isaac, and W. Pearse. *caper: Comparative Analyses of Phylogenetics and Evolution in R*, 2018. R package version 1.0.1.

[53] M. Pagel. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 255(1342):37–45, 1994.

[54] M. Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877, 1999.

[55] M. Pagel and A. Meade. Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. *The American Naturalist*, 167(6):808–825, 2006.

[56] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[57] E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528, 2018.

[58] S. E. Pav. *madness: Automatic Differentiation of Multivariate Operations*, 2019. R package version 0.2.6.

[59] M. W. Pennell and L. J. Harmon. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Annals of the New York Academy of Sciences*, 1289(1):90–105, 2013.

[60] A. Quinn. When is a cladist not a cladist? *Biology & Philosophy*, 32(4):581–598, 2017.

[61] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[62] L. J. Revell. Size-correction and principal components for interspecific comparative studies. *Evolution: International Journal of Organic Evolution*, 63(12):3258–3268, 2009.

[63] L. J. Revell. Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution*, 1(4):319–329, 2010.

[64] L. J. Revell. phytools: An r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3:217–223, 2012.

[65] L. J. Revell and L. J. Harmon. Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. *Evolutionary Ecology Research*, 10(3):311–331, 2008.

[66] L. J. Revell and A. S. Harrison. Pcca: a program for phylogenetic canonical correlation analysis. *Bioinformatics*, 24(7):1018–1020, 2008.

[67] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017.

[68] F. J. Rohlf. Comparative methods for the analysis of continuous variables: geometric interpretations. *Evolution*, 55(11):2143–2160, 2001.

[69] F. J. Rohlf. A comment on phylogenetic correction. *Evolution*, 60(7):1509–1515, 2006.

[70] G. Rolshausen, T. J. Davies, and A. P. Hendry. Evolutionary rates standardized for evolutionary space: perspectives on trait evolution. *Trends in ecology & evolution*, 33(6):379–389, 2018.

[71] B. Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[72] S. Shekhar, M. R. Evans, J. M. Kang, and P. Mohan. Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):193–214, 2011.

[73] T. Therneau and B. Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2019. R package version 4.1-15.

[74] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[75] M. R. Tolkoff, M. E. Alfaro, G. Baele, P. Lemey, and M. A. Suchard. Phylogenetic factor analysis. *Systematic biology*, 67(3):384–399, 2017.

[76] V. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

[77] R. Wehrens and J. Kruisselbrink. Flexible self-organizing maps in kohonen 3.0. *Journal of Statistical Software*, 87(7):1–18, 2018.

[78] M. Westoby, M. R. Leishman, and J. M. Lord. On misinterpreting thephylogenetic correction'. *Journal of Ecology*, 83(3):531–534, 1995.

[79] I. Žliobaitė, A. Bifet, J. Read, B. Pfahringer, and G. Holmes. Evaluation methods and decision theory for classification of streaming data with temporal dependence. *Machine Learning*, 98(3):455–482, 2015.

[80] I. Žliobaitė, J. Rinne, A. B. Tóth, M. Mechenich, L. Liu, A. K. Behrensmeyer, and M. Fortelius. Herbivore teeth predict climatic limits in kenyan ecosystems. *Proceedings of the National Academy of Sciences*, 113(45):12751–12756, 2016.