Journal of Theoretical Biology 407 (2016) 371–386

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Phylogenetic effective sample size

Krzysztof Bartoszek

Department of Mathematics, Uppsala University, Uppsala 751 06, Sweden

HIGHLIGHTS

- A way to quantify the number of independent signal in a phylogenetic data set.
- Can be extended to non-normal models of trait evolution.
- Can be used for model selection and applied for quantifying biodiversity.
- Can be used to assess importance of clade and phylogenetic inertia.
- The R software package mvSLOUCH is extended with the pESS.

ARTICLE INFO

Article history:
Received 25 July 2015
Received in revised form
11 May 2016
Accepted 18 June 2016
Available online 21 June 2016

MSC:
62B10
62P10
92-08
92B10
92B15
94A17

Keywords:
Biodiversity
Effective sample size
Measurement error
Ornstein–Uhlenbeck process
Phylogenetic comparative methods
Quantitative trait evolution

ABSTRACT

In this paper I address the question—*how large is a phylogenetic sample?* I propose a definition of a phylogenetic effective sample size for Brownian motion and Ornstein–Uhlenbeck processes—the *regression effective sample size*. I discuss how mutual information can be used to define an effective sample size in the non-normal process case and compare these two definitions to an already present concept of effective sample size (the mean effective sample size). Through a simulation study I find that the AIC_c is robust if one corrects for the number of species or effective number of species. Lastly I discuss how the concept of the phylogenetic effective sample size can be useful for biodiversity quantification, identification of interesting clades and deciding on the importance of phylogenetic correlations.

1. Introduction

One of the reasons to introduce phylogenetic comparative methods (PCMs) in the words of Martins and Hansen (1996) was to address the problem of statistical dependence. They called the issue the "degrees of freedom" or "effective sample size" problem. If we have n species related by a phylogenetic tree, unless it is a star phylogeny, then our effective sample size is less than n (in extreme cases even one). Taking into consideration the number of independent observations is important in evaluating the accuracy of parameter estimation or hypothesis tests. The performance of such statistical procedures depends on the number of independent data points and not on the observed number of data points (Martins and Hansen, 1996). Ignoring the correlations (and hence inflating the sample size) results in too narrow confidence intervals, inflated p-values and power. All of this leads to type I and II errors of which the user may be oblivious of.

In a phylogenetic context the calculation of the effective number of observations has not been often addressed directly. In statistical literature effective sample size (ESS) is usually parameter specific, it can be understood as "the number of independent measurements one would need to reach the same amount of information about a parameter as in the original data" (Faes et al., 2009)—in other words how many independent points

E-mail address: bartoszekkj@gmail.com

http://dx.doi.org/10.1016/j.jtbi.2016.06.026
0022-5193/

do we have for estimating a particular parameter. Nunn (2011, p. 145) points out that often phylogenetic comparative methods have been viewed in a restricted manner as a "degrees of freedom" correction procedure that "reduce the number of data points", due to the nonindependence. Most phylogenetic comparative methods work in the following way—one assumes a model and maximizes the likelihood under that model. Hence, the issue of ESS, as mentioned above, has been taken care of but only for the estimation problem. In other situations, as Nunn (2011) following Pagel (1993) reminds, the "degrees of freedom analogy can be misleading". It is more important how the variance is partitioned among species. In fact, in the case of model selection, or when one wants to know how many "independent" taxa one has e.g. for conservation purposes the situation becomes much more complex. As we will see, it is more important how the covariance is structured.

Smith (1994) directly approached the problem of effective sample size. He studied interspecies phenotypic data by a nested ANOVA and "*Determination of the taxonomic levels that account for most of the variation can be used to select a single level at which it is most reasonable to consider the data points as independent*". From the perspective of modern phylogenetic comparative methods this is a "hack", as Smith (1994) himself wrote "the method improves the nonindependence problem but does not eliminate it". From our perspective his work is important, as from the nested ANOVA setup, he partitioned the variance into components from different levels of the phylogeny and then defined the effective sample size as

$$\begin{aligned} n_e = {} & (\#\text{of superfamilies})(\text{PVC for superfamilies}) \\ & + (\#\text{of families})(\text{PVC for families}) \\ & + (\#\text{of genera})(\text{PVC for genera}) \\ & + (\#\text{of species})(\text{PVC for species}) \end{aligned} \quad (1)$$

where PVC is percentage of variance component. Smith (1994) importantly notices that in principle "*The method does not require that levels of the nested hierarchy are defined by taxonomic categories.*" In this work I develop the idea described in Smith (1994)'s own words: to "*consider each species as some fraction of a free observation varying between 0 and 1.0, a value could be computed … that would reflect the balance between constraint and independent evolution. This value is defined as the effective sample size (effective N) for the data set and trait, as opposed to the traditionally used observed sample size (observed N).*" Building up on the modern development of stochastic models for phylogenetic comparative methods, I do not have to restrict myself to partitioning the data into hierarchical levels containing different fractions of the variance, but rather look holistically at the dependence pattern induced by the tree and model of evolution. This might make it impossible (but maybe not always) to assign to each species (or taxonomic level) its fraction of free observations but as we shall see it will allow me to calculate the sum of fractions of free observations.

An analysis of phylogenetically structured phenotypic data often has as its goal to identify the mode of evolution, i.e. is the trait (s) adapting (and if so to what trait/phenotype) or rather exhibiting neutral evolution. Information criteria like the Akaike Information Criterion (AIC Akaike, 1974), Akaike Information Criterion corrected for small sample size (AIC_c Hurvich and Tsai, 1989) or Bayesian Information Criterion (BIC Schwarz, 1978) are commonly used to identify the model better supported by the data. However, if one goes back to the derivation of the AIC_c (Hurvich and Tsai, 1989) and BIC (Schwarz, 1978) one can see that the n observations are assumed independent. Therefore a phylogenetic comparative model seems to violate this assumption, in the best case by inflating the sample size. In a way such an inflation corresponds to not penalizing enough for additional parameters. However in their original paper Hurvich and Tsai (1989) derive the same AIC_c formula for autoregressive models so this warrants further study in the phylogenetic setting where the covariance structure is hierarchical.

Therefore, using the number of species (unless the phylogeny is a star) results in a risk of overfitting for small phylogenies or those with most speciation events near the tips. In this work I propose a way of taking into account the *effective number of species* during the model selection procedure. The newest version of mvSLOUCH (available from http://cran.r-project.org/web/packages/mvSLOUCH/index.html) allows for automatic model selection if one treats n as the true sample size and also if one corrects for the dependencies using an effective sample size. Importantly, mvSLOUCH allows for an arbitrary pattern of missing data—no observation is removed and the likelihood is based on all provided information. Using this new version of mvSLOUCH, I include in this work a simulation study and analyze a number of data sets to see how much a difference does it make whether, one uses the observed or effective number of species for model selection. In most cases, the two ways of counting species lead to the same conclusion. However, for small samples (see Table 3) using the effective number of species can result in a different outcome. In fact, we should expect this to be so, a good correction method should be robust—with enough observations the data (or rather likelihood) should decide no matter how one corrects. It is only with few observations (and hence little power) that correction methods should play a role by pointing to different possibilities of interpreting the observed data.

2. Effective sample size

Effective sample size is intuitively meant to represent the number of independent particles of data in the sample. If the sample is correlated, then each observation will only have a certain fraction of the information it carries particular to itself. The rest of the information will be shared with one/some/all other points in the sample. We would like to quantify what proportion of the whole sample is made up of these independent bits of information. If this proportion is p, then our phylogenetic effective sample size (pESS) will be $n_e = pn$. However our situation is a bit different. It is reasonable to assume that we have at least one observation—at least one species described by at least a single trait. One way is to define p to be between 1 and $1/n$. Alternatively we can define as

$$n_e = 1 + p(n - 1), \quad (2)$$

where $p \in [0, 1]$. I will call this p of Eq. (2) the phylogenetic ESS factor. The value n_e/n is useful in practice to compare between different sized phylogenies and I will call it the relative phylogenetic ESS.

Martins and Hansen (1996) point out, that in the discrete trait case, the ESS cannot be greater than the number of independent evolutionary changes regardless of the number of observed species. Maddison and FitzJohn (2015) very recently remind us of this again. Phylogenetic comparative methods are there to take care of "pseudoreplicates" due to the tree induced correlations. However, especially in the discrete case, tests of significance might have inflated power as one uses the number of species instead of the (unknown) number of independent evolutionary changes. Unfortunately, at the moment, there does not seem to be any solution for this problem (Maddison and FitzJohn, 2015). Hopefully the phylogenetic effective sample size concept presented here could indicate a direction for finding one. An alternative potential approach in the discrete case, is phylogenetic informativeness based on the number of mutations (i.e. changes) shared by tip taxa under the Poisson process (Mulder and Crawford, 2015; Townsend,