Phylogenetic comparative methods: applying modern probabilistic methods to evolutionary biology

Krzysztof Bartoszek *

January 31, 2015

- Lecture 1 "Introduction to phylogenetic comparative methods, a probabilistic and statistical problem introduced by biologists"

- Lecture 2 "Branching diffusion processes — a mathematicians view on phylogenetic comparative methods: asymptotic and inference questions"

Since Darwin biologists have recognized the need for a specialized mathematical apparatus to study inter–species measurements. However it was only Felsenstein [1985]'s independent contrasts method [although similar models were discussed earlier, e.g. Edwards, 1970] that proposed an appropriate framework for correctly analyzing trait data obtained from different taxonomic units. What makes such data different from the usual independent, identically distributed sample setting is that species level measurements are not independent. Due to a shared evolutionary history one can easily observe that more recently diverged species tend to be similar. Therefore using standard statistical methodologies, will make it impossible to distinguish between similarities due to shared environment or trait function (selection and adaptation) and phylogenetic inertia.

To remedy this situation Felsenstein [1985] assumed that a trait evolves as a Brownian motion on top of a phylogenetic tree. Along a branch we have usual Brownian motion evolution until a speciation point is reached. After

---

*Department of Mathematics, Uppsala University, Uppsala, Sweden, bartoszekkj@gmail.com

speciation along the daughter branches two independent Brownian motions start running. Their initial values are the value of the process just before the speciation event. Very quickly this method (called independent contrasts in the biological community) became a field standard.

However Felsenstein [1988], Hansen [1997] pointed out that a Brownian motion model is not appropriate for traits under stabilizing selection — a pure Brownian motion has no stationary distribution. Therefore they proposed an Ornstein–Uhlenbeck model. Only now due to the increase of computational power is this process taking its place in an evolutionary biologist's toolbox [Butler and King, 2004, Labra et al., 2009, Hansen and Orzack, 2005, Hansen et al., 2008]. This framework is currently undergoing rapid expansion allowing for measurement error [Felsenstein, 2008, Hansen and Bartoszek, 2012, Rohlfs et al., 2013], multiple interacting traits [Bartoszek et al., 2012] or different drift and diffusion parameters for different clades [Beaulieu et al., 2012].

Usually a phylogenetic comparative method will assume that the tree describing the relationships between species is fully resolved. This however need not always be the case. We do not know all the currently alive species, they are often reclassified or they are undergoing speciation. What may happen also in many insect orders is that molecular based tree inference methods can resolve the tree on the family level but there is not enough information to say much about the clades' subtrees. Hence there is also a need for tree–free methods. These methods only assume a branching process model conditioned on the number of contemporary tips for the phylogenetic tree. Of course with an unknown tree we can infer much less about the process driving the phenotype's evolution. However we still may infer a remarkable amount of information and also gain new insights into the species' evolution.

Using this tree–free framework we may study the expected similarity of species and how quickly they loose it via the so–called interspecies correlation coefficient [Bartoszek, 2014, Bartoszek and Sagitov, 2015, Mulder and Crawford, 2015, Sagitov and Bartoszek, 2012]. We may also derive phylogenetic confidence intervals for the optimal trait values and effectively estimate the stationary variance of an adapting trait [Bartoszek and Sagitov, 2015] or diffusion parameter of a Brownian motion one [Bartoszek and Sagitov, 2014, Crawford and Suchard, 2013]. These methods also allow one to observe a phase transition in phylogenetic inertia. If the speciation process is fast enough then ancestral dependencies have an effect for long stretches [Adamczak and Miłoś, 2011, in press, Ané et al., 2014, Bartoszek and Sagitov, 2015]

In the first lecture hour I will introduce the field of phylogenetic comparative methods, discuss the biological motivation and mathematical approach of using diffusion type stochastic differential equations to model inter–species data. I will discuss multivariate and measurement error extensions to these methods and illustrate why these are important. The second lecture hour will be more theoretical concentrating on tree–free methods. I will discuss their probabilistic background and what model parameters can be inferred using them.

References

R. Adamczak and P. Miłoś. CLT for Ornstein–Uhlenbeck branching particle system. *ArXiv e-prints*, 2011.

R. Adamczak and P. Miłoś. U–statistics of Ornstein–Uhlenbeck branching particle system. *J. Th. Probab.*, in press.

C. Ané, L. S. T. Ho, and S. Roch. Phase transition on the convergence rate of parameter estimation under an Ornstein–Uhlenbeck diffusion on a tree. *ArXiv e-prints*, 2014.

K. Bartoszek. Quantifying the effects of anagenetic and cladogenetic evolution. *Math. Biosci.*, 254:42–57, 2014.

K. Bartoszek and S. Sagitov. A consistent estimator of the evolutionary rate. *ArXiv e-prints*, 2014. Accepted *J. Theor. Biol.*

K. Bartoszek and S. Sagitov. Phylogenetic confidence intervals for the optimal trait value. *J. App. Prob.*, 52(4), 2015.

K. Bartoszek, J. Pienaar, P. Mostad, S. Andersson, and T. F. Hansen. A phylogenetic comparative method for studying multivariate adaptation. *J. Theor. Biol.*, 314:204–215, 2012.

J. M. Beaulieu, D. C. Jhwueng, C. Boettiger, and B. C. O'Meara. Modeling stabilizing selection: Expanding the Ornstein–Uhlenbeck model of adaptive evolution. *Evolution*, 66:2369–2389, 2012.

M. A. Butler and A. A. King. Phylogenetic comparative analysis: a modelling approach for adaptive evolution. *Am. Nat.*, 164(6):683–695, 2004.

F. W. Crawford and M. A. Suchard. Diversity, disparity, and evolutionary rate estimation for unresolved Yule trees. *Syst. Biol.*, 62(3):439–455, 2013.

A. W. F. Edwards. Estimation of the branch points of a branching diffusion process. *J. Roy. Stat. Soc. B*, 32(2):155–174, 1970.

J. Felsenstein. Phylogenies and the comparative method. *Am. Nat.*, 125(1): 1–15, 1985.

J. Felsenstein. Phylogenies and quantitative characters. *Annu. Rev. Ecol. Syst.*, 19:445–471, 1988.

J. Felsenstein. Comparative methods with sampling error and within-species variation: contrasts revisited and revised. *Am. Nat.*, 171:713–725, 2008.

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

T. F. Hansen and K. Bartoszek. Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. *Syst. Biol.*, 61(3):413–425, 2012.

T. F. Hansen and S. H. Orzack. Assessing current adaptation and phylogenetic inertia as explanations of trait evolution: the need for controlled comparisons. *Evolution*, 59(10):2063–2072, 2005.

T. F. Hansen, J. Pienaar, and S. H. Orzack. A comparative method for studying adaptation to a randomly evolving environment. *Evolution*, 62: 1965–1977, 2008.

A. Labra, J. Pienaar, and T. F. Hansen. Evolution of thermal physiology in *Liolaemus* lizards: Adaptation, phylogenetic inertia, and niche tracking. *Am. Nat.*, 174(2):204–220, 2009.

W. H. Mulder and F. W. Crawford. On the distribution of interspecies correlation for Markov models of character evolution on Yule trees. *J. Theor. Biol.*, 364:275–283, 2015.

R. V. Rohlfs, P. Harrigan, and R. Nielsen. Modeling gene expression evolution with an extended Ornstein–Uhlenbeck process accounting for within-species variation. *Mol. Biol. Evol.*, 31(1):201–211, 2013.

S. Sagitov and K. Bartoszek. Interspecies correlation for neutrally evolving traits. *J. Theor. Biol.*, 309:11–19, 2012.