Phylogenetic comparative methods: applying modern probabilistic methods to evolutionary biology

Krzysztof Bartoszek

Department of Mathematics
Uppsala University

Będlewo, 28 May 2015

Species proximity

Figs. 2 and 3 (J. Felsenstein. **Phylogenies and the comparative method.** *Am. Nat.*, 125(1): 1-15, 1985)

Figs. 5, 6 and 7 (J. Felsenstein. **Phylogenies and the comparative method.** *Am. Nat.*, 125(1): 1-15, 1985)

Evolutionary models — neutral evolution

- Independent contrasts, **Brownian motion** (Felsenstein 1985)
- $\mathrm{d}X_t = \sigma \mathrm{d}B_t,\ X_0 = X_0$
- $\mathrm{E}\left[X\right](t) = X_0, \quad \mathrm{Var}\left[X\right](t) = \sigma^2 t \to \infty$
- $\mathrm{Cov}\left[X_i, X_j\right](t) = \sigma^2 t_{ij}$

Evolutionary models — adaptive evolution

- Adaptation, **Ornstein–Uhlenbeck** (Hansen 1997)
- $\mathrm{d}X_t = -\alpha(X_t - \theta)\mathrm{d}t + \sigma \mathrm{d}B_t,\ X_0 = X_0$
- $\mathrm{E}\left[X\right](t) = e^{-\alpha t} X_0 + (1 - e^{-\alpha t})\theta \to \theta$
- $\mathrm{Var}\left[X\right](t) = \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t}) \to \frac{\sigma^2}{2\alpha}$
- $\mathrm{Cov}\left[X_i, X_j\right] = \frac{\sigma^2}{2\alpha}(e^{-2\alpha t_{ij}} - e^{-2\alpha t})$

Interacting traits — slouch

- Adaptation to a randomly fluctuating optimum
- Allometry $y = ax^b \rightarrow \log y = \log a + b \cdot \log x$
- Hansen, Pienaar, Orzack (2008)
- $$\begin{aligned} \mathrm{d}Y(t) &= -\alpha \left(Y(t) - \psi(t) - \sum_{l=1}^{k} b_l X_l(t) \right) \mathrm{d}t \\ &\quad + \sigma_y \mathrm{d}B_y(t) \\ \mathrm{d}X_l(t) &= \sigma_l \mathrm{d}B_l(t) \end{aligned}$$

Interacting traits — mvSLOUCH

- Coadaptation of traits
- Bartoszek, Pienaar, Mostad, Andersson, Hansen (2012)
-
$$\mathrm{d}\vec{X}(t) = \mathbf{\Sigma}\mathrm{d}\vec{B}_x(t)$$
-
$$\mathrm{d}\vec{Y}(t) = -\mathbf{A}\left(\vec{Y}(t) - \vec{\psi}(t)\right)\mathrm{d}t + \mathbf{\Sigma}_y\mathrm{d}\vec{B}_y(t)$$
-
$$\begin{aligned}\mathrm{d}\vec{Y}(t) &= -\mathbf{A}\left(\vec{Y}(t) - \vec{\psi}(t) - \mathbf{B}\vec{X}(t)\right)\mathrm{d}t \\ &\quad +\mathbf{\Sigma}_y\mathrm{d}\vec{B}_y(t) \\ \mathrm{d}\vec{X}(t) &= \mathbf{\Sigma}\mathrm{d}\vec{B}_x(t)\end{aligned}$$

Interacting traits — mvSLOUCH

- Eigenvalues of $\mathbf{A}$ — λ
- all $\lambda > 0$ — adaptation
- evolutionary regression $\mathrm{E}\left[\vec{Y}|\vec{X}\right](t)$
- optimal regression $\mathbf{B}$
- $\mathrm{E}\left[\vec{Y}|\vec{X}\right](t) \to \mathbf{B}$ in OUBM mode

Simpson's paradox

$$\mathrm{Cov}\left[Y_i, Y_j\right](t)$$

Fig. E1 (K. Bartoszek, J. Pienaar, P. Mostad, S. Andersson, T.F. Hansen. **A phylogenetic comparative method for studying multivariate adaptation.** *J. Theor. Biol.*, 314:204–215, 2012)

Plard et. al. (2011) — Cervidae

Fig. 2A (K. Bartoszek, J. Pienaar, P. Mostad, S. Andersson, T.F. Hansen. **A phylogenetic comparative method for studying multivariate adaptation.** *J. Theor. Biol.*, 314:204–215, 2012)

- antler length, male and female body size
- breeding group size, mating strategy

Coadaptation — mvSLOUCH

Plard et. al. (2011) OLS explains the data best
i.e. *no* phylogenetic effects, but only compared to BM.

Figs. 1 and 3 (K. Bartoszek, J. Pienaar, P. Mostad, S. Andersson, T.F. Hansen. **A phylogenetic comparative method for studying multivariate adaptation.** *J. Theor. Biol.*, 314:204–215, 2012)

mvSLOUCH analysis

Figs. 2C and 2D (K. Bartoszek, J. Pienaar, P. Mostad, S. Andersson, T.F. Hansen. **A phylogenetic comparative method for studying multivariate adaptation.** *J. Theor. Biol.*, 314:204–215, 2012)

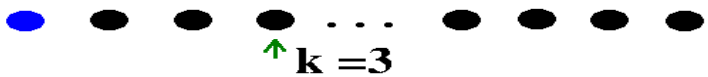
- antler length = female body size + breeding group size
- male body size = female body size + breeding group size
- female body size : BM

But what if we do not observe the tree ?

Fig. 1 (S. Sagitov, K. Bartoszek **Interspecies correlation for neutrally evolving traits.** *J. Theor. Biol.*, 309:11–19, 2012)

Gernhard (2008), Sagitov and KB (2012)

- Conditional on T all $n-1$ s_is are independent.
- $$q_n(T|n) = n\lambda^n(\lambda-\mu)^2 \frac{(1-e^{-(\lambda-\mu)T})^{n-1} e^{-(\lambda-\mu)T}}{(\lambda-\mu e^{-(\lambda-\mu)T})^{n+1}}$$
- $$F(s|T) = \frac{1-\lambda e^{-(\lambda-\mu)s} - \mu e^{-(\lambda-\mu)T}}{(\lambda-\mu e^{-(\lambda-\mu)s})(1-e^{-(\lambda-\mu)T})} \mathbf{1}(s \le T)$$
- $$\begin{aligned} \mathrm{E}\left[T-\tau\right] &= \mathrm{E}\left[T - \int_0^T 1 - F^{\kappa}(s|T)\mathrm{d}s\right] \\ &= \sum_{k=1}^{n-1} \frac{n-k}{\binom{n}{2}} \int_0^{\infty} \left(\int_0^T F^k(s|T)\mathrm{d}s\right) q_n(T)\mathrm{d}T \end{aligned}$$



Conditioned Yule tree

$$H_{n,k} = \sum_{i=1}^{n} \frac{1}{i^k} \quad b_{n,x} = \prod_{i=1}^{n} \frac{i}{i+x}$$

- $\mathrm{E}[T] = H_{n,1}/\lambda \sim \ln n/\lambda$
- $\mathrm{Var}[T] = H_{n,2}/\lambda^2 \to \frac{\pi^2}{6\lambda^2}$
- $\mathrm{E}[\tau] = \frac{n+1}{\lambda(n-1)} H_{n,1} - \frac{2n}{\lambda(n-1)} \sim \ln n/\lambda$
- $\mathrm{Var}[\tau] = \frac{(n^2-1)H_{n,2} - 2(n+1)H_{n,1}^2 + 4(n+1)H_{n,1} - 4n}{\lambda^2(n-1)^2} \to \frac{\pi^2}{6\lambda^2}$
- $\mathrm{Cov}[T,\tau] = \frac{1}{\lambda^2(n-1)}(2n - (n+1)H_{n,2}) \to \frac{1}{\lambda^2}(2 - \frac{\pi^2}{6})$
- $\mathrm{E}\left[e^{-\gamma T}\right] = b_{n,\gamma}, \quad \mathrm{E}\left[e^{-\gamma\tau}\right] = \frac{2-(n+1)(\gamma+1)b_{n,\gamma}}{(n-1)(\gamma-1)}$

Application: Interspecies correlation ($\lambda = 1$)

How similar do we expect two random species to be?

$$\rho_n = \frac{\operatorname{Cov}[X_1, X_2]}{\operatorname{Var}[X]}$$

Brownian motion

$$\mathrm{d}X(t) = \sigma \mathrm{d}B(t) \quad X(0) = X_0$$

$$\mathrm{E}[X(t)] = X_0 \qquad \operatorname{Var}[X(t)] = \sigma^2 t$$

$$\rho_n \sim 2(\ln n)^{-1}$$

Ornstein–Uhlenbeck process

$$\mathrm{d}X(t) = -\alpha(X(t) - \theta)\mathrm{d}t + \sigma \mathrm{d}B(t) \quad X(0) = X_0$$

$$\mathrm{E}[X(t)] = e^{-\alpha t} X_0 + (1 - e^{-\alpha t})\theta \qquad \operatorname{Var}[X(t)] = \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t})$$

$$\rho_n \sim \begin{cases} Cn^{-2\alpha} & 0 < \alpha < 0.5 \\ 2n^{-1} \ln n & \alpha = 0.5 \\ Cn^{-1} & \alpha > 0.5 \end{cases}$$

Application: Classical estimators ($\lambda = 1$)

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

	$\mathrm{E}\left[\bar{X}_n\right]$	$\mathrm{Var}\left[\bar{X}_n\right]$	$\mathrm{E}\left[S_n^2\right]$	$\mathrm{Var}\left[S_n^2\right]$
$\alpha = 0$	X_0	$2\sigma^2$	$\sigma^2 \ln n + O(1)$	$(\frac{\pi^2}{6} + 1)\sigma^4$
$\alpha \in (0, \frac{1}{2})$	$\theta + O(n^{-\alpha})$	$c_\alpha n^{-2\alpha}$	$\frac{\sigma^2}{2\alpha} + O(n^{-2\alpha})$	$O(n^{-2\alpha})$
$\alpha = \frac{1}{2}$	$\theta + O(n^{-\alpha})$	$2\frac{\ln n}{n}$	$\frac{\sigma^2}{2\alpha} + O(\frac{\ln n}{n})$	$O(\frac{\ln n}{n})$
$\alpha > \frac{1}{2}$	$\theta + O(n^{-\alpha})$	$c_\alpha n^{-1}$	$\frac{\sigma^2}{2\alpha} + O(n^{-2\alpha})$	$O(n^{-1})$

$\mathrm{Var}\left[S_n^2\right] \to 0$

$$\mathrm{E}\left[S_n^2\right] \to \frac{\sigma^2}{2\alpha}$$

$$S_n^2 = \frac{1}{n}\sum_i X_i^2 - \frac{2}{n(n-1)}\sum_i \sum_{j>i} X_i X_j$$

$$\begin{aligned}\mathrm{E}\left[Z_1 Z_2 Z_3 Z_4\right] = {} & m^4 \\ & + m^2(c_{12} + c_{13} + c_{14} + c_{23} + c_{24} + c_{34}) \\ & + c_{12}c_{34} + c_{13}c_{24} + c_{14}c_{23}.\end{aligned}$$

$\mathrm{Var}\left[S_n^2\right] \to 0$

$$
\begin{aligned}
\mathrm{E}\left[(S_n^2)^2\right] &= \tfrac{1}{n^2}\Big(\sum_i \mathrm{E}\left[X_i^4\right] + 2\sum_i \sum_{j>i} \mathrm{E}\left[X_i^2 X_j^2\right]\Big) \\
&-\tfrac{4}{n^2(n-1)}\Big(\sum_i \sum_{j>i} \mathrm{E}\left[X_i^3 X_j\right] \\
&+\sum_i \sum_{j>i} \mathrm{E}\left[X_i X_j^3\right] + \sum_i \sum_{j>i} \sum_{k\neq i,j} \mathrm{E}\left[X_i^2 X_j X_k\right]\Big) \\
&+\tfrac{4}{n^2(n-1)^2}\Big(\sum_i \sum_{j>i} \mathrm{E}\left[X_i^2 X_j^2\right] \\
&+\sum_i \sum_{j>i} \sum_{k\neq i,j} \mathrm{E}\left[X_i^2 X_j X_k\right] \\
&+\sum_i \sum_{j>i} \sum_{k\neq i,j} \mathrm{E}\left[X_i X_j^2 X_k\right] \\
&+\sum_i \sum_{j>i} \sum_{k\neq i,j} \sum_{m>k;\ m\neq i,j} \mathrm{E}\left[X_i X_j X_k X_m\right]\Big).
\end{aligned}
$$

$$b_{n,x} \sim \Gamma(x+1) n^{-x}$$

Application: weak convergence

Theorem

Let $\delta = \frac{X_0 - \theta}{\sqrt{\sigma_a^2/2\alpha}}$ *be a normalized difference between the ancestral and optimal values. Consider the normalized sample mean* $\overline{Y}_n = \frac{\overline{X}_n - \theta}{\sqrt{\sigma_a^2/2\alpha}}$ *of the Yule-Ornstein-Uhlenbeck process with* $\overline{Y}_0 = \delta$. *As* $n \to \infty$ *the process* $\overline{Y}_n$ *has the following limit behaviour.*

(i) If $\alpha > 0.5$, *then* $\sqrt{n} \cdot \overline{Y}_n$ *is asymptotically normally distributed with zero mean and variance* $\frac{2\alpha+1}{2\alpha-1}$.

(ii) If $\alpha = 0.5$, *then* $\sqrt{n/\ln n} \cdot \overline{Y}_n$ *is asymptotically normally distributed with zero mean and variance* 2.

(iii) If $\alpha < 0.5$, *then* $n^\alpha \cdot \overline{Y}_n$ *converges a.s. and in* L^2 *to a random variable* $Y_{\alpha,\delta}$ *with* $\mathrm{E}\left[Y_{\alpha,\delta}\right] = \delta\Gamma(1+\alpha)$ *and* $\mathrm{E}\left[Y_{\alpha,\delta}^2\right] = \left(\delta^2 + \frac{4\alpha}{1-2\alpha}\right)\Gamma(1+2\alpha)$.

Weak convergence proof sketch

Lemma

Let $\mathcal{Y}_n$ be the σ–algebra containing information about our Yule tree. We have:

- $$\begin{aligned}
\mathrm{E}\left[\overline{Y}_n|\mathcal{Y}_n\right] &= \delta e^{-\alpha T}, \\
\mathrm{E}\left[\overline{Y}_n^2|\mathcal{Y}_n\right] &= n^{-1} + (1-n^{-1})\,\mathrm{E}\left[e^{-2\alpha\tau}|\mathcal{Y}_n\right] \\
&\quad - e^{-2\alpha T} + \delta^2 e^{-2\alpha T}, \\
\mathrm{Var}\left[\overline{Y}_n|\mathcal{Y}_n\right] &= n^{-1} + (1-n^{-1})\,\mathrm{E}\left[e^{-2\alpha\tau}|\mathcal{Y}_n\right] \\
&\quad - e^{-2\alpha T}.
\end{aligned}$$

- *For all positive α we have* $\mathrm{Var}\left[\mathrm{E}\left[e^{-2\alpha\tau}|\mathcal{Y}_n\right]\right] = O(n^{-3})$ *as* $n \to \infty$.

Weak convergence proof sketch

Proof.
E.g.

$$\begin{aligned}\operatorname{Var}\left[Y_1 + \ldots + Y_n | \mathcal{Y}_n\right] &= n(1 - e^{-2\alpha T}) \\ &\quad + 2\sum_{i<j}(e^{-2\alpha\tau_{ij}} - e^{-2\alpha T}) \\ &= n - n^2 e^{-2\alpha T} \\ &\quad + n(n-1)\operatorname{E}\left[e^{-2\alpha\tau} | \mathcal{Y}_n\right].\end{aligned}$$

□

Weak convergence proof sketch

Let $\mathcal{Y}_n$ be the σ–algebra about our Yule tree.
(i)
It suffices to show :

$$\begin{aligned}(\mu_n, \sigma_n^2) &:= \left(\sqrt{n}\,\mathrm{E}\left[\overline{Y}_n|\mathcal{Y}_n\right],\ n\,\mathrm{Var}\left[\overline{Y}_n|\mathcal{Y}_n\right]\right) \\ &\xrightarrow{P} \left(0, \frac{2\alpha+1}{2\alpha-1}\right), \quad n \to \infty,\end{aligned}$$

since then, due to the conditional normality of $\overline{Y}_n$, we will get the following convergence of characteristic functions

$$\mathrm{E}\left[e^{i\gamma\sqrt{n}\cdot\overline{Y}_n}\right] = \mathrm{E}\left[e^{i\mu_n\gamma-\sigma_n^2\gamma^2/2}\right] \to e^{-\frac{2\alpha+1}{2(2\alpha-1)}\gamma^2}.$$

Weak convergence proof sketch

We can write,

$$\mu_n = \sqrt{n}\delta e^{-\alpha T},$$
$$\sigma_n^2 = 1 + (n-1)\,\mathrm{E}\left[e^{-2\alpha\tau}|\mathcal{Y}_n\right] - ne^{-2\alpha T}.$$

and further

$$\mathrm{E}\left[\sigma_n^2\right] = 1 - nb_{n,2\alpha} + \frac{2-(n+1)(2\alpha+1)b_{n,2\alpha}}{2\alpha-1} \to \frac{2\alpha+1}{2\alpha-1}.$$

Due to fast enough decay of variances:

$$1 + (n-1)\,\mathrm{E}\left[e^{-2\alpha\tau}|\mathcal{Y}_n\right] \overset{L^2}{\to} \frac{2\alpha+1}{2\alpha-1},$$
$$ne^{-2\alpha T} \overset{L^2}{\to} 0,$$
$$\mu_n \overset{L^2}{\to} 0$$

hence

$$(\mu_n, \sigma_n^2) \overset{P}{\to} (0, \frac{2\alpha+1}{2\alpha-1}).$$

Application: confidence interval

$$\mathrm{E}\left[S_n^2\right] \to \frac{\sigma_a^2}{2\alpha} \qquad \mathrm{Var}\left[S_n^2\right] \to 0$$

allows us to write confidence intervals

$$\text{for } \alpha > 0.5 \qquad \overline{X}_n \pm z_{1-x/2} \cdot \sqrt{S_n^2} \cdot \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{2\alpha+1}{2\alpha-1}},$$

$$\text{for } \alpha = 0.5 \qquad \overline{X}_n \pm z_{1-x/2} \cdot \sqrt{S_n^2} \cdot \frac{\sqrt{2\ln n}}{\sqrt{n}},$$

$$\text{for } \alpha < 0.5 \qquad (\overline{X}_n - q_{x/2} \cdot \sqrt{S_n^2} \cdot n^{-\alpha},$$
$$\overline{X}_n + q_{1-x/2} \cdot \sqrt{S_n^2} \cdot n^{-\alpha}).$$

Including jumps

Fig. 1 (K. Bartoszek **Quantifying the effects of anagenetic and cladogenetic evolution.** *Math. Biosc.*, 254:42–57, 2014)

- Jump has mean 0, variance σ_c^2
- Jumps decorrelate the contemporary sample
- Quadratic variation

$$\mathrm{E}\left[[X]\right] = \sigma_a^2 \mathrm{E}\left[T\right] + \sigma_c^2 \mathrm{E}\left[\Upsilon^*\right]$$

Including jumps

$$H_{n,k} = \sum_{i=1}^{n} \frac{1}{i^k} \quad b_{n,x} = \prod_{i=1}^{n} \frac{i}{i+x}$$

- $\mathrm{E}[\Upsilon] = 2H_{n,1} - 2 \sim 2\ln n$
- $\mathrm{Var}[\Upsilon] = 2\left(H_{n,1} - 1 - 2(H_{n,2} - 1)\right) \sim 2\ln n$
- $\mathrm{E}[\upsilon] = \frac{4}{n-1}\left(n - H_{n,1}\right) - 2 \to 2 \quad \mathrm{Var}[\upsilon] = 6 + O(n^{-1}\ln n)$
- $s > 0$, for υ we need $n \geq 2$,

$$\begin{aligned} \mathrm{E}\left[s^{\Upsilon}\right] &= \frac{1}{n}\frac{1}{b_{n-1,2s-1}} \\ \mathrm{E}\left[s^{\upsilon}\right] &= \begin{cases} \frac{1}{(n-1)(3-2s)}\left((n+1) - \frac{2}{nb_{n-1,2s-1}}\right) \\ \frac{2(n+1)}{n-1}\left(H_{n+1} - \frac{5}{3}\right), s = 1.5 \end{cases} \end{aligned}$$

Network models

Joint work with G. Jones, B. Oxelman, S. Sagitov

Fig. 1 (K. Bartoszek, G. Jones, B. Oxelman, S.S. **Time to a single hybridization event in a group of species with unknown ancestral history.** *J. Theor. Biol.*, 322:1–6, 2013)

- speciation rate λ
- hybridization rate β
- if $\frac{\beta_n}{2\lambda_n} n \to 0$ then $\tau_n \xrightarrow{\mathcal{D}}$ exponential(2λ)

and also working with S. Glemin, I. Kaj, M. Lascoux (UU)

Thank you !!!!

- S. Andersson, G. Jones, P. Mostad, B. Oxelman, S. Sagitov, M. Prager (GU, CTH)
- S. Glemin, I. Kaj, M. Lascoux (UU)
- V. Mitov, T. Stadler (ETH Zürich)
- T.F. Hansen, K. Voje (Oslo Univ.)
- P. Liò, H. Xiao (Cambridge Univ.)
- J. Pienaar (Alabama Univ.)

References

- K.B., J. Pienaar, P. Mostad, S. Andersson, T.F. Hansen. **A phylogenetic comparative method for studying multivariate adaptation.** *J. Theor. Biol.*, 314:204–215, 2012
- S.S., K.B. **Interspecies correlation for neutrally evolving traits.** *J. Theor. Biol.*, 309:11–19, 2012
- K.B., G. Jones, B. Oxelman, S.S. **Time to a single hybridization event in a group of species with unknown ancestral history.** *J. Theor. Biol.*, 322:1–6, 2013
- K.B., S.S. **Phylogenetic confidence intervals for the optimal trait value.** *J. Appl. Probab.*, 52, 2015
- K.B. **Quantifying the effects of anagenetic and cladogenetic evolution.** *Math. Biosc.*, 254:42–57, 2014
- K.B., S.S. **A consistent estimator of the evolutionary rate.** *J. Theor. Biol.*, 371:69-78, 2015