Luminy 15 June 2012

Serik Sagitov

Chalmers University and University of Gothenburg

20 slides

Interspecies correlation for Brownian traits

Joint work with Krzysztof Bartoszek: http://arxiv.org/abs/1201.5364

Comparative phylogenetics

Trait evolution model

Sample mean and variance of the trait values

Conditioned Yule tree

Aldous-Popovic-Stadler tree

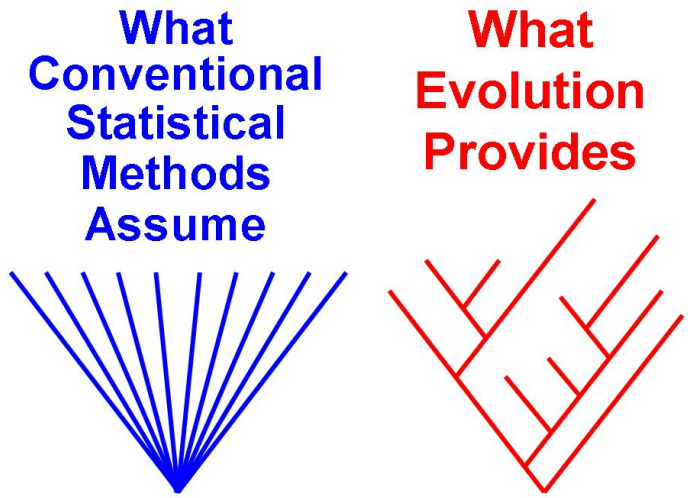
Interspecies correlation: exact and approximate formulae

Graphical illustrations

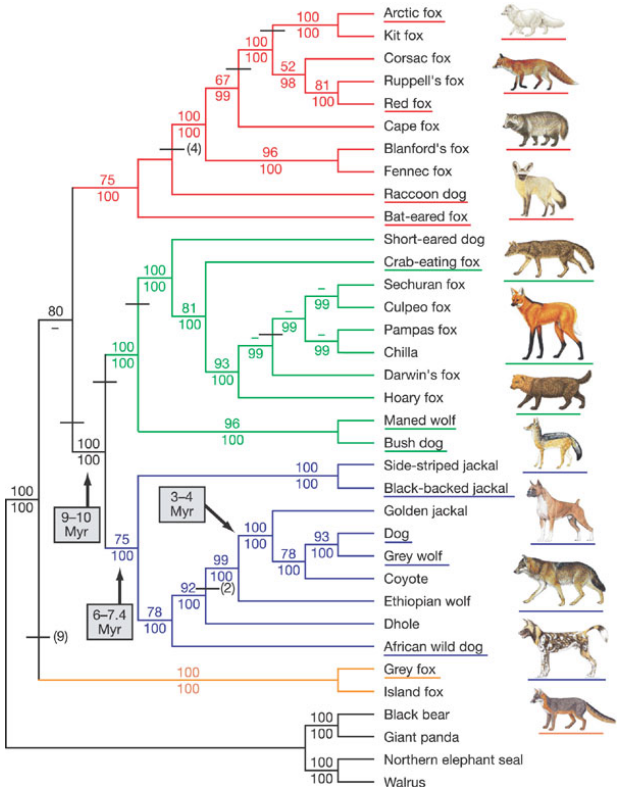
Ongoing and future projects

Comparative phylogenetics:

studies various trait values like log-bodysize $(X_1, \ldots, X_n)$ for a sample of related species



We want a simple formula for the correlation coefficient of (X_i, X_j)



Our paper "Interspecies correlation for neutrally evolving traits" is accepted by *Journal of Theoretical Biology* on 5 June 2012

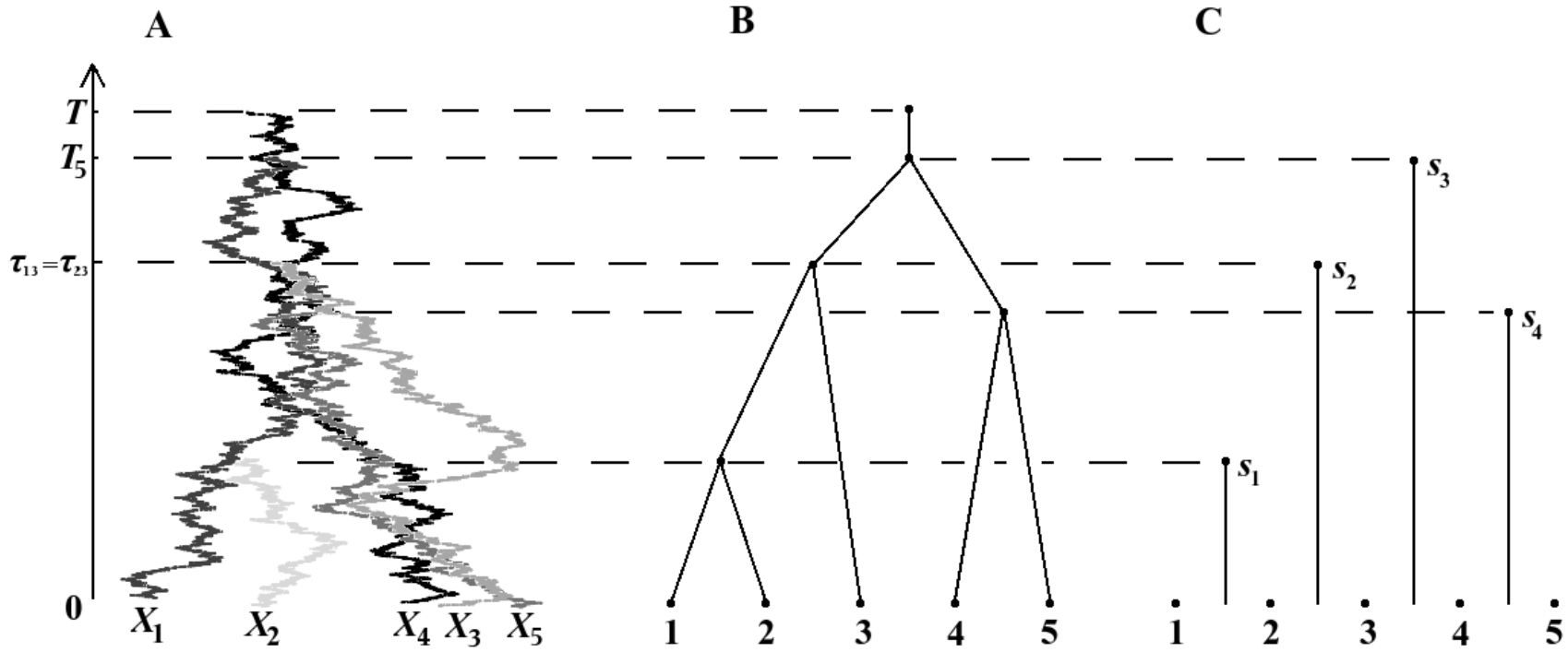
Extracts from a reviewer report for the JTB

To a mathematician the methodology may not seem so novel – perhaps any expert could do these calculations.

To a data-oriented biologist models like this may be regarded as so oversimplified and unrealistic that doing formal statistical tests of significance to see if data is consistent with the model, would have little scientific justification.

But such criticism could be leveled at ALMOST ALL WORK in this area of mathematical biology.

Trait evolution model



A: the trait value evolution for $n = 5$ species

B: the species tree $\boxed{T = \text{time to the origin}}$

C: speciation times $\boxed{\tau = \text{MRCA time for a random pair of species}}$

Coalescent processes model the gene trees.
The conditional branching processes model the species trees.

Basic Yule-BM model

G.U.Yule (1924):

let us model speciation by a
pure birth process with parameter λ.

J.Felsenstein (1985):

the trait value evolves along a lineage as a Brownian motion with variance σ^2 and ancestral state X_0.

Sample mean and variance

Estimates of the ancestral state X_0 and the variance σ^2

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} \qquad \text{sample mean}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{sample variance}$$

In the BM case, given the mean time to the origin $\mathrm{E}[T]$, we have

$$\begin{aligned}
\mathrm{E}\left[\bar{X}\right] &= X_0 \\
\mathrm{Var}\left[\bar{X}\right] &= \sigma^2 n^{-1}(1 + (n-1)\rho_n)\,\mathrm{E}[T] \\
\mathrm{E}\left[S^2\right] &= \sigma^2(1-\rho_n)\,\mathrm{E}[T]
\end{aligned}$$

we call ρ_n the interspecies correlation coefficient

$$\rho_n = \frac{1}{\binom{n}{2}\,\mathrm{Var}[X]} \sum_{1 \le i < j \le n} \mathrm{Cov}[X_i, X_j]$$

Conditioned Yule tree

Consider the Yule process with known age $\{T = t\}$

$$P(Z_t = n) = (1 - e^{-\lambda t})^{n-1} e^{-\lambda t}$$

where the geometric form is explained by the contour process of the Yule tree. The speciation times are iid with distribution function

$$F_t(s) = \frac{1 - e^{-\lambda s}}{1 - e^{-\lambda t}}, \ 0 \le s \le t$$

Assuming the uniform $(0, \infty)$ prior for T one gets a posterior density for T given $\{Z_T = n\}$

$$q_n(t) = n\lambda(1 - e^{-\lambda t})^{n-1} e^{-\lambda t}$$

Posterior distribution for ordered speciation times $T_1 > \ldots > T_{n-1}$

$$(T, T_1, \ldots, T_{n-1}) \stackrel{d}{=} (n \text{ ordered iid rv with Exp}(\lambda) \text{ distribution})$$

In the BM case the interspecies correlation coefficient can be computed as

$$\rho_n = \frac{\mathrm{E}\left[T - \tau\right]}{\mathrm{E}\left[T\right]} \tag{1}$$

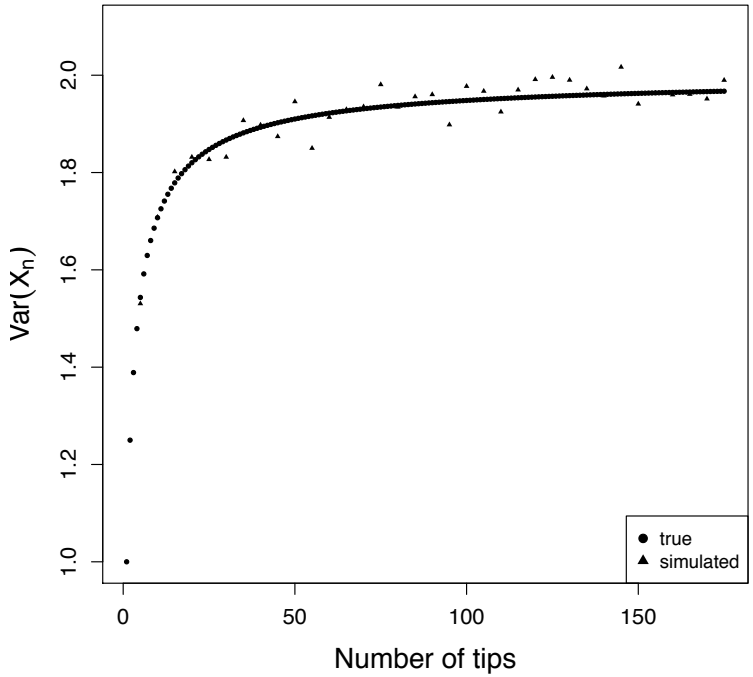
where τ is the time to the most recent common ancestor for a pair of randomly chosen extant species, and

$$\mathrm{E}\left[T\right] = \int_0^\infty t q_n(t) dt$$

$$\mathrm{E}\left[T - \tau\right] = \sum_{k=1}^{n-1} \frac{2(n-k)}{n(n-1)} \int_0^\infty \left(\int_0^t F_t^k(s) ds \right) q_n(t) dt$$

Yule-BM case: $\rho_n = \frac{2}{n-1}\left(\frac{n}{a_n} - 1\right)$, where $a_n = \sum_{i=1}^{n} \frac{1}{i}$

Sample mean and variance in the Yule-BM case



In the Yule-BM case with $\lambda = 1$, $\mu = 0$, and $\sigma^2 = 1$

$$\mathrm{Var}\left[\bar{X}\right] = 2 - \frac{a_n}{n} \quad \text{and} \quad \mathrm{E}\left[S^2\right] = \frac{(n+1)a_n - 2n}{n-1}$$

Aldous-Popovic-Stadler tree

Conditionally on n tips and time to origin $\{T = t\}$ the $n-1$ speciation times are iid



Aldous-Popovic (2005)

- a birth-death process with speciation rate λ
- and the same extinction rate $\mu = \lambda$, the critical case
- conditioned on n extant species
- improper Uniform $(0, \infty)$ prior for the time of origin T

Stadler (2008)

- the conditioned supercritical birth-death $\mu < \lambda$
- with $\mu = 0$ we get the conditioned Yule tree

Supercritical case $\mu = 1$ and $\lambda > 1$

$$\rho_n = \frac{2}{n-1}\left(\frac{n(1+e_{n,\lambda})}{a_n + e_{n,\lambda} - \ln\frac{\lambda}{\lambda-1}} - \frac{\lambda}{\lambda-1}\right),$$

where

$$e_{n,\lambda} = \lambda^n\left(\ln\frac{\lambda}{\lambda-1} - \sum_{i=1}^{n}\frac{1}{i\lambda^i}\right) = \sum_{i=1}^{\infty}\frac{1}{(n+i)\lambda^i} = \int_0^1 \frac{x^n dx}{\lambda - x}$$

so that $e_{n,\lambda} \in (0, \frac{1}{n(\lambda-1)})$ for $1 < \lambda < \infty$.

The Yule case can be recovered from here by letting $\lambda \to \infty$.

Critical case $\mu = 1$ and $\lambda = 1$

- uniform prior over $(0, \infty)$ gives posterior $\mathrm{E}[T] = \infty$
- proper uniform prior over $(0, N)$ gives $\mathrm{E}[T] = ne_{n,m}$

Important notation $m = 1 + 1/N$

Exact formula for ρ_n given the proper uniform prior over $(0, N)$

$$\rho_n = 2 - \frac{2N}{n-1}\left(1 + \frac{1}{e_{n,m}}\right) + \frac{2N(N+1)}{n(n-1)}\left(1 + \frac{a_n - \ln(N+1)}{e_{n,m}}\right)$$

Question for biologists: if the unit of time corresponds to one speciation event, what is a realistic upper bound for N?

Supercritical case: uniformly in $\lambda \geq \lambda_0$ for any $\lambda_0 > 1$

$$\rho_n = \frac{2}{\ln n + \gamma - \ln \frac{\lambda}{\lambda - 1} + o(1)}, \quad n \to \infty$$

Nearly critical case with $\lambda = 1 + 1/N$ and $N \to \infty$

- Approximation 1: n is fixed

$$\rho_n = 1 - \frac{1}{2(\ln N - a_n + 1) + o(1)}$$

- Approximation 2: $n \to \infty$ so that $n/N \to \alpha$

$$\rho_n \to 2\left(\frac{1 + I_\alpha}{\ln \alpha + \gamma + I_\alpha} - \frac{1}{\alpha}\right)$$

$$\boxed{I_\alpha = \int_\alpha^\infty \frac{e^{\alpha - x} dx}{x}}$$

so that $e^{-\alpha} I_\alpha = \int_\alpha^\infty \frac{e^{-x} dx}{x}$ is the exponential integral.

Critical case with a Uniform$(0, N)$ prior

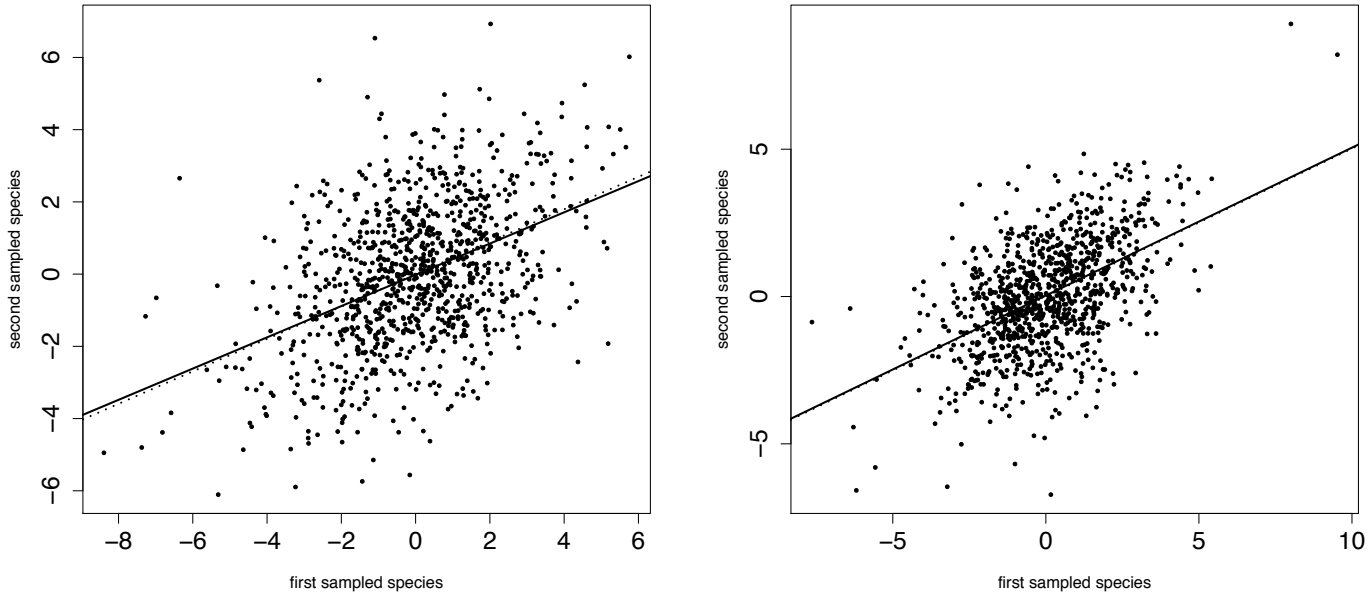
- Approximation 1: n is fixed and $N \to \infty$

$$\rho_n = 1 - \frac{1}{2(\ln N - a_n) + o(1)}$$

- Approximation 2: $n \to \infty$ and $N \to \infty$ so that $n/N \to \alpha$

$$\rho_n \to 2 - \frac{2}{\alpha}\left(1 + \frac{1}{I_\alpha}\right) + \frac{2}{\alpha^2}\left(1 + \frac{\ln \alpha + \gamma}{I_\alpha}\right)$$

Graphical illustrations

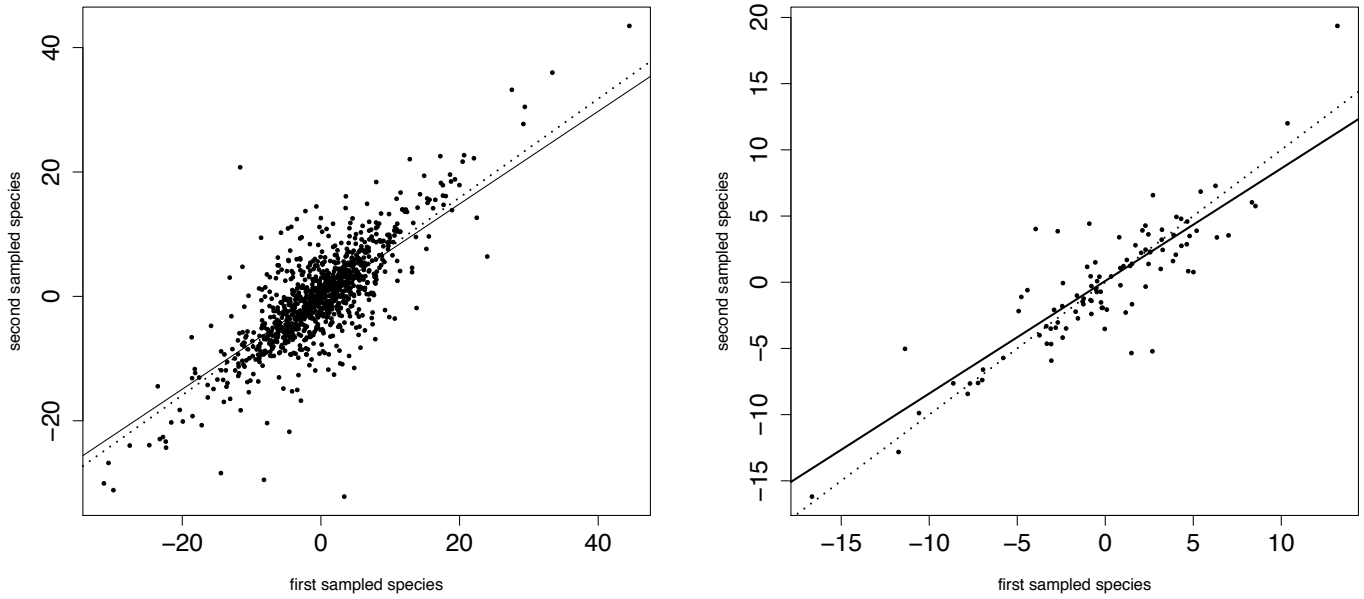


Regression line fitted to the simulated data (thick line) compared to the line $y = \rho_n x$ (dotted line) with ρ_n given by the exact formula:

Left: the Yule case with $\lambda = 1$ and $\mu = 0$.

Right: the supercritical case with $\lambda = 2$ and $\mu = 1$.

Graphical illustrations

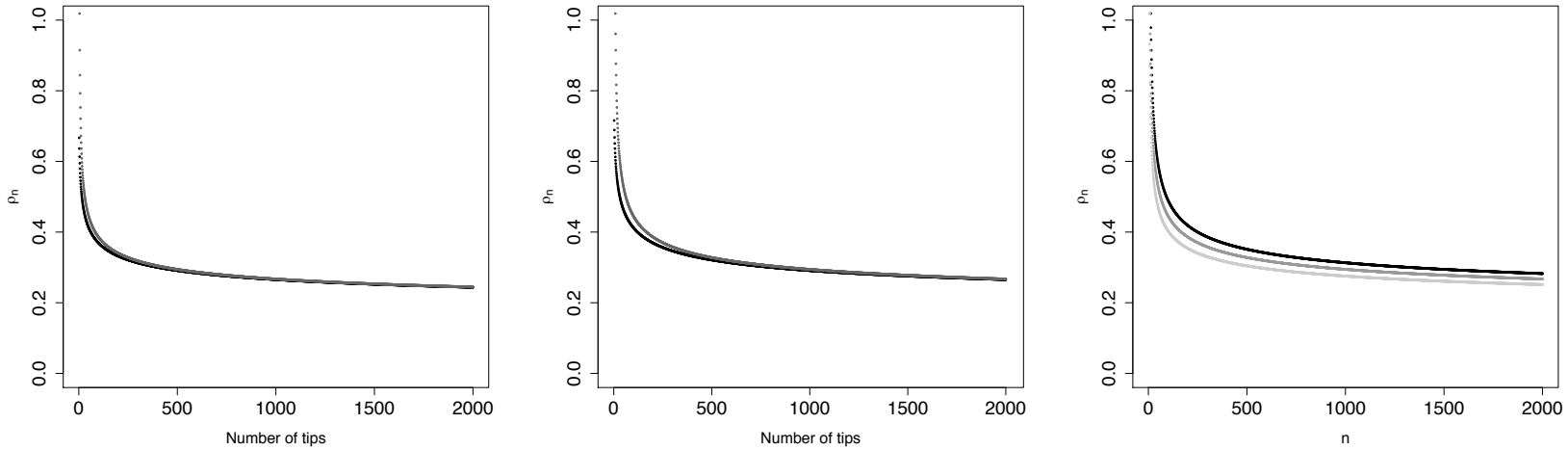


Regression line fitted to the simulated data (thick line) compared to the line $y = \rho_n x$ (dotted line) with ρ_n given by the exact formula:

Left: the near–critical case with $\lambda = 1.01$ and $\mu = 1$.

the critical case with improper prior $\lambda = \mu = 1$.

Graphical illustrations



Exact and approximate formulae for ρ_n in the supercritical case

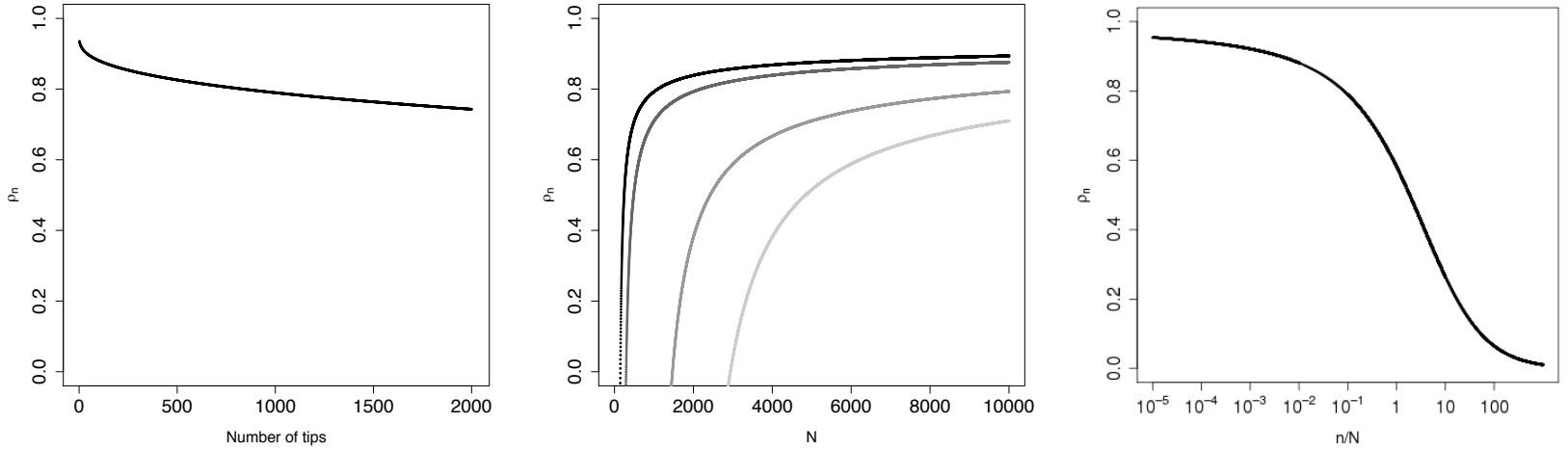
Left: $\mu = 0$, $\lambda = 1$, black line - exact, gray - approximate

Center: $\mu = 1$, $\lambda = 2$, black line - exact, gray - approximate

Right: three approximate lines for $\mu = 1$

black $\lambda = 1.5$, gray $\lambda = 2$, light gray $\lambda = 5$

Graphical illustrations



Exact and approximate formulae for ρ_n in the critical case $\lambda = \mu = 1$ with a Uniform$(0, N)$ prior

Left: exact formula for $N = 10^4$

Center: black to gray approx. for $n = 50, 100, 500, 1000$

Right: approximation as $n/N \to$ const

Ongoing project 1 (with K.Bartoszek): Confidence intervals for X_0

[conditioned Yule tree] + [Ornstein-Uhlenbeck traits]

Ongoing project 2 (with K.Bartoszek):

[conditioned Yule tree] +
[BM or OU-traits with jumps at speciation events]

Ongoing project 3 (with K.Bartoszek, G.Jones, B.Oxelman):
suppose a tetraploid comes as a hybrid of a pair in a set of n diploids, find the mean time to the hybridization event assuming

conditioned Yule tree for n diploid species with parameter λ

Poissonian clock for hybridization events with rate μ per pair of (ancestral) diploids

Future projects

Conditionally on n tips and time to origin $\{T = t\}$
the $n - 1$ speciation times are iid

Use more general tree models possessing this property:

- continuous time binary splitting tree
  CMJ-tree with Poisson reproduction
  (A.Lambert, 2010)
- discrete time tree with multifurcations
  linear-fractional BGW-tree with countably many types
  (SS, 2013)

Acknowledgement. This work was supported by the Swedish Research Council grant 621-2010-5623.