Serik Sagitov

Chalmers University and University of Gothenburg

25 slides

A phylogenetic confidence interval for the optimal trait value

http://arxiv.org/abs/1207.6488

joint work with

Krzysztof Bartoszek

Comparative phylogenetics

Interspecies correlation

Yule-Brownian-Motion model

Yule-Ornstein-Uhlenbeck model

Adaptation to an optimal value

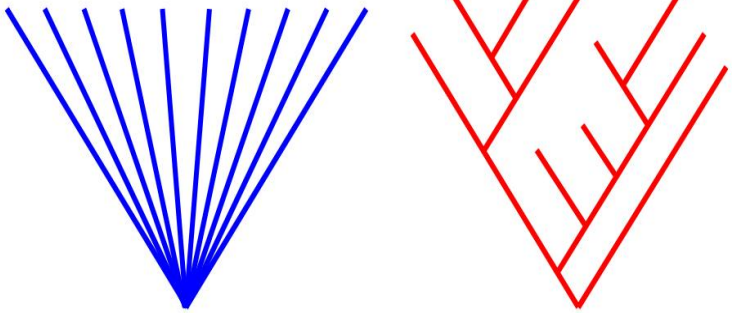
Sample mean and variance of the trait values

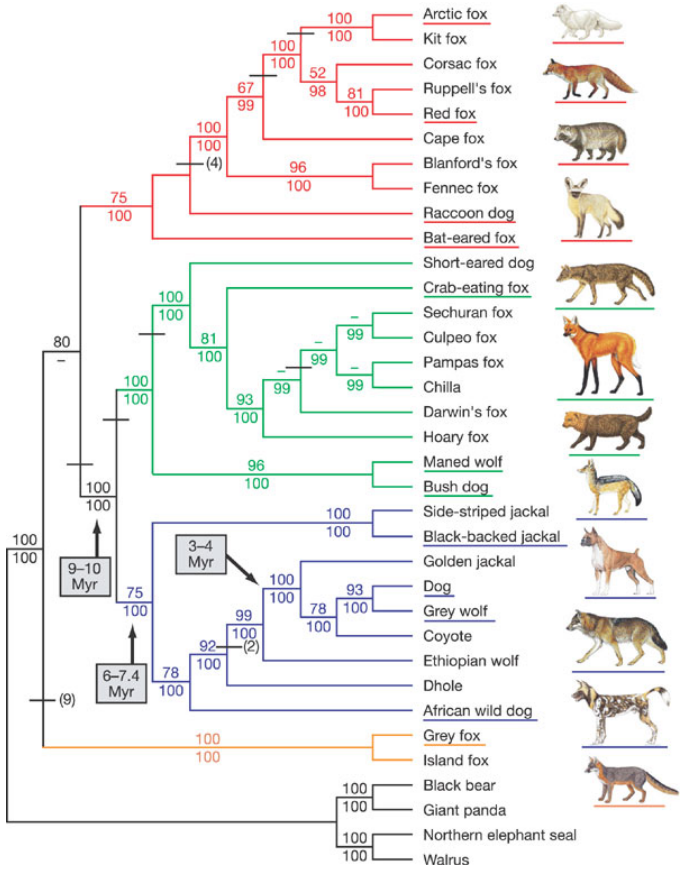
Phylogenetic confidence intervals

Exact and approximate formulae

Related projects

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





We will assume throughout that the species tree is unknown.

Interspecies correlation

In our previous paper Interspecies correlation for neutrally evolving traits.J. Theor. Biol. 309 (2012) 11-19 we obtain explicit formulae for

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

assuming that the trait evolves along a lineage as a Brownian Motion.

> Aldous-Popovic (2005), Stadler (2008): unknown species trees are modeled by branching processes conditioned on n tips.



In particular, we found for the Yule-BM-model

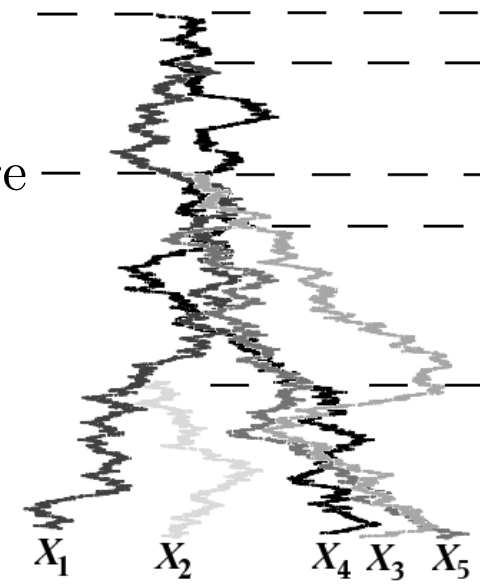
$$\rho_n = \frac{2}{n-1}\left(\frac{n}{a_n} - 1\right),$$

where $a_n = \sum_{i=1}^{n}(1/i)$ are the harmonic numbers.

YBM-model for the trait evolution

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 neutrally as a Brownian $B(t)$ motion

with variance σ^2 and ancestral state $B(0) = X_0$.

Coalescent processes model the gene trees.
Conditioned branching processes model the species trees.

Yule-Ornstein-Uhlenbeck model

In this paper we work with the YOU-model:
(conditioned Yule tree)+(OU trait evolution)

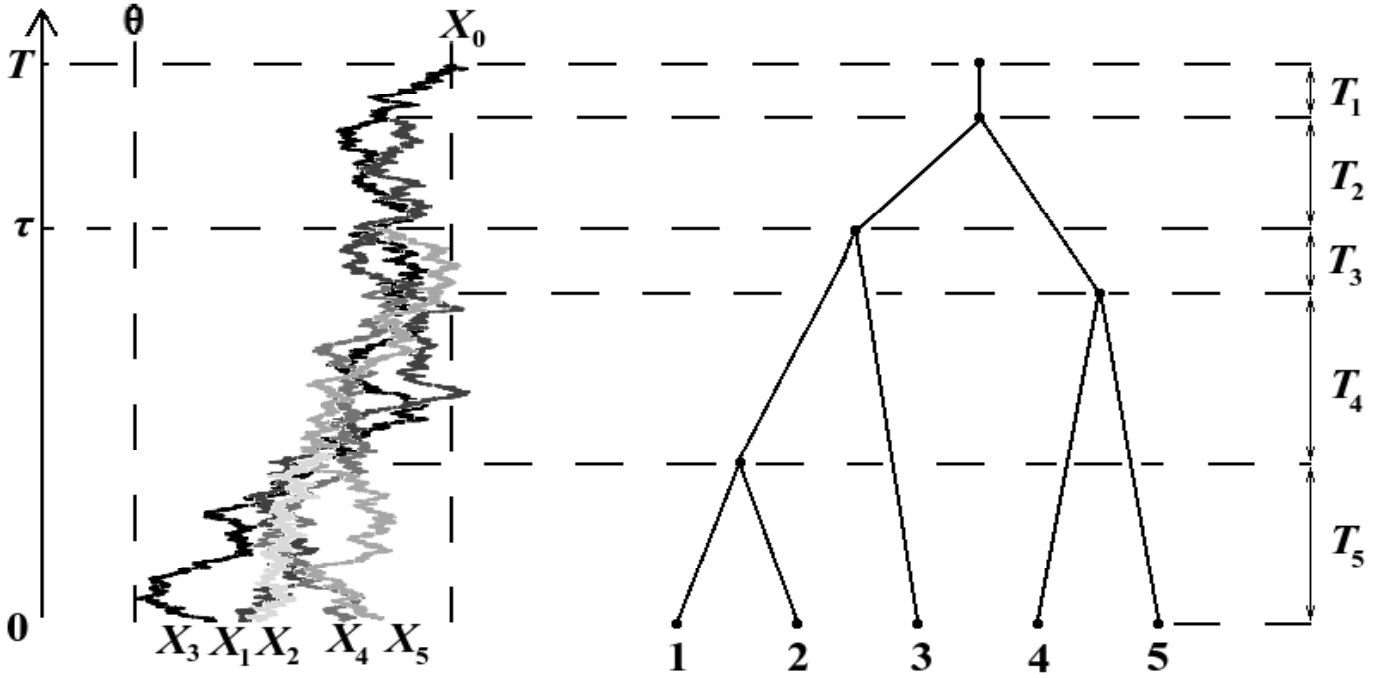
L.Ornstein and G.Uhlenbeck (1930)

$$dX(t) = -\alpha(X(t) - \theta)dt + \sigma dB(t)$$
$$X(0) = X_0$$

Just six parameters fully describe the YOU-model:

- number of species n
- speciation rate λ, wlog we assume $\lambda = 1$
- optimal trait value θ,
- adaptation rate $\alpha > 0$ $\quad \boxed{\alpha = 0 \Rightarrow X(t) = B(t)}$
- ancestral state X_0
- noise size $\sigma > 0$ $\quad \boxed{\sigma = 0 \Rightarrow X(t) = \theta + e^{-\alpha t}(X_0 - \theta)}$

Yule-Ornstein-Uhlenbeck model



T = time to the origin, T_i = inter-speciation times

τ = MRCA time for a random pair of species

θ = optimal trait value, X_0 = ancestral trait value

$(X_1, \ldots, X_n)$ = observed trait values identically distr. but dependent

OU-process: the distribution of $X(t)$ is normal with

$$\mathrm{E}\,[X(t)] = \theta + e^{-\alpha t}(X_0 - \theta)$$
$$\mathrm{Var}\,[X(t)] = \sigma_0^2(1 - e^{-2\alpha t})$$

$$\boxed{\sigma_0^2 = \frac{\sigma^2}{2\alpha}}$$

Stationary regime: $\mathrm{E}\,[X(t)] \to \theta$ and $\mathrm{Var}\,[X(t)] \to \sigma_0^2$ as $t \to \infty$.

The Yule tree conditioned on n tips:

run forward a linear birth process from a single branch and stop at the n-th speciation disregarding the nascent branch.

Interspeciation times

T_i = time duration with exactly i branches in the tree
T_i are independent exponential with rates $i = 1, \ldots, n$.

Backward description reminds a stretched Kingman's coalescent.

Classical summary statistics: sample mean and sample variance

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \quad \text{and} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Denoting X, Y a random pair of sampled trait values we get

$$\begin{aligned} \mathrm{E}\left[\bar{X}_n\right] &= \mathrm{E}\left[X\right] & \rho_n &= \mathrm{Cov}\left[X,Y\right] / \mathrm{Var}\left[X\right] \\ \mathrm{Var}\left[\bar{X}_n\right] &= \mathrm{Var}\left[X\right] - (n-1)n^{-1}(1-\rho_n)\,\mathrm{Var}\left[X\right] \\ \mathrm{E}\left[S_n^2\right] &= (1-\rho_n)\,\mathrm{Var}\left[X\right] \end{aligned}$$

and it becomes crucial to know the Laplace transforms of (T, τ):

$$\begin{aligned} \mathrm{E}\left[X\right] &= \theta + (X_0 - \theta)\,\mathrm{E}\left[e^{-\alpha T}\right] \\ \mathrm{Var}\left[X\right] &= \sigma_0^2(1 - \mathrm{E}\left[e^{-2\alpha T}\right]) + (X_0 - \theta)^2\,\mathrm{Var}\left[e^{-\alpha T}\right] \\ \mathrm{Cov}\left[X,Y\right] &= \sigma_0^2(\mathrm{E}\left[e^{-2\alpha\tau}\right] - \mathrm{E}\left[e^{-2\alpha T}\right]) + (X_0 - \theta)^2\,\mathrm{Var}\left[e^{-\alpha T}\right] \end{aligned}$$

Laplace transforms of T and τ

A Yule tree conditioned on n tips has $n-1$ splittings.

Let κ_n be the splitting number (counted forward in time) corresponding to the coalescent of two randomly sampled tips.

Key Lemma. We claim that

$$\mathrm{P}(\kappa_n = k) = \frac{2(n+1)}{(n-1)(k+2)(k+1)}, \ k = 1, \ldots, n-1$$

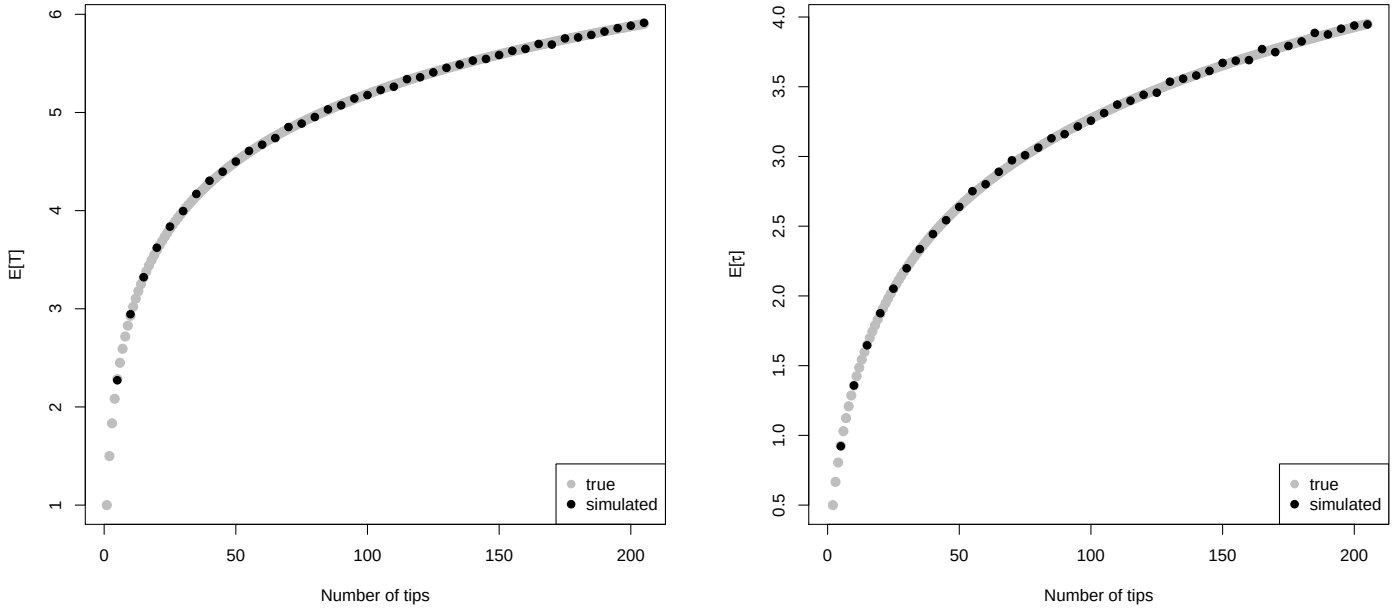
and that for any $x > -1$ and $y > -1$

$$\mathrm{E}\left[e^{-x(T-\tau)-y\tau}\right] = \frac{2(n+1)b_{n,y}}{(n-1)} \sum_{k=1}^{n-1} \frac{b_{k,x}}{(k+2)(k+1)b_{k,y}}$$

$$\boxed{b_{n,x} = \frac{1}{1+x} \cdot \frac{2}{2+x} \cdots \frac{n}{n+x}}$$

Moments of T and τ

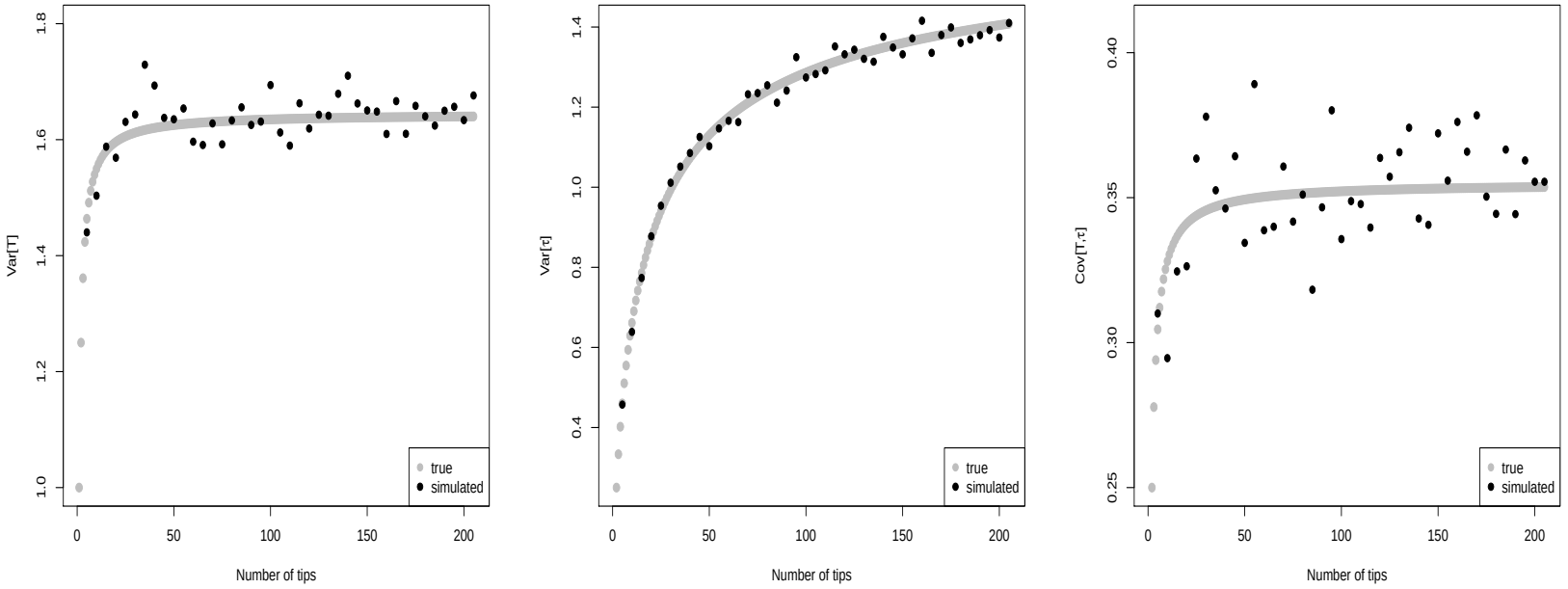
Using the Key Lemma we derive recursively all the joint moments $\mathrm{E}\left[(T-\tau)^k \tau^m\right]$ for $k \geq 0$ and $m \geq 0$.



Simulated and true values of $\mathrm{E}[T]$ and $\mathrm{E}[\tau]$.

Each point comes from 10^4 simulated Yule trees.

Moments of T and τ



Simulated and true values of $\mathrm{Var}\left[T\right]$, $\mathrm{Var}\left[\tau\right]$, and $\mathrm{Cov}\left[T, \tau\right]$.

Each point comes from 10^4 simulated Yule trees.

Exact formulae for the moments of $\bar{X}_n$ and S_n^2

Using the Key Lemma we compute

$$\mathrm{E}\left[e^{-\alpha T}\right] = b_{n,\alpha} = \frac{1}{1+\alpha} \cdot \frac{2}{2+\alpha} \cdots \frac{n}{n+\alpha}$$

$$\mathrm{E}\left[e^{-\alpha \tau}\right] = \frac{2-(n+1)(y+1)b_{n,\alpha}}{(n-1)(\alpha-1)}$$

which imply

$$\mathrm{E}\left[\bar{X}_n\right] = b_{n,\alpha} X_0 + (1-b_{n,\alpha})\theta$$

$$\mathrm{Var}\left[\bar{X}_n\right] = \sigma_0^2 \cdot \frac{2\alpha+1-(4\alpha n+2\alpha+1)b_{n,2\alpha}}{(2\alpha-1)n} + (X_0-\theta)^2(b_{n,2\alpha}-b_{n,\alpha}^2)$$

$$\mathrm{E}\left[S_n^2\right] = \sigma_0^2\left(1+\frac{(2\alpha+1)(n+1)b_{n,2\alpha}-2}{(2\alpha-1)(n-1)}\right)$$

For example, with $\alpha = 1$ we get $\quad \mathrm{E}\left[\bar{X}_n\right] = \frac{1}{n+1} X_0 + \frac{n}{n+1}\theta.$

$\bar{X}_n$ and S_n^2 are asymptotically unbiased estimates of the optimum θ and the stationary OU-variance σ_0^2 respectively.

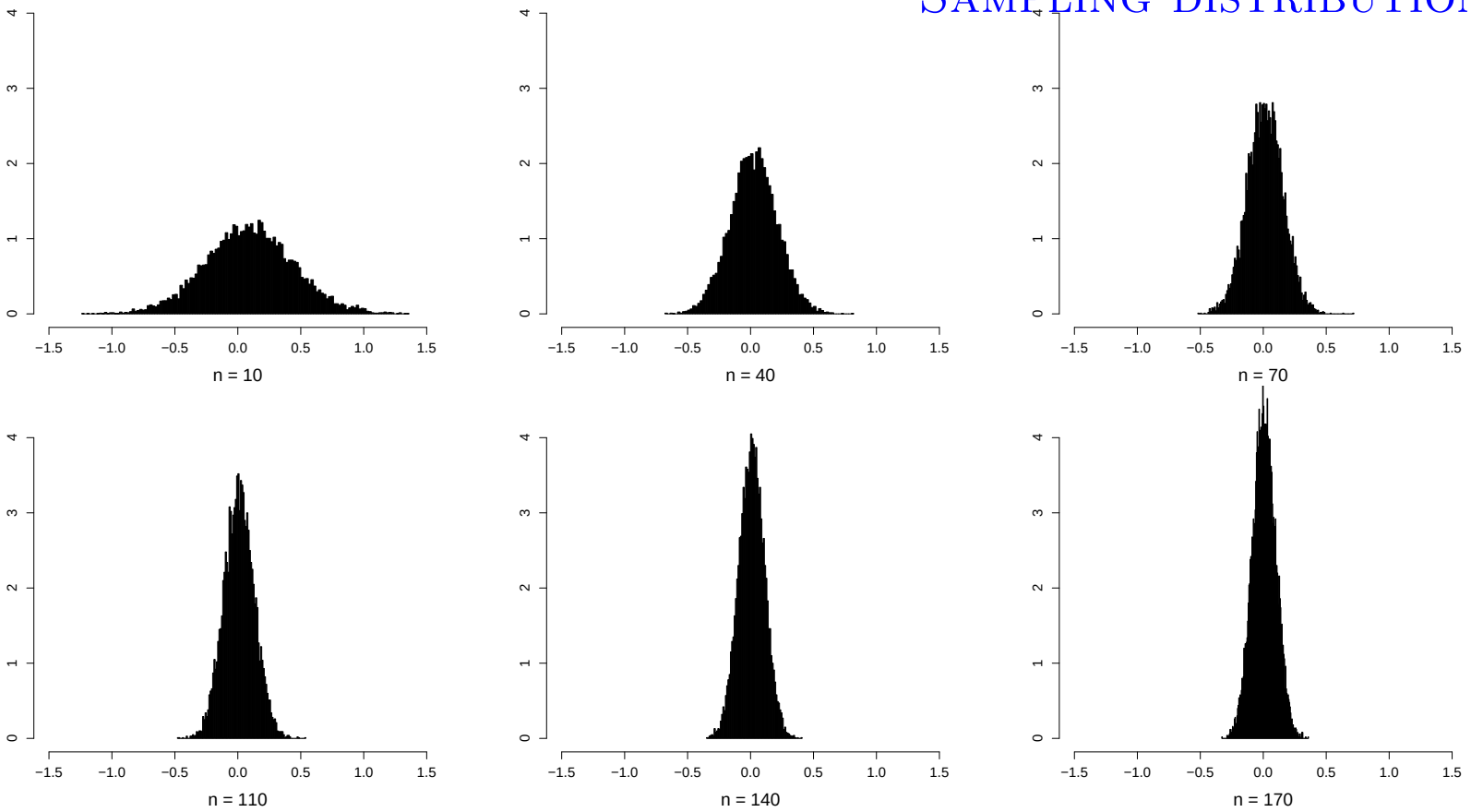
Concistency of the estimate $\bar{X}_n$

$$\mathrm{Var}\left[\bar{X}_n\right] \sim \sigma^2 \cdot \begin{cases} C_{\alpha,\delta} \cdot n^{-2\alpha}, & \text{if } 0 < \alpha < 0.5, \\ 2n^{-1} \ln n, & \text{if } \alpha = 0.5, \\ \frac{2\alpha+1}{2\alpha(2\alpha-1)n}, & \text{if } \alpha > 0.5, \end{cases}$$

where $\delta = \frac{|X_0 - \theta|}{\sigma}$ is the normalized deviation of X_0 from θ, and

$$C_{\alpha,\delta} = \frac{2}{1-2\alpha}\Gamma(2\alpha+1) + \delta^2\Gamma(2\alpha+1) - \delta^2\Gamma^2(\alpha+1).$$

Sampling distributions



Histograms of $\bar{X}_n$ for $\alpha = 1$, $X_0 = 1$, $\theta = 0$, $\sigma = 1$.

Each histogram is based on 10^4 independently simulated samples.

For sufficiently strong adaptation when $\alpha > 0.5$ as $n \to \infty$ we have

$$\mathrm{E}\left[\bar{X}_n\right] = \theta + O(\delta n^{-\alpha})$$

$$\mathrm{Var}\left[\bar{X}_n\right] \sim \frac{2\alpha+1}{2\alpha-1} \cdot \frac{\sigma_0^2}{n}$$

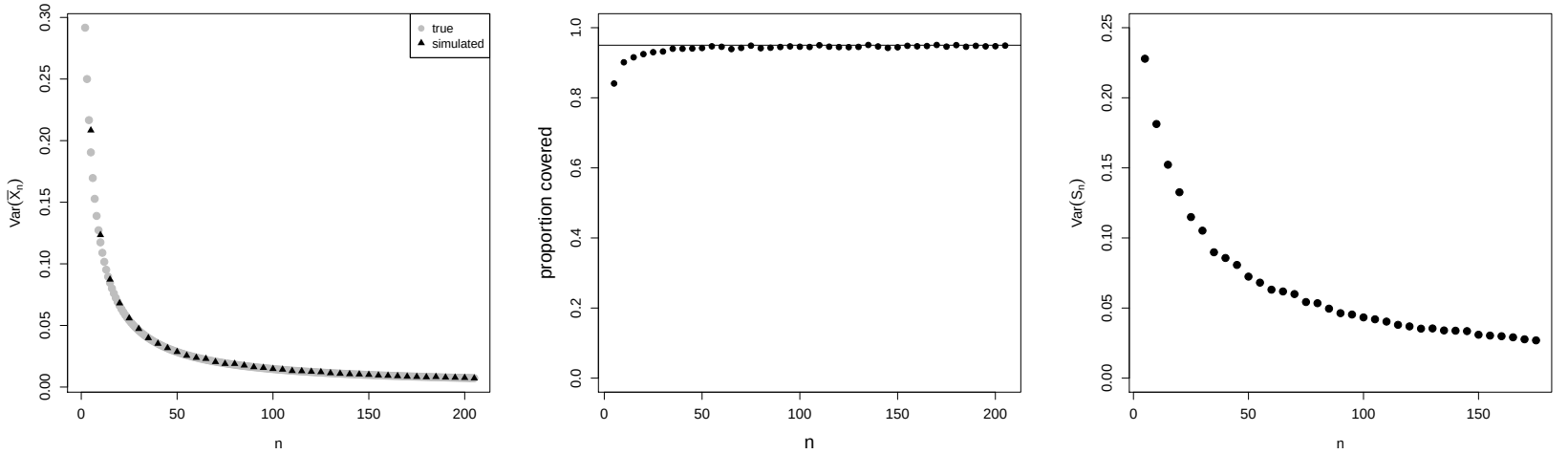
$$\mathrm{E}\left[S_n^2\right] \to \sigma_0^2 \qquad\qquad \sigma_0^2 = \sigma^2/(2\alpha)$$

Let $\alpha > 0.5$ and $\delta = \frac{|X_0-\theta|}{\sigma}$ stay bounded as $n \to \infty$. We propose a convenient approximate formula for a 95% confidence interval for θ:

$$\boxed{\bar{X}_n \pm 1.96 \cdot S_n \cdot \frac{K_\alpha}{\sqrt{n}}} \qquad K_\alpha = \sqrt{\frac{2\alpha+1}{2\alpha-1}}$$

Compared to $(\bar{X}_n \pm 1.96 \cdot S_n/\sqrt{n})$ it has an extra factor $K_\alpha > 1$ reflecting a positive correlation among the sample observations.

Phylogenetic confidence interval



Simulations results for $\alpha = 1$, $\sigma = 1$, $X_0 = 1$, $\theta = 0$.

Left: variance of the sample mean, simulated and theoretical values.

Center: each dot gives the coverage proportion of 10^4 confidence intervals. The horizontal line: predicted coverage of 95%.

Right: simulated values of Var $\left[S_n^2\right]$ decrease toward zero with n indicating that S_n^2 is a consistent estimate of the stationary variance.

A strict mathematical justification of this formula requires two results which will be addressed elsewhere.

On one hand, have we have to prove that

$$\mathrm{Var}\left[S_n^2\right] \to 0, \quad n \to \infty.$$

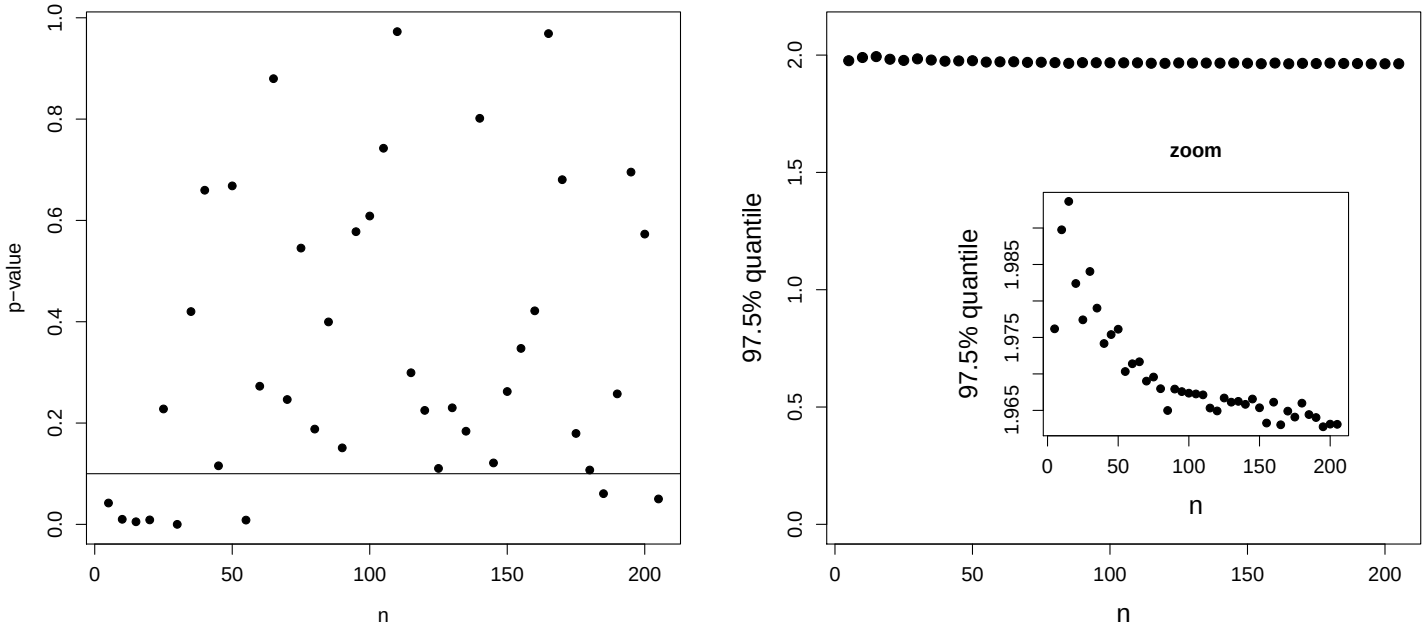
On the other hand, the following Central Limit Theorem has to be verified

Conjecture. *As* $n \to \infty$, *the standardized sample mean*

$$Z_n = (\bar{X}_n - \mathrm{E}\left[\bar{X}_n\right])(\mathrm{Var}\left[\bar{X}_n\right])^{-1/2}$$

is asymptotically normally distributed with parameters (0,1).

Asymptotic normality



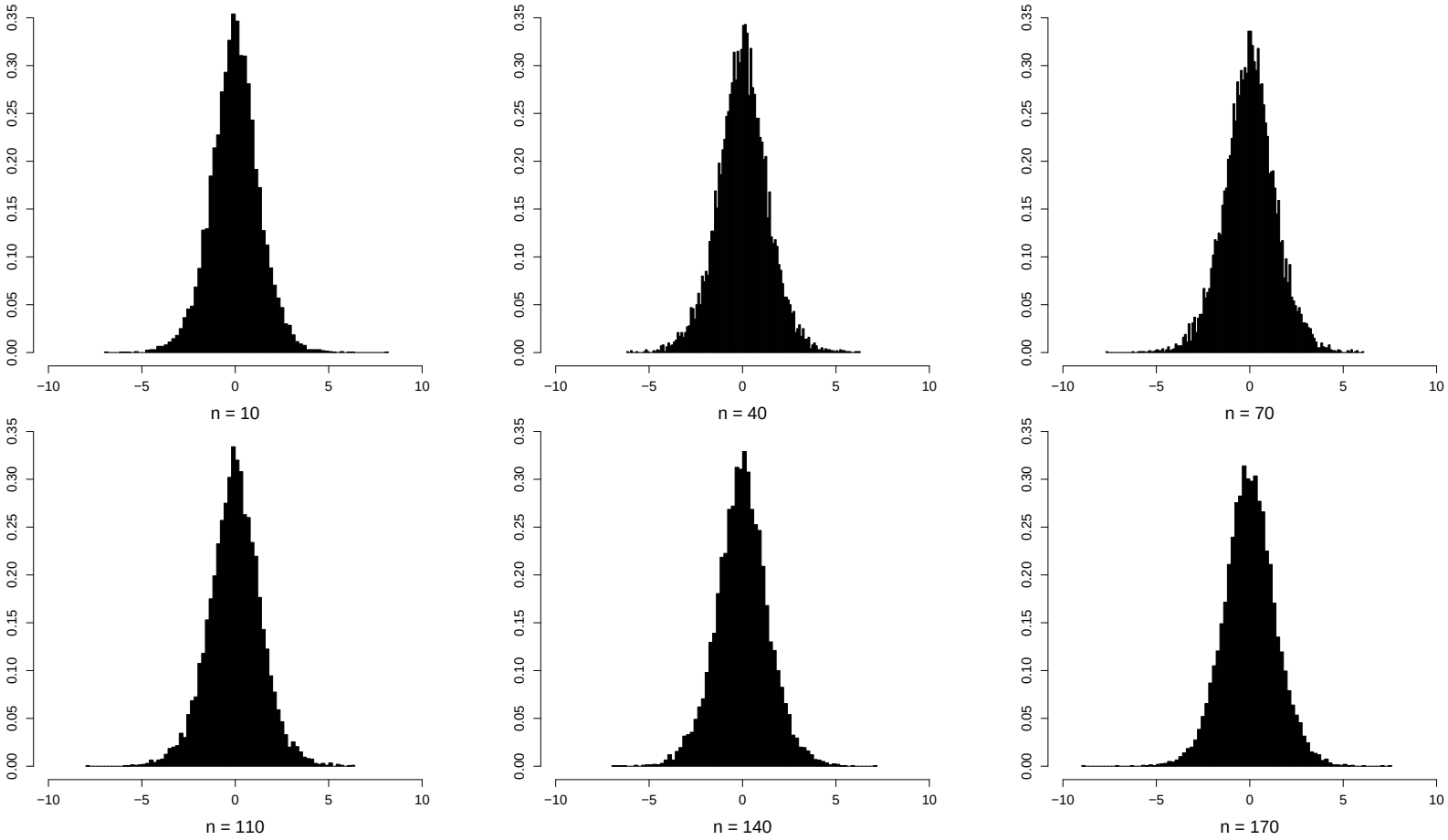
Simulation results for $\alpha = 1$, $X_0 = 1$, $\theta = 0$, $\sigma = 1$.

Left: p–value of Shapiro–Wilk test for normality of Z_n.

Right: A more thorough 0.975 level quantile check of normality for Z_n is given by a hybrid Monte–Carlo–numerical approach.

X_0 ESTIMATE FOR THE YMB-MODEL

If $\alpha = 0$ and $\sigma = 1$, then $\mathrm{E}\left[\bar{X}_n\right] = X_0$ with $\mathrm{Var}\left[\bar{X}_n\right] \to 2$ as $n \to \infty$.

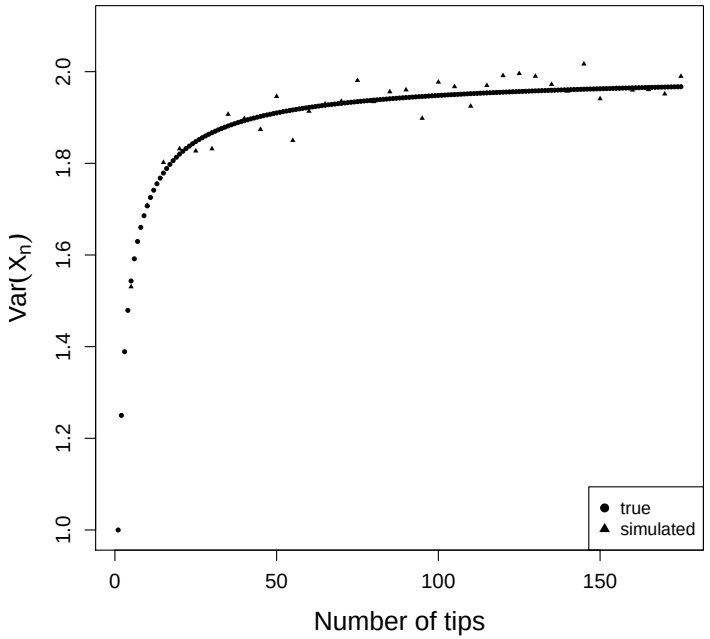


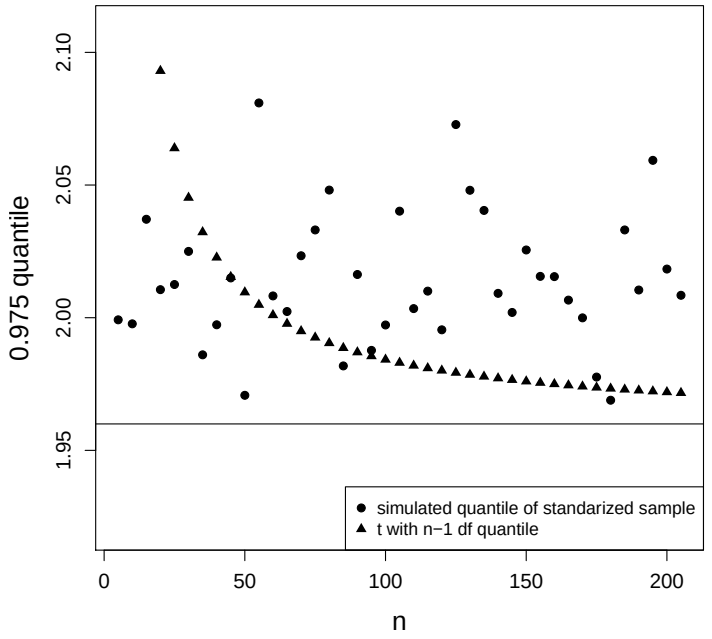
X_0 ESTIMATE FOR THE YMB-MODEL

For YBM-model $\bar{X}_n$ is an unbiased but not consistent estimate X_0.

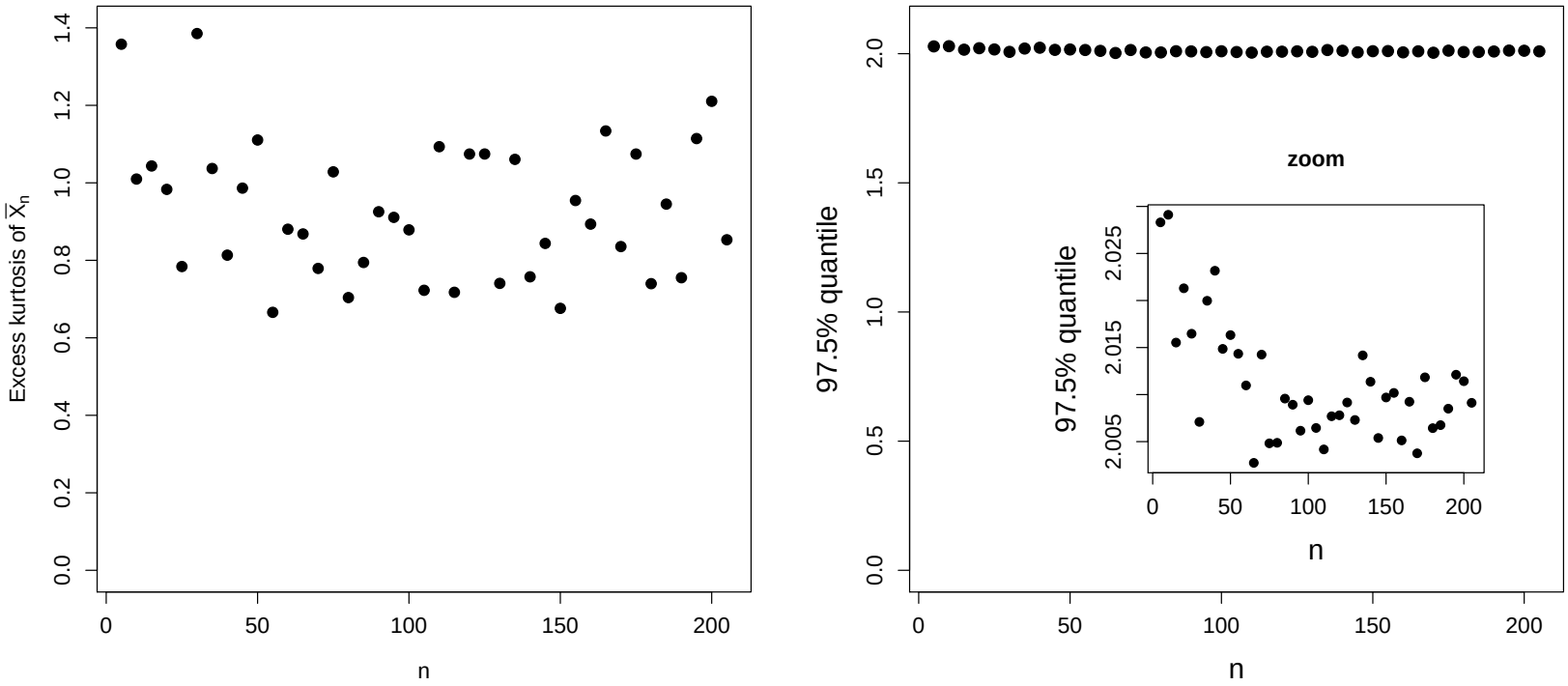
Assuming σ is known one can think of an approximate 95% CI for X_0

$$\bar{X}_n \pm q_n \frac{\sigma}{\sqrt{n}} \sqrt{2n - a_n}$$





X_0 ESTIMATE FOR THE YMB-MODEL



Asymptotic distribution of $\bar{X}_n$ for the Brownian motion model with $X_0 = 0$ and $\sigma = 1$ is symmetric but not normal.

From

$$\rho_n = 1 - \frac{\sigma_0^2(1 - \mathrm{E}\left[e^{-2\alpha\tau}\right])}{\sigma_0^2(1 - \mathrm{E}\left[e^{-2\alpha T}\right]) + (X_0 - \theta)^2 \,\mathrm{Var}\left[e^{-\alpha T}\right]}$$

we obtain putting $\delta = \frac{|X_0 - \theta|}{\sigma}$

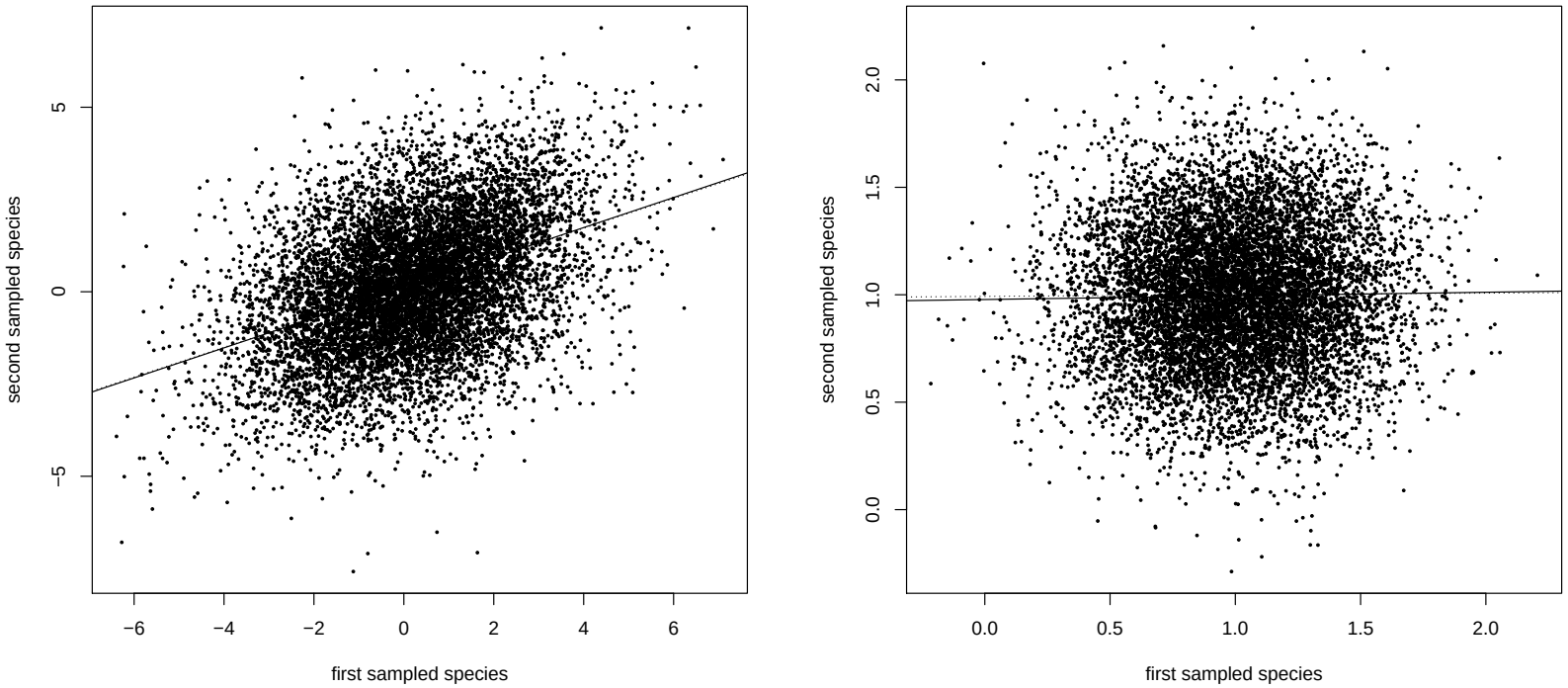
$$\boxed{\rho_n = 1 - \frac{2\alpha(n-1) + (n+1)((2\alpha+1)b_{n,2\alpha} - 1)}{(n-1)(2\alpha-1)(1 + (2\alpha\delta^2 - 1)b_{n,2\alpha}) - 2\alpha\delta^2 b_{n,\alpha}^2)}}$$

As $n \to \infty$

$$\rho_n \sim \begin{cases} 2(\ln n)^{-1}, & \text{if } \alpha = 0 \\ c_{\alpha,\delta}\, n^{-2\alpha}, & \text{if } 0 < \alpha < 0.5 \\ 2n^{-1} \ln n, & \text{if } \alpha = 0.5 \\ \frac{2}{2\alpha - 1} n^{-1}, & \text{if } \alpha > 0.5 \end{cases}$$

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

Interspecies correlation



Simulation results for $n = 30$, $\sigma = 1$, $X_0 = 0$, $\theta = 1$. Thick lines fitted to simulated data are indistinguishable from the predicted (dashed) lines $y = \rho_{30}x + (1 - b_{30,\alpha})(1 - \rho_{30})$. Left: $\alpha = 0.05$. Right: $\alpha = 5$.

Ongoing project 1 (with K.Bartoszek)

YOU-model with jumps at speciation events.

Ongoing project 2 (with K.Bartoszek, G.Jones, B.Oxelman): suppose a tetraploid comes as a hybrid of a pair in a set of n diploids. The time to the hybridization event is asymptotically exponential.

Assumption 1: conditioned Yule tree for n diploid species with parameter λ. Assumption 2: Poissonian clock for hybridization events with rate β per a pair of diploids.


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

References

1. Sagitov S. and Bartoszek K.

   Interspecies correlation for neutrally evolving traits. J. Theor. Biol. 309 (2012) 11-19

2. Bartoszek K. and Sagitov S.

   A phylogenetic confidence interval for the optimal trait value. Submitted to J. Math. Biol. http://arxiv.org/abs/1207.6488

3. Graham J., Sagitov S., and Oxelman B.

   Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting.

   Submitted to Syst. Biol. http://arxiv.org/abs/1208.3606

4. Bartoszek K., Jones G., Oxelman B., and Sagitov S.

   Time to a single hybridization event in a group of species with unknown ancestral history. In progress.