How to validate a Bayesian evolutionary model

Fábio K. Mendes[1†*], Remco Bouckaert[2†],
Luiz M. Carvalho[3†], Alexei J. Drummond[4]

[1]Department of Biology, Washington University in St. Louis

[2]School of Computer Science, The University of Auckland

[3]Escola de Matemática Aplicada, Fundação Getulio Vargas

[4]School of Biological Sciences, The University of Auckland

[*]Corresponding author: f.mendes@auckland.ac.nz

[†]Authors contributed equally to this work

November 10, 2023

Supplementary Material

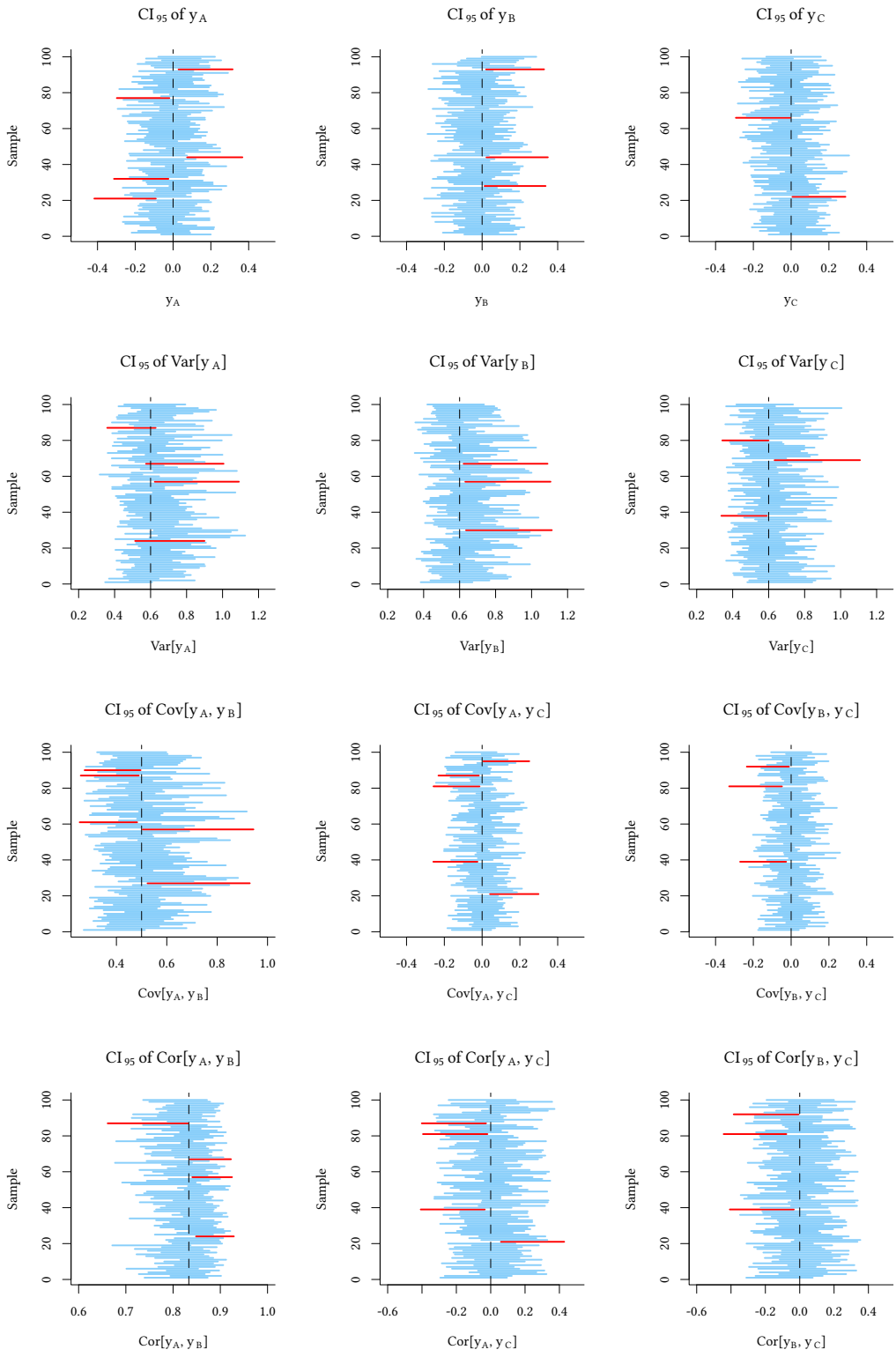
1 Validating a phylogenetic Brownian motion simulator

In this section we focus on validating a simulator for the phylogenetic Brownian motion model ("PhyloBM"; Felsenstein, 1973). As explained in the main text, our goal is to verify that the expected value of certain summary statistics (given a specific combination of parameter values) falls within its α-confidence intervals approximately α% of the time. We will build confidence intervals about statistics calculated from several PhyloBM samples of size 10,000, and then ask if the "population" value of a statistic – given by the parameters of the multivariate normal sampling distribution – is contained within its confidence interval frequently enough.

For summary statistics, we pay attention to the trait value's (i) species mean, (ii) species variance, (iii) among-species covariance, and (iv) among-species correlation coefficient. Supplementary figure 1 shows one hundred confidence intervals for each of these statistics, under multivariate normal MVN($\boldsymbol{y_0}, r\boldsymbol{T}$), where $\boldsymbol{y_0} = \{0, 0, 0\}$, $r = 0.1$ and $\boldsymbol{T}$ is given by the tree in Fig. 2 in the main text. Supplementary table 1 summarizes how often each statistic fell within its 95% confidence interval. These results indicate the PhyloBM simulator produces appropriate confidence intervals and behaves as expected.

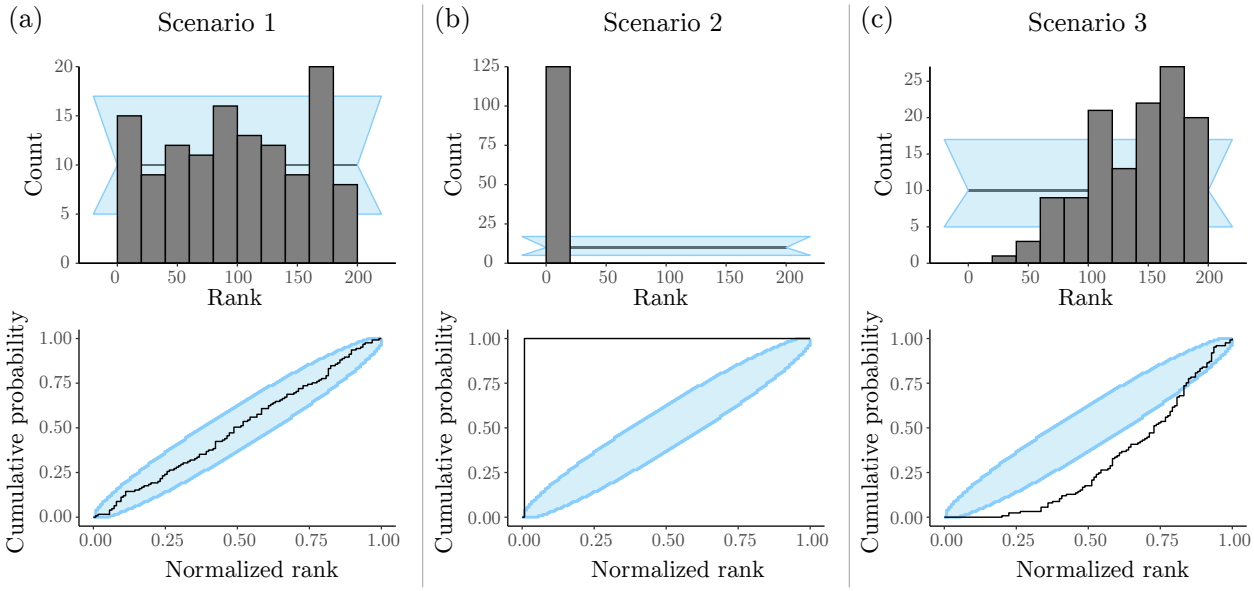
Supplementary Table 1: The number of times k that a summary statistic was contained within its corresponding 95% confidence interval. Each statistic was calculated from 100 datasets of size 10,000 simulated under the PhyloBM model described in the text and in Box 1 in the main text.

Statistic	Species s (and v)	k
$\mathrm{E}[y_s]$	A	95
	B	97
	C	98
$\mathrm{Var}[y_s]$	A	93
	B	97
	C	97
$\mathrm{Cov}[y_s, y_v]$	A and B	95
	A and C	95
	B and C	97
$\mathrm{Cor}[y_s, y_v]$	A and B	96
	A and C	96
	B and C	97



Supplementary Figure 1: One hundred 95% confidence intervals (blue and red lines) built for summary statistics of interest. Red lines represent intervals that do not contain the value expected under the MVN sampling distribution characterizing the PhyloBM model. Parameter values are described in the text.

2 Supplementary Figures



Supplementary Figure 2: Rank-uniformity validation (RUV) of the Bayesian hierarchical model in Fig. 1 in the main text. Panels in the top row show the histograms of $n = 100$ ranks, for parameter Λ in each scenario, obtained after 10% burnin and thinning of posterior samples down to 200 out of 10,000. Panels in the bottom row show the corresponding ECDF plots, for parameter Λ in each scenario. (a) In "Scenario 1", the model was correctly specified, and we simulated trees with 3 to 300 taxa using rejection sampling (approximately one in ten trees were rejected). (b) In "Scenario 2", the model was incorrectly specified in inference (see main text), and we used the same data sets simulated in "Scenario 1". (c) In "Scenario 3", the model was correctly specified, but rejection sampling was more intense (we rejected a large number of trees, approximately 90%, keeping those having between 100 to 200 tips).

3 Model validation with rejection sampling: a simple example

In this section, we experiment with a simple hierarchical Gaussian (toy) model to further examine the effect of rejection sampling in coverage validation and rank-uniformity validation (RUV). This experiment is motivated by results described in the main text, namely, the model in scenario 3 (Fig. 1, with extreme rejection sampling) passing coverage validation but not RUV.

Let us devise the following data-generating process for obtaining different levels of model misspecification:

$$\begin{aligned} \mu &\sim \text{Normal}\,(0,1)\;T[t,], \\ y_1,\ldots,y_K &\overset{\text{iid}}{\sim} \text{Normal}\,(\mu,1)\,, \end{aligned}$$

where K is the sample size, and notation $T[,t]$ indicates the distribution is truncated **below** at $t \in \mathbb{R}$, i.e.,

$$\pi_t(\mu) = \frac{\phi(\mu)}{1-\Phi(t)}\mathbb{I}(\mu > t).$$

and ϕ and Φ are the pdf and CDF of a standard normal, respectively. Here, μ and $y_1,\ldots,y_K$ correspond to θ_i and d_i in Fig. 4 (main text), respectively.

Inference is then done under the misspecified model

$$\begin{aligned} \mu &\sim \text{Normal}\,(0,1)\,, \\ y_1,\ldots,y_K &\overset{\text{i.i.d.}}{\sim} \text{Normal}\,(\mu,1)\,. \end{aligned}$$

It is well-known that:

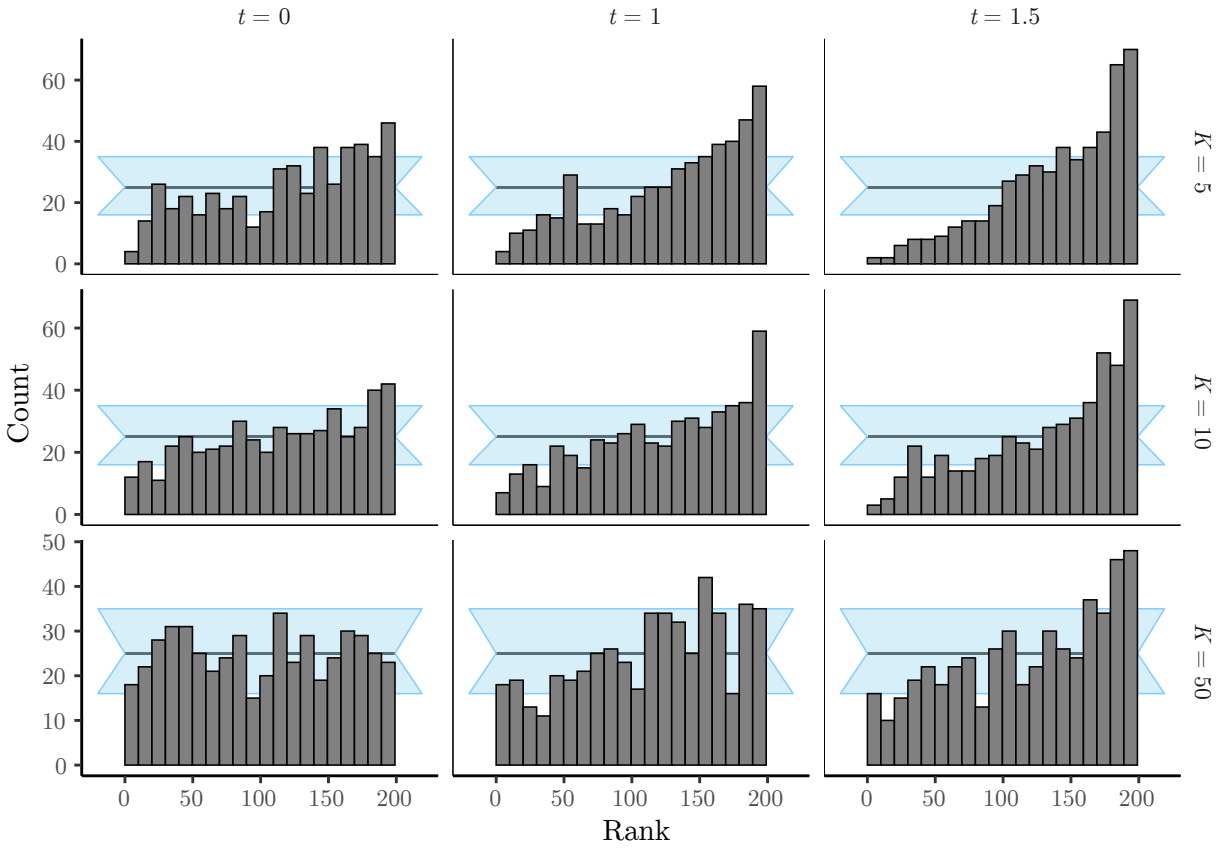
$$\mu \mid y_1,\ldots,y_K \sim \text{Normal}\left(\frac{\sum_{k=1}^{K} y_k}{K+1}, \frac{1}{K+1}\right), \tag{1}$$

i.e., the posterior is known in closed-form so MCMC is not required and one can sample directly and independently from it. One can thus control how extreme the model misspecification will be during inference by increasing t. Here we explore $t = \{0, 1, 1.5\}$ and $K = \{5, 10, 50\}$ in order to show how the method behaves in various misspecification scenarios. We ran $n = 1000$ replications of the experiment, draw 200 i.i.d. samples directly from the posterior in (1).

The resulting rank histograms (Supplementary Fig. 3) show how the posterior consistently underestimates the true mean, as indicated by a pattern of ranks bunching up towards the right-hand side of the histogram. The pattern is clearer in situations where the evidence in the data is weaker (e.g., when K is small), and the effect of the prior on the posterior is stronger. Moreover, the misspecification becomes more apparent the larger t is, that is, the more

extreme the misspecification becomes.

It is worth noting a couple of things. First, unlike what we observed and reported for the model in the main text, coverage validation does suggest an incorrectly specified model for certain combinations of t and K (Supplementary Table 2). Second, when the truncation is less extreme ($t = 0$) and there is a substantial amount of data ($K = 50$), RUV fails to detect any problems. This is an important insight about validation protocols: their sensitivity is context-dependent, and responds to the combination of model and data regime.



Supplementary Figure 3: Rank-uniformity validation (RUV) of the hierarchical Gaussian toy model. We show the rank histograms in three misspecification scenarios (t is 0.0, 1.0 or 1.5) and three data regimes (K is 5, 10 or 50). Results are based on $n = 1000$ replicates of $L = 200$ i.i.d. draws each. Horizontal black lines shows the expected count for each rank (the same for all ranks due to uniformity), and the light-blue bands represent the 95% confidence interval for the counts.

Supplementary Table 2: Coverage results for the 95%-HPD under the hierarchical Gaussian toy model. For each combination of truncation (t) and data set size (K), we show the estimated coverage over $n = 1000$ replicates, and whether the coverage procedure passes. A pass is determined according to table 2 in the main text.

Truncation (t)	Number of observations (K)	Estimated coverage	Pass
0	5	0.93	No
1	5	0.92	No
1.5	5	0.90	No
0	10	0.95	Yes
1	10	0.91	No
1.5	10	0.91	No
0	50	0.96	Yes
1	50	0.93	No
1.5	50	0.93	No

4 Proof for coverage validation

In this section we provide a mathematical argument that coverage-based validation is sound, i.e., that sampling from the prior, simulating data and then using the same prior for computing the posterior should give Bayesian credible intervals (BCIs) with nominal frequentist coverage.

For a number n of replicates, simulate parameter values θ_i, and then given those values, simulate data d_i:

$$\theta_i \sim f_\Theta(\cdot),$$
$$d_i \sim f_{D|\Theta}(\cdot|\Theta = \theta_i).$$

Now for notational convenience, define $a_i := a(d_i, \alpha)$ as the HPD lower bound and, similarly, $b_i := b(d_i, \alpha)$ as the HPD upper bound. Recall $I_\alpha(d_i)$ is such that:

$$Q_{d_i}(b_i) - Q_{d_i}(a_i) = p_1 - p_2 = \alpha,$$

where $Q_d(x)$ is the posterior CDF (conditional on data d) and $p_1, p_2 \in (0,1)$, with $p_1 < p_2$. A natural quantity to compute is:

$$S_n = n^{-1} \sum_{i=1}^{n} \mathbb{I}\left(\theta_i \in I_\alpha(d_i)\right),$$

i.e., the attained coverage of the Bayesian intervals.

Let $F_U(x) = x$ be the CDF of a Uniform$(0, 1)$ random variable. Now we can consider what happens when the number of simulations grows, i.e., the limit $\lim_{n\to\infty} S_n$. We may re-write the limit as:

$$
\begin{aligned}
\lim_{n\to\infty} S_n &= \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbb{I}\left(\theta_i \in I_\alpha\left(d_i\right)\right), \\
&= \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \left\{\mathbb{I}\left(\theta_i \leq b_i\right) - \mathbb{I}\left(\theta_i \leq a_i\right)\right\}, \\
&= \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbb{I}\left(\theta_i \leq b_i\right) - n^{-1} \sum_{i=1}^{n} \mathbb{I}\left(\theta_i \leq a_i\right), \\
&= \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbb{I}\left(Q_{d_i}^{-1}\left(\theta_i\right) \leq p_1\right) - \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbb{I}\left(Q_{d_i}^{-1}\left(\theta_i\right) \leq p_2\right), \\
&= F_U(p_1) - F_U(p_2) = \alpha,
\end{aligned}
$$

where the last line follows from the fact that the CDF of θ_i is uniformly distributed on $(0, 1)$ (Theorem 1 in Cook et al., 2006) and almost sure convergence of the ECDF to the true CDF due to the Glivenko-Cantelli theorem (Billingsley, 1986, page 275).

References

Billingsley, P. (1986). *Probability and measure*. John Wiley & Sons, second edition.

Cook, S. R., Gelman, A., and Rubin, D. B. (2006). Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graph. Stat.*, 15(3):675–692.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.*, 25:471–92.