A simple and general method for accounting for phylogenetic uncertainty via Rubin's rules in comparative analysis

Shinichi Nakagawa[1,2,4*] and Pierre de Villemereuil[3,4]

[1]Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia

[2]Department of Zoology, University of Otago, 340 Great King Street, Dunedin 9054, New Zealand

[3]Laboratoire d'Écologie Alpine (LECA-UMR CNRS 5553), Université Joseph Fourier, BP 53, 38041 Grenoble, France

[4]Both authors contributed equally

Running head: Phylogenetic uncertainty and Rubin's rules

*Correspondence: S. Nakagawa

e-mail: s.nakagawa@unsw.edu.au

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1216v1 | CC-BY 4.0 Open Access | rec: 6 Jul 2015, publ: 6 Jul 2015

Abstract

Phylogenetic comparative methods (PCMs), especially ones based on linear models, have played a central role in understanding species' trait evolution. These methods, however, usually assume that phylogenetic trees are known without error or uncertainty, but this assumption is most likely incorrect. So far, Markov chain Monte Carlo, MCMC-based Bayesian methods have successfully been deployed to account for such phylogenetic uncertainty in PCMs. Yet, the use of these methods seems to have been limited, probably due to difficulties in their implementation. Here, we propose an approach with which phylogenetic uncertainty is incorporated in a simple, readily implementable and reliable manner. Our approach uses Rubin's rules, which are an integral part of a standard multiple imputation procedure, often employed to recover missing data. In our case, we see the true phylogenetic tree as a missing piece of data, and apply Rubin's rules to amalgamate parameter estimates from a number of models using a set of phylogenetic trees (e.g. a Bayesian posterior distribution of phylogenetic trees). Using a simulation study, we demonstrate that our approach using Rubin's rules performs better in accounting for phylogenetic uncertainty than alternative methods such as MCMC-based Bayesian and Akaike information criterion, AIC-based model averaging approaches; that is, on average, our approach has the best 95% confidence/credible interval coverage among all. A unique property of the multiple imputation procedure is that the index, named 'relative efficiency', could be used to quantify the number of trees required for incorporating phylogenetic uncertainty. Thus, by using the relative efficiency, we show the required tree number is surprisingly small (~50 trees) at least in our simulation. In addition to these advantages above, our approach could be combined seamlessly with PCMs that utilize multiple imputation to recover missing data. Given the ubiquity of missing data, it is likely that the use of the multiple imputation procedure with Rubin's rules will be popular to deal with phylogenetic uncertainty as well as missing data in comparative data.

Keywords – Bayesian statistics; data augmentation; likelihood methods; phylogenetic comparative methods; missing data; model averaging, multiple imputation

(Introduction)

Phylogenetic comparative methods, PCMs, have been playing a central role in investigating trait evolution across species (reviewed in Garamszegi 2014). The most popular methods in comparative biology are based on linear regression such as independent contrasts (Felsenstein 1985), phylogenetic generalized least squares (PGLS; Grafen 1989), or phylogenetic (generalized) linear mixed models (Lynch 1991; Hadfield and Nakagawa 2010). When one phylogenetic tree is used in analysis, all these methods assume that the phylogeny of organisms is known without error.

However, no phylogenetic trees (or hypotheses) are known without error. Errors come in the form of uncertainty in branch length, topology, and also in the model of assumed character evolution. Researchers have been investigating the impact of these types of uncertainty on statistical inference (e.g., Díaz-Uriarte & Garland 1996; Symonds 2002). These studies generally suggest the importance of incorporating 'phylogenetic uncertainty' in PCMs; note that by using one tree, point estimates (e.g. regression coefficients) are not necessarily biased (Stone 2011), but uncertainty estimates (e.g. standard error or confidence intervals) are not accurate. Therefore, a number of methods have been proposed to include phylogenetic uncertainty (e.g. Losos 1994; Martin 1996; Housworth & Martin 2001; Huelsenbeck et al. 2000). Among these methods, probably the best one is to use Bayesian Markov Chain Mote Carlo, MCMC (Huelsenbeck et al. 2000; Huelsenbeck and Rannala 2003; de Villemereuil et al. 2012); the Bayesian MCMC methods utilize phylogenetic trees sampled from posterior tree set obtained from Bayesian phylogenetic tree estimation programs such as BEAST (Brummond and Rambaut 2007) and MrBayes (Ronquist and Huelsenbeck 2003).

Nonetheless, these methods are not always met with enthusiasm in the evolutionary biology community (cf. Pagel et al. 2004, Pagel and Meade 2006). Difficulties we see are two-fold: (i) currently, few easy-to-use implementations for such Bayesian MCMC methods are widely available, at least, for regression-based PCMs (but see Hadfield 2010, de Villemereuil et al. 2012); and (ii) even if implemented, Bayesian MCMC-based analysis may take a long time to process many phylogenetic trees (e.g., see Figure 6 in de Villemereuil et al. 2012). Recently, Garamszegi and Mundry (2014) have proposed a frequentist solution, which employs model averaging with Akaike information criterion (AIC) in PGLS incorporating many phylogenetic trees (see also Mahler et al. 2010). Such a method overcomes the aforementioned difficulties. However, Garamszegi and Mundry (2014) acknowledge the lack of theoretical basis for this proposal, and that theoretical or simulation-based confirmation of their method is necessary.

Here, we propose another solution to account for phylogenetic uncertainty. Our method is simple, generally applicable, and, what is more, it is fairly reliable and readily implementable (see below). Also, it is firmly based on missing data theory (reviewed in Little and Rubin 2002), and utilizes Rubin's rules, which have been proposed as a part of the multiple imputation procedure (Rubin 1987). Evolutionary biologists and ecologists have just recently recognized the usefulness of techniques based on missing data theory (reviewed in Nakagawa and Freckleton 2008; Nakagawa 2015). Also, the importance of these missing-data methods has been discussed in the phylogenetic comparative literature (e.g. Garamszegi and Møller, 2011; de Villemereuil and Nakagawa 2014). Especially, multiple imputation has been successfully employed in a number of comparative studies to recover missing data (e.g. Fisher et al. 2003; González-Suárez et al. 2012; Liker et al. 2014; Pollux et al. 2014). Yet, so far, nobody seems to have made a use of Rubin's rules to deal with phylogenetic uncertainty, only to get correct estimates and standard errors for the regression parameters.

Below, we first describe Rubin's rules associated with multiple imputation, and explain the rationale and potential advantages of our proposed method. Then, we conduct a simulation study using PGLS with a Bayesian posterior tree set to compare the performance of our proposed method to other methods such as methods using only one phylogenetic tree and the AIC-based method. We finish with a discussion, focusing on how our method can be combined seamlessly with PCMs that use multiple imputation to recover missing data.

Multiple imputation and Rubin's rules

Multiple imputation is a three-step process: imputing data, analyzing imputed data and pooling results. In the first step, *m* copies of 'complete' data sets are generated from an incomplete original data set. Popular techniques for the imputation steps use EM/EMB (expectation maximization with bootstrap) and MCMC algorithms, both of which are implemented in R packages such as Amelia (Honaker et al. 2011), mice (van Buuren and Groothuis-Oudshoorn 2011) and mi (Su et al. 2011); for more details regarding the algorithms, see Schafer (1997), Enders (2010) and van Buuren (2012). In the second step (analysis), we run separate statistical analyses on *m* data sets. In the final step (pooling), we use Rubin's rules (see below) to aggregate *m* sets of results to produce parameter estimates along with their uncertainty.

As an example of applying this three-step process to PCMs, let us assume that we have complete data for species traits (see Discussion for cases where missing data exist). Then, what remains missing is the 'true phylogenetic tree'; note that this is the central reason for us using (a part of) multiple imputation to account for phylogenetic uncertainty. Currently, a standard approach to creating candidate trees is to use Bayesian phylogenetic methods, as mentioned above, such as BEAST and MrBayes, which yield a posterior distribution of phylogenetic trees (for a guidance on building phylogenetic trees, see Garamszegi and Gonzalez-Voyer 2014). Alternatively, we can use published Bayesian tree sets as in Jetz et al. (2012) for birds, and Arnold et al. (2010) for primates. We consider this tree generation stage as our imputation step (the first step). The second step can be conducted using any frequentist or Bayesian statistical procedures including PCMs, such as independent contrasts, PGLS and phylogenetic mixed models. Say, we will run PGLS with *m* randomly sampled phylogenetic trees from a Bayesian posterior tree set, which will result in *m* sets of results. Then, by combining these result sets via Rubin's rules (the final step), we will have integrated phylogenetic uncertainty in our estimates from PGLS.

Rubin’s rules are a set of formulas for combining multiple statistical results, and they are as follows (Rubin 1987). With *m* imputations, parameters can be estimated by:

$$\bar{\mathbf{b}} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{b}^{j} \ , \qquad (1)$$

where $\bar{\mathbf{b}}$ is a *k* length vector and an average of $\mathbf{b}^j$, and $\mathbf{b}^j$ is the *j*th set (of *m*) of *k* parameter estimates (e.g. regression coefficients). An overall variance-covariance matrix of $\bar{\mathbf{b}}$ is obtained by:

$$\mathbf{V} = \bar{\mathbf{W}} + \left(1 + \frac{1}{m}\right)\mathbf{B} \, , \qquad (2)$$

$$\bar{\mathbf{W}} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{W}^{j} \, , \qquad (3)$$

$$\mathbf{B} = \frac{1}{m-1}\sum_{j=1}^{m}(\mathbf{b}^{j} - \bar{\mathbf{b}})(\mathbf{b}^{j} - \bar{\mathbf{b}})^{T} \ , \qquad (4)$$

where **V** is the overall (total) variance(-covariance) matrix for $\bar{\mathbf{b}}$, the within-imputation variance(-covariance) matrix, $\bar{\mathbf{W}}$ is the average of the variance-covariance matrix, $\mathbf{W}^j$ for **b**, and **B** is the between-imputation variance(-covariance) matrix for $\mathbf{b}^j$; note that the standard error of the *i*th parameter (of *k*) is $\sqrt{\mathbf{V}_{ii}}$ (subscript denotes the *i*th row and *i*th column, or *i*th diagonal element). Also, the term, (1+1/*m*) in Equation (2) can be seen as a correction for *m* not being infinite. An important concept in multiple imputation is called, ‘fraction of missing information’, usually denoted by *γ* and given by:

$$\bar{\gamma} = \left(1 + \frac{1}{m}\right)\mathrm{tr}\left(\mathbf{B}\mathbf{V}^{-1}\right)\frac{1}{k} \, , \qquad (5)$$

where $\bar{\gamma}$ is the initial estimate of the fraction of missing information, ranging from 0 to 1 (see below; cf. Equation (12)), and the term, tr($\mathbf{B}\mathbf{V}^{-1}$) denotes the trace of the resulting matrix from $\mathbf{B}\mathbf{V}^{-1}$. We can appreciate why $\bar{\gamma}$ is termed ‘the fraction of missing information’ because it represents a proportion of the between-imputation variance to the total (overall) variance (note that it may be

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1216v1 | CC-BY 4.0 Open Access | rec: 6 Jul 2015, publ: 6 Jul 2015

easier to see this in Equation (8) below). In other words, it represents the proportion of the parameter uncertainty due to using different trees. We can obtain statistical significance and confidence intervals based on *t* distributions with the degrees of freedom of the following:

$$\bar{v} = (m-1)\frac{1}{\bar{\gamma}^2}, \tag{6}$$

where $\bar{v}$ is the degrees of freedom to be used for *t* values ($\mathbf{b}_i / \sqrt{\mathbf{V}_{ii}}$). However, since the parameters will not be influenced equally by the phylogenetic uncertainty, it is probably better to obtain a fraction of missing information value for each parameter ($\mathbf{b}_i$) rather than omnibus values as in Equations (5 and 6) (Lipsitz et al. 2002). Such separate values of the degree of freedom (v_i) can be obtained by:

$$\gamma_i = \left(1 + \frac{1}{m}\right)\left(\frac{\mathbf{B}_{ii}}{\mathbf{V}_{ii}}\right), \tag{7}$$

$$v_i = (m-1)\frac{1}{\gamma_i^2}. \tag{8}$$

However, the formulation of v_i or $\bar{v}$ assumes a very large sample size, *n* (which is the length of data when no data are missing; Rubin and Schenker 1986; Rubin 1987). Barnard & Rubin (1986) proposed the following adjustment in the degrees of freedom (cf. Lipsitz et al. 2002):

$$v_i^* = \left(\frac{1}{v_i} + \frac{1}{v_{\mathrm{obs}(i)}}\right)^{-1}, \tag{9}$$

$$v_{\mathrm{obs}(i)} = (1-\gamma_i)\left(\frac{v_{\mathrm{com}}+1}{v_{\mathrm{com}}+3}\right)v_{\mathrm{com}}, \tag{10}$$

$$v_{\mathrm{com}} = n - k \tag{11}$$

where v_i^* is the degrees of freedom for *i*th parameter, especially suitable when sample size, *n* is small. The degrees of freedom, v_{obs} denotes the observed degrees of freedom, whereas v_{com} denotes the complete degrees of freedom (i.e. the degrees of freedom for the complete data set assuming no missing data). In the next section, we will compare the performance of both v_i (hereafter denoted "original df") and v_i^* (hereafter denoted "corrected df").

Once we have an estimate of the corrected degrees of freedom, we can obtain a refined estimate of the fraction of missing information, γ_i^* for each parameter:

$$\gamma_i^* = \left(1+\frac{1}{m}\right)\frac{\mathbf{B}_{ii}}{\mathbf{V}_{ii}} + \frac{2}{(v_i^*+3)\mathbf{V}_{ii}} . \tag{12}$$

Then, we can use γ_i^* to find a very useful quantity called 'relative efficiency', which is given by:

$$\varepsilon_i = \left(1+\frac{\gamma_i^*}{m}\right)^{-1} , \tag{13}$$

where ε_i is relative efficiency of the *i*th parameter and ranges from 0 to 1. Relative efficiency represents the efficacy of multiple imputation process, compared to the case of *m* being infinite. In other words, this number can be used to assess how many imputations (*m*) are needed to account for uncertainty due to missing data. In our case, relative efficiency can indicate how many phylogenetic trees we should use for analysis (typically, the number of required trees to account for phylogenetic uncertainty are chosen arbitrarily). Notably, to achieve fairly high relative efficiency, the required number of *m* is surprisingly low, even when the fraction of missing information is relatively large. For example, with γ= 0.5 and m = 5, relative efficiency is 90.91%, while it is 95.24% when γ = 0.5 and m = 10. Rubin's (1987) initial recommendation of *m* was low (3-10) probably due to computational limitation at that time, but current thinking is to use much larger *m*, aiming at over 99% relative efficiency (e.g. Graham et al. 2007, von Hippel 2009, Nakagawa 2015). As you see in Equation (13), we obtain a relative efficiency value (ε_i) for every parameter and such values vary among parameters. For assessing efficiency of a model, we will use the relative efficiency (ε^*) that is obtained from the largest value of the fraction of missing information, following McKnight et al. (2007); that is:

$$\varepsilon^{*} = \left(1 + \frac{\max(\gamma_i^{*})}{m}\right)^{-1}, \qquad (14)$$

where $\max(\gamma_i^{*})$ denotes the maximum (largest) value of γ_i^{*}; the use of the maximum value of γ_i^{*} ensures all parameters will achieve at least a certain relative efficiency level or above. We can easily automate calculations involving the above formulae with currently available R packages for multiple imputation such as mice (reviewed in Nakagawa and Freckleton 2011; see also Penone et al. 2014).

A simulation study

In order to assess the overall quality of our new method and compare it to existing ones, we performed a simulation study using 100 trees estimated from a real data set, using BEAST (reduced species and tree samples from Wells et al., Submitted; see also de Villemereuil et al., 2012). We simulated data sets in which a variable *y* was linearly predicted from a variable *x*, with an intercept of 5 and a slope of 2. The error structure of this relationship was constrained by a phylogenetic tree among the 100 trees (hereafter called the 'true tree'), following a Brownian motion model. We simulated the data sets either using a small sample size (10 species, chosen to keep a strong phylogenetic structure; Fig. S1) or a larger one (50 species, Fig. S2, which is the same as used in de Villemereuil et al. 2012) and different residual standard deviation (sigma, σ from 2 to 15). Hence the simulation scheme used here was exactly identical to the one used in de Villemereuil et al. (2012). We compared GLS using the true tree or two types of consensus trees (majority rule or consensus), with both multiple GLS with pooling of the results using AIC averaging (as in Garamszegi and Mundry 2014) and pooling with Rubin's rules as described above (either using the original degrees of freedom, df, or the corrected df as in Equation (9)). We also compared the present results with the simulation results on the Bayesian methods of de Villemereuil *et al.* (2012). We simulated 10,000 replicates for each condition. Note that in the case of the Bayesian methods, only the results based on 3,000 replicates of 50 species are available.

As shown in Fig. 1, the estimation of the intercept and the slope was accurate for all methods, but the accuracy of the residual variance (σ^2, depicted as the residual standard deviation σ in Fig. 1) differed greatly between the methods. Especially, the two methods using consensus trees overestimated the residual variance (σ^2), even more so as the sample size became larger. Regarding the precision (i.e. estimates sampling variance), it was lower for larger residual variances (σ^2) and lower sample size, as expected. The precision of the estimates also differed between the methods: again the two methods using consensus trees yielded very imprecise estimates, whereas the precision of the other methods was comparable. All methods incorporating phylogenetic uncertainty were very similar in terms of precision.

However, as shown in Fig. 2, the coverage of the confidence intervals, CIs (credible intervals for the Bayesian method) were very different between these methods. The CI coverage of the methods using consensus trees was very liberal (note that we focus on the slope as this is the parameter of prime interest). The CI coverage of the methods using Rubin's rules depended strongly on the type of degrees of freedom used. Whereas the original df yielded anti-conservative confidence intervals for N=10 species and perfectly calibrated ones for N=50, the CIs obtained from the corrected df were always conservative, and especially so for N=10. The confidence interval coverage of the AIC model averaging method was much like those of Rubin's rule with the original df, but were slightly more liberal. In particular, they were not perfectly calibrated for N=50 species. Note that the Bayesian method was even slightly more liberal (discrepancy from the 5% expectancy was less often significant for this method, as there were only 3,000 replicates).

In order to assess the behavior of the proposed method using Rubin's rules to account for phylogenetic uncertainty, we also conducted a study using different sample size for the trees (T=10, 20, 50 or 100, all sets including the true tree) and computed the relative efficiency as shown in Equation (14). This analysis revealed two interesting patterns (Fig. 3). First, no efficiency lower than 0.93 was recorded for a total of 80,000 simulated data sets, even for a sample size of trees as low as T=10. Accordingly, the number of trees used had little effect on the coverage of confidence intervals, even for T=10 (Fig. S3-4). Second, in order to reach a relative efficiency over 0.99, on average, only 50 trees were necessary. With 100 trees, the relative efficiency was always over 0.99. All the results above (both regarding efficiency and coverage) were not changed when we used random sets of trees excluding the true tree (Fig. S5-7). Hence, a relatively small number of trees are sufficient to account for the phylogenetic uncertainty very efficiently, at least, in our simulation scheme.

Discussion

The aim of this article is to introduce a simple and generally applicable method to account for phylogenetic uncertainty in phylogenetic comparative methods, PCMs. Via a simulation study using a simple PGLS, we compared the proposed method using Rubin's rules with other existing methods, and we also assessed the number of trees required to accurately account for phylogenetic uncertainty. Two main results have emerged from our simulation study.

First, all methods accounting for phylogenetic uncertainty performed fairly well although not perfectly. In contrast, methods ignoring phylogenetic uncertainty performed poorly. These findings are concordant with the previous work by de Villemereuil et al. (2012). Our proposed method using Rubin's rules, and with the original degrees of freedom (df), provided the expected coverage of the confidence interval, CI, for the slope with N = 50. Thus, when sample sizes are fairly high, i.e. over 50, we recommend the use of Rubin's rules with the original df. However, with sample size less than 50, one may want to use Rubin's rules with corrected df, bearing on mind that the CI coverage may be slightly wider than expected (i.e. conservative).

The second main result is that the number of phylogenetic trees needed to correct for phylogenetic uncertainty is surprisingly low. The required number of trees is far less than 1000 (as in Garamszegi and Mundry 2014), and probably less than 100 (as in de Villemereuil et al. 2012). It is likely to be a matter of dozens. In our simulation, regardless of the inclusion of the 'true tree', sets of 50 randomly selected trees achieved, on average, over 99% relative efficiency; in other words, the using 50 trees should be almost as good as using an infinite number of trees. Thus, we recommend the use of over 50 phylogenetic trees in a PCM to account for phylogenetic uncertainty. However, for any given analysis and tree set, we recommend checking the number of trees needed to reach a relative efficiency of 99% (Nakagawa 2015). In practice, indeed, the required number of trees required to achieve high efficiency will strongly depend on the phenotypic data (e.g. phylogenetic signal), the complexity of the model and the variability in the tree estimates (e.g. strong topological and branch length uncertainty). We note that the statistical literature has discussed other criteria apart from the relative efficiency to determine how many imputations one requires (see Graham et al. 2007; White 2009).

As mentioned, the MCMC-based Bayesian method (e.g. de Villemereuil et al. 2012) and AIC model averaging method (Garamszegi and Mundry 2014) accounted for phylogenetic uncertainty fairy well, although both produce slightly liberal credible/confidence intervals, CIs. The method based on Rubin's rules (or multiple imputation) has the advantage of superior performance and simplicity over these two other methods at least in terms of incorporating phylogenetic uncertainty. This is, given that the imputation step is 'proper', which is the case here as long as the trees come from a Bayesian posterior distribution and the estimates are Maximum Likelihood Estimators (e.g. BEAST/PGLS combination, for example; for the definition on proper multiple imputation, see Rubin 1987 and Nielsen 2003). However, there is another clear benefit of using the proposed method. That is, multiple imputation can simultaneously handle missing data and phylogenetic uncertainty in a comparative data set. Given the pervasive nature of missing data, we suggest multiple imputation may be useful for virtually every comparative data set (Nakagawa and Freckelton 2008; Garamszegi and Møller, 2011).

Notably, implementing multiple imputation with comparative data is a technically complex affair, because the imputation process requires the inclusion of a correlation matrix based on phylogeny. Thus, it requires a special package like PhyloPars (Bruggeman et al. 2009), and multiple imputation of comparative data was not possible with other general-purpose multiple imputation packages, which provide more flexibility over the former. However, Penone et al. (2014) recently showed that phylogenetic information could be added to multiple imputation in the form of phylogenetic eigenvectors (Diniz et al. 1998; see also Guénard et al. 2003), which can be seen as additional predictor variables (for the imputation step not for the analysis step). This means that, to conduct multiple imputation for comparative data, one can use general and flexible packages such as mice (van Buuren and Groothuis-Oudshoorn 2011), as was used for our simulation. However, more work is necessary to confirm such multiple imputation using phylogenetic eigenvectors are actually comparable to ones including phylogenetic correlation matrices.

The procedure known as 'data augmentation' can also be used for dealing with missing data. We note that the term data augmentation is used in a number of ways in the statistical literature, but here we follow the usage by McKnight et al. (2007); that is, in this procedure, uncertainty of missing data is incorporated in to parameter estimates during analysis, thus, including methods like the full information maximum likelihood (FIML) method (see Enders 2001; for the original usage of the term, data augmentation, see Tanner and Wong 1987). A data augmentation procedure is implemented, for instance, in MCMCglmm (Hadfield 2010). However, there is one disadvantage to data augmentation that does not affect multiple imputation. Data augmentation assumes the use of just identified or over-identified models (Enders 2001, 2010). That is, a particular model (for imputation) includes enough or more predictor variables, so that missing values can be recovered accurately from these predictors.

In contrast, because multiple imputation separates the steps of data imputation and analysis, we do not need to clutter a statistical model for analysis (i.e. the analysis step) with many variables, which assist in recovering missing values (known as auxiliary variables; Enders 2010; Nakagawa 2015). Technically speaking, auxiliary variables are supported to make missing values to fulfill the assumption of ‘missing at random’ (Little and Rubin 2002; for an accessible account, see Nakagawa and Freckelton 2008). In a multiple imputation procedure, we need add auxiliary variables only to a statistical model for imputation (i.e. the imputation step). For example, known data on species body size can be used during the imputation step to help recover missing data on species longevity, given the strong correlation between the two. However and importantly, because multiple imputation separates imputation and analysis, body size does not need to be a part of the final model. The use of multiple imputation probably has wider applications over data augmentation. Most importantly, to integrate phylogenetic uncertainty in a comparative data set with missing data, one just needs to conduct extra imputations (e.g. more m as in Equation (1)) to include the adequate number of trees (e.g. 50).

In conclusion, the method using Rubin’s rules is readily usable for all comparative biologists, and it can be integrated with both frequentist and Bayesian PCMs alike. Clearly, the use of multiple imputation is extremely useful not only for imputing missing data in PCMs, but also for integrating phylogenetic uncertainty, as we have shown above. As yet, we are unaware of any study combining the standard merits of multiple imputation with phylogenetic uncertainty in comparative analysis. However, we expect such a dual use to be common in the near future.

Acknowledgements

We thank Travis Ingram, Losia Lagisz, Alistair Senior and Simon Blomberg for comments, which have improved the manuscript. We are also grateful to Jessie A. Wells for letting us use the phylogenetic trees. SN was supported by a Rutherford Discovery Fellowship (New Zealand) and a Future Fellowship (Australia).

References

Arnold C, Matthews LJ, Nunn CL. 2010. The 10k Trees Website: A New Online Resource for Primate Phylogeny. Evol Anthropol, 19:114-118.

Barnard J, Rubin DB. 1999. Small-sample degrees of freedom with multiple imputation. Biometrika, 86:948-955.

Bruggeman J, Heringa J, Brandt BW. 2009. PhyloPars: estimation of missing parameter values using phylogeny. Nucleic Acids Res, 37:W179-W184.

de Villemereuil P, Nakagawa S. 2014. General Quantitative Genetic Methods for Comparative Biology. In: Garamszegi LZ editor. Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology, Springer Berlin Heidelberg, p. 287-303.

de Villemereuil P, Wells JA, Edwards RD, Blomberg SP. 2012. Bayesian models for comparative analysis integrating phylogenetic uncertainty. Bmc Evol Biol, 12.

Diaz-Uriarte R, Garland T. 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: Sensitivity to deviations from Brownian motion. Syst Biol, 45:27-47.

Diniz JAF, De Sant'ana CER, Bini LM. 1998. An eigenvector method for estimating phylogenetic inertia. Evolution; international journal of organic evolution, 52:1247-1262.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. Bmc Evol Biol, 7.

Enders CK. 2010. Applied missing data analysis. New York, Guilford Press.

Enders CK, Bandalos DL. 2001. The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. Struct Equ Modeling, 8:430-457.

Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat, 125:1-15.

Fisher DO, Blomberg SP, Owens IPF. 2003. Extrinsic versus intrinsic factors in the decline and extinction of Australian marsupials. Proceedings of the Royal Society of London Series B-Biological Sciences, 270:1801-1808.

Garamszegi L, Gonzalez-Voyer A. 2014. Working with the Tree of Life in Comparative Studies: How to Build and Tailor Phylogenies to Interspecific Data sets. In: Garamszegi LZ editor. Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology, Springer Berlin Heidelberg, p. 19-48.

Garamszegi L, Mundry R. 2014. Multimodel-Inference in Comparative Analyses. In: Garamszegi LZ editor. Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology, Springer Berlin Heidelberg, p. 305-331.

Garamszegi LZ. 2014. Modern phylogenetic comparative methods and their application in evolutionary biology. New York, Springer, p. pages cm.

Garamszegi LZ, Moller AP. 2011. Nonrandom Variation in Within-Species Sample Size and Missing Data in Phylogenetic Comparative Studies. Syst Biol, 60:876-880.

Gonzalez-Suarez M, Lucas PM, Revilla E. 2012. Biases in comparative analyses of extinction risk: mind the gap. J Anim Ecol, 81:1211-1222.

Grafen A. 1989. The Phylogenetic Regression. Philos T Roy Soc B, 326:119-157.

Graham JW, Olchowski AE, Gilreath TD. 2007. How many imputations are really needed? - Some practical clarifications of multiple imputation theory. Prev Sci, 8:206-213.

Guénard G, Legendre P, Peres-Neto PR (2013) Phylogenetic eigenvector mapping: a framework to model and predict species trait. Methods Ecol Evol, 4: 1120-1131.

Hadfield J. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J Stat Softw, 33:1 - 22.

Hadfield JD, Nakagawa S. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J Evolution Biol, 23:494-508.

Honaker J, King G, Blackwell M. 2011. Amelia II: A Program for Missing Data. J Stat Softw, 45:1-47.

Huelsenbeck JP, Rannala B. 2003. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. Evolution; international journal of organic evolution, 57:1237-1247.

Huelsenbeck JP, Rannala B, Masly JP. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. Science, 288:2349-2350.

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012. The global diversity of birds in space and time. Nature, 491:444-448.

Liker A, Freckleton RP, Szekely T. 2014. Divorce and infidelity are associated with skewed adult sex ratios in birds. Curr Biol, 24:880-884.

Lipsitz SR, Parzen M, Zhao LP. 2002. A degrees-of-freedom approximation in multiple imputation. J Stat Comput Sim, 72:309-318.

Little RJA, Rubin DB. 2002. Statistical analysis with missing data. 2nd ed. Hoboken, N.J., Wiley.

Losos JB. 1994. An Approach to the Analysis of Comparative Data When a Phylogeny Is Unavailable or Incomplete. Syst Biol, 43:117-123.

Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. Evolution; international journal of organic evolution, 45:1065-1080.

Mahler DL, Revell LJ, Glor RE, Losos JB. 2010. Ecological Opportunity and the Rate of Morphological Evolution in the Diversification of Greater Antillean Anoles. Evolution; international journal of organic evolution, 64:2731-2745.

McKnight PE, McKnight KM, Sidani S, Figueredo AJ. 2007. Missing data: a gentle introduction. New York, NY, The Guilford Press.

Nakagawa S. 2015. Missing data: mechanisms, methods and messages In: Fox GA, Negrete-Yankelevich S, Sosa VJ editors. Ecological Statistics: contemporary theory and application. Oxford, Oxford University Press, p. 81-105.

Nakagawa S, Freckleton R. 2011. Model averaging, missing data and multiple imputation: A case study for behavioural ecology. Behav Ecol Sociobiol, 65:103-116.

Nakagawa S, Freckleton RP. 2008. Missing inaction: The dangers of ignoring missing data. Trends Ecol Evol, 23:592-596.

Nielsen SF. 2003. Proper and improper multiple imputation. Int Stat Rev, 71:593-607.

Pagel M, Meade A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am Nat, 167:808-825.

Pagel M, Meade A, Barker D. 2004. Bayesian estimation of ancestral character states on phylogenies. Syst Biol, 53:673-684.

Penone C, Davidson AD, Shoemaker KT, Di Marco M, Rondinini C, Brooks TM, Young BE, Graham CH, Costa GC. 2014. Imputation of missing data in life-history trait data sets: which approach performs the best? Methods in Ecology and Evolution, 5:961-970.

Pollux BJA, Meredith RW, Springer MS, Garland T, Reznick DN. 2014. The evolution of the placenta drives a shift in sexual selection in livebearing fish. Nature, 513:233-236.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 19:1572-1574.

Rubin DB. 1987. Multiple imputation for nonresponse in surveys. New York, NY, J. Wiley & Sons.

Rubin DB, Schenker N. 1986. Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. J Am Stat Assoc, 81:366-374.

Schafer JL. 1997. Analysis of incomplete multivariate data. London, Chapman & Hall.

Stone EA. 2011. Why the Phylogenetic Regression Appears Robust to Tree Misspecification. Syst Biol, 60:245-260.

Su YS, Gelman A, Hill J, Yajima M. 2011. Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. J Stat Softw, 45:1-31.

Symonds MRE. 2002. The effects of topological inaccuracy in evolutionary trees on the phylogenetic comparative method of independent contrasts. Syst Biol, 51:541-553.

Tanner MA, Wing HW. 1987. The Calculation of Posterior Distributions by Data Augmentation. J Am Stat Assoc, 82:528-540.

van Buuren S. 2012. Flexible imputation of missing data. Boca Raton, FL, CRC Press.

van Buuren S, Groothuis-Oudshoorn K. 2011. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw, 45:1-67.

Wells JA, Blomberg SP, Scharaschkin T, Lowe AJ. Submitted. Time trees for the Wet Tropics: Effects of dates and densities on divergence times for Australia's rainforest flora. Plos One.

White IR, Royston P, Wood AM. 2011. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med, 30:377-399.

Figure legends

Figure 1. The distributions of estimates for the intercept, slope and residual standard deviation parameters for the different methods, different simulated standard deviation (σ) and different sample sizes (N=10, blue; N=50, red). The simulated intercept is 5 and the simulated slope is 2. The boxes depict the 50% inter-quantile interval, the whiskers depict the 95% inter-quantile interval and the horizontal bar is the average estimate. The number of replicates is 10,000, except for the Bayesian method with only 3,000 replicates.

Figure 2. Complementary of the coverage (1 - coverage) for 95% confidence/credible intervals for the different methods, different simulated standard deviation (σ) and different sample sizes (N=10, blue; N=50, red). Expected values of 1 – coverage (5%) are depicted by a dashed line. Stars show significant discrepancy (binomial test) from 5% assuming 10,000 replicates (six-branch stars) or 3,000 (five-branch stars, only for the Bayesian method).

Figure 3. Relative efficiency distribution for different tree sample size (T), different simulated standard deviation (σ) and different sample sizes (N). The boxes depict the 50% inter-quantile interval, the whiskers depict the 95% inter-quantile interval and the horizontal bar is the average estimate. The red lower dot is the minimal relative efficiency yielded during the simulations. The number of replicates is 10,000.

Figure S1. A visual representation of the ‘true tree’ for the 10 species used for the simulation study.

Figure S2. A visual representation of the ‘true tree’ for the 50 species used for the simulation study.

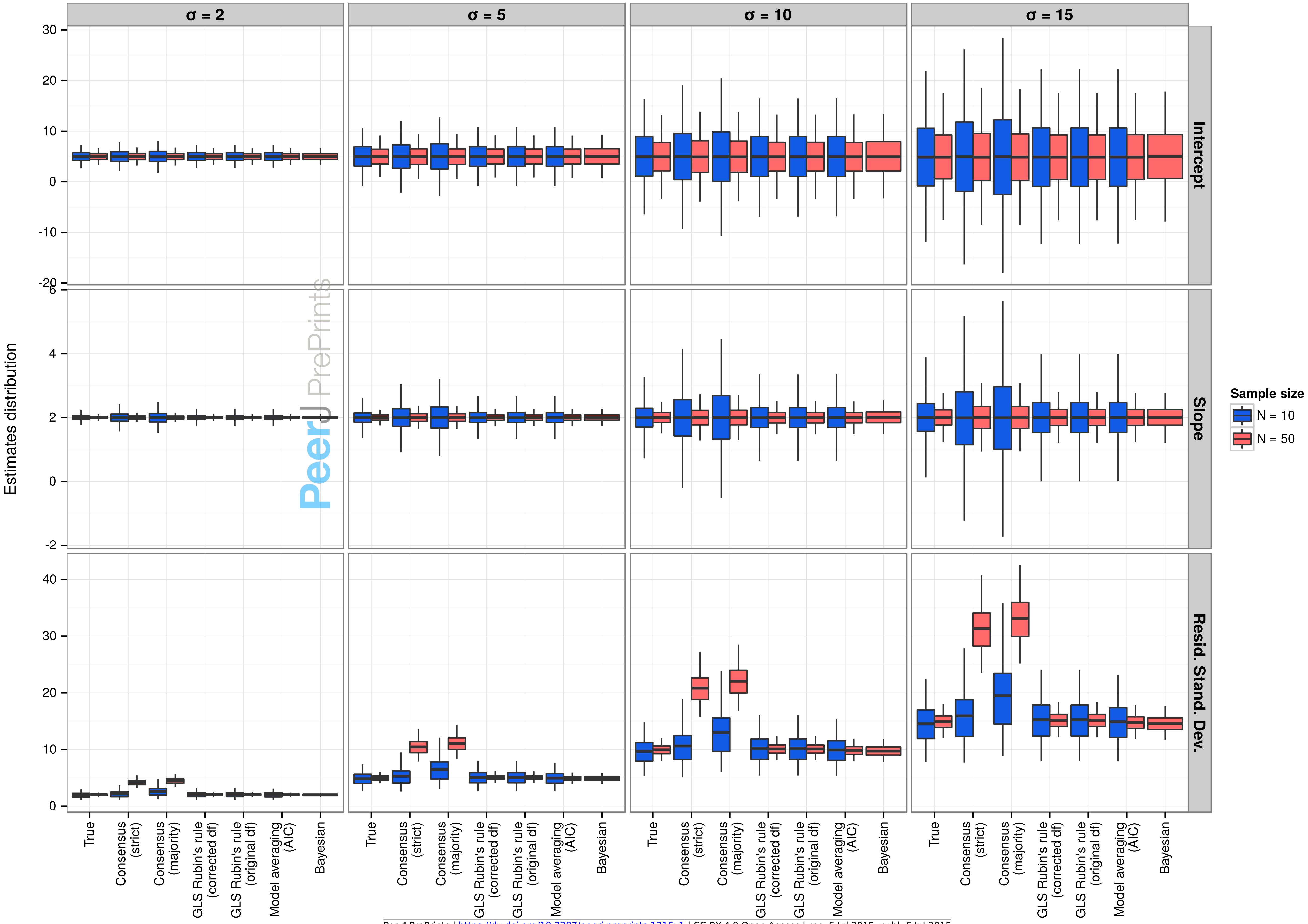
Figure S3. The distribution of estimates for the relative efficiency simulation, against different tree sample size (T), different simulated standard deviation (σ) and different sample size (N=10, blue; N=50, red). The simulated intercept is 5 and the simulated slope is 2. The boxes depict the 50% inter-quantile interval, the whiskers depict the 95% inter-quantile interval and the horizontal bar is the average estimate. The number of replicates is 10,000.

Figure S4. Complementary of the coverage (1 - coverage) for 95% confidence/credible intervals for the relative efficiency simulation, against different tree sample size (T), different simulated standard deviation (σ) and different sample sizes (N=10, blue; N=50, red). The number of replicates is 10,000.

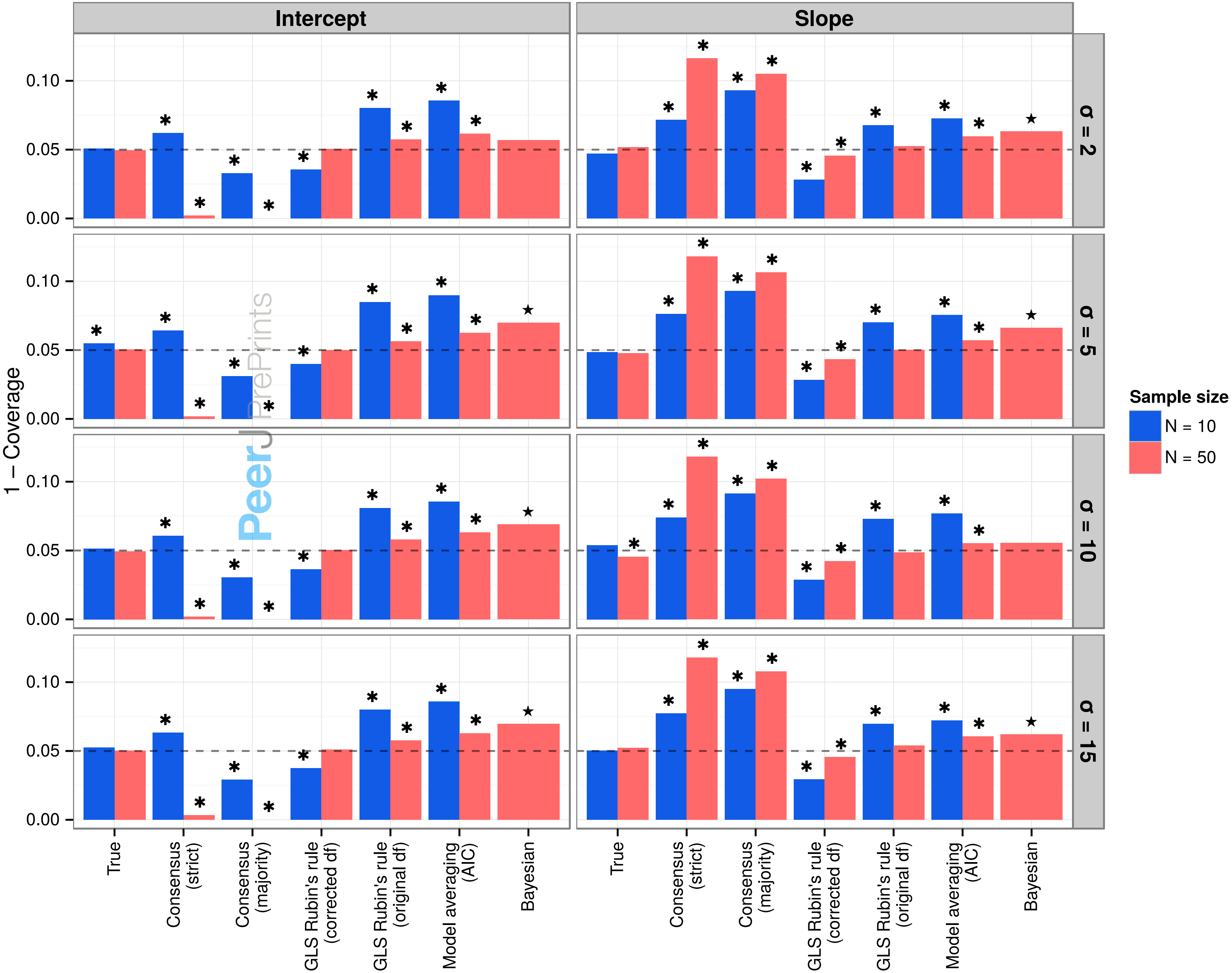
Figure S5. Relative efficiency distribution for random tree sets excluding the true tree, against different tree sample size (T), different simulated standard deviation (σ) and different sample sizes (N). The boxes depict the 50% inter-quantile interval, the whiskers depict the 95% inter-quantile interval and the horizontal bar is the average estimate. The red lower dot is the minimal relative efficiency yielded during the simulations. The number of replicates is 10,000.

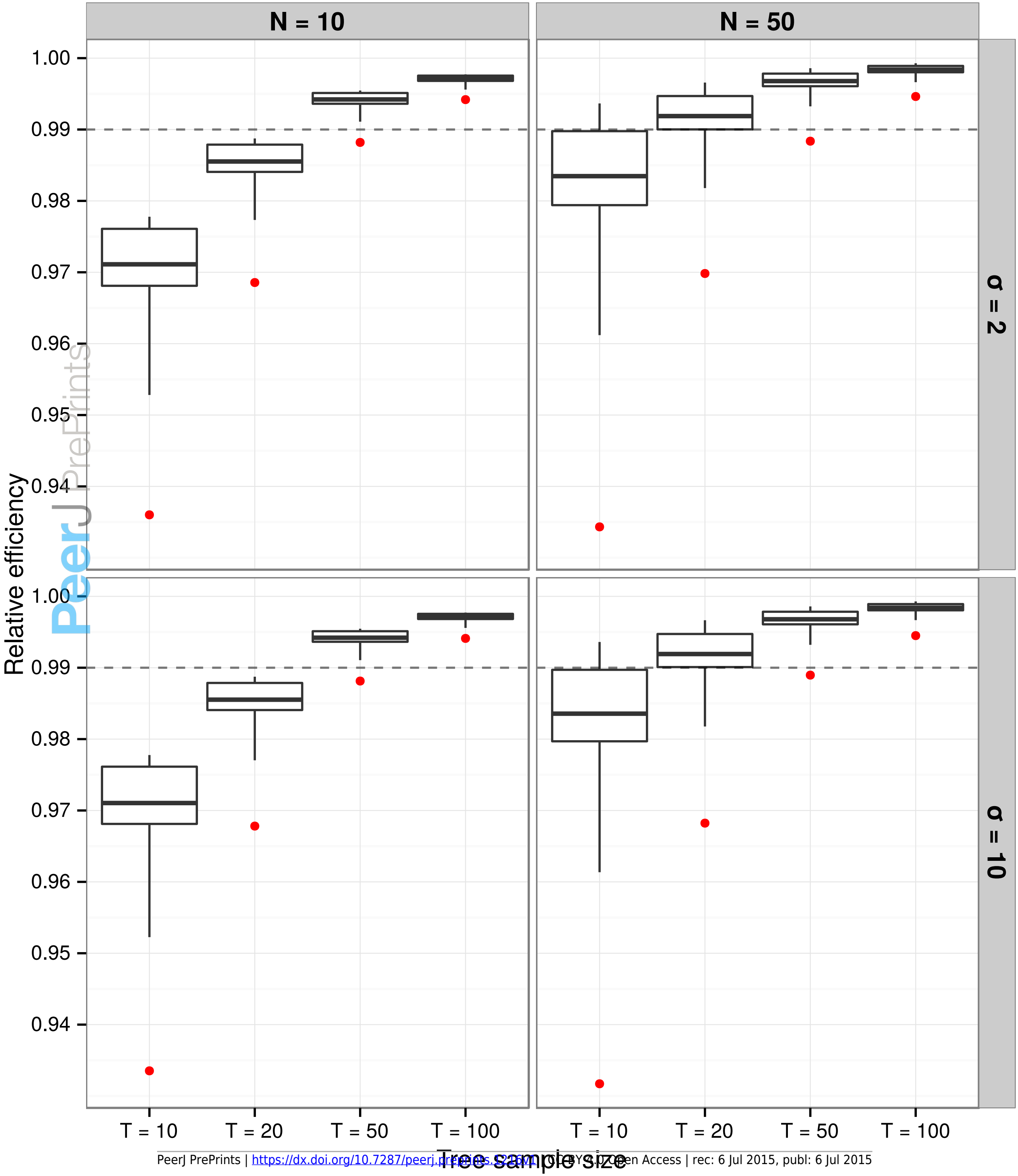
Figure S6. The distribution of estimates for random tree sets excluding the true tree, against different tree sample size (T), different simulated standard deviation (σ) and different sample size (N=10, blue; N=50, red). The simulated intercept is 5 and the simulated slope is 2. The boxes depict the 50% inter-quantile interval, the whiskers depict the 95% inter-quantile interval and the horizontal bar is the average estimate. The number of replicates is 10,000.

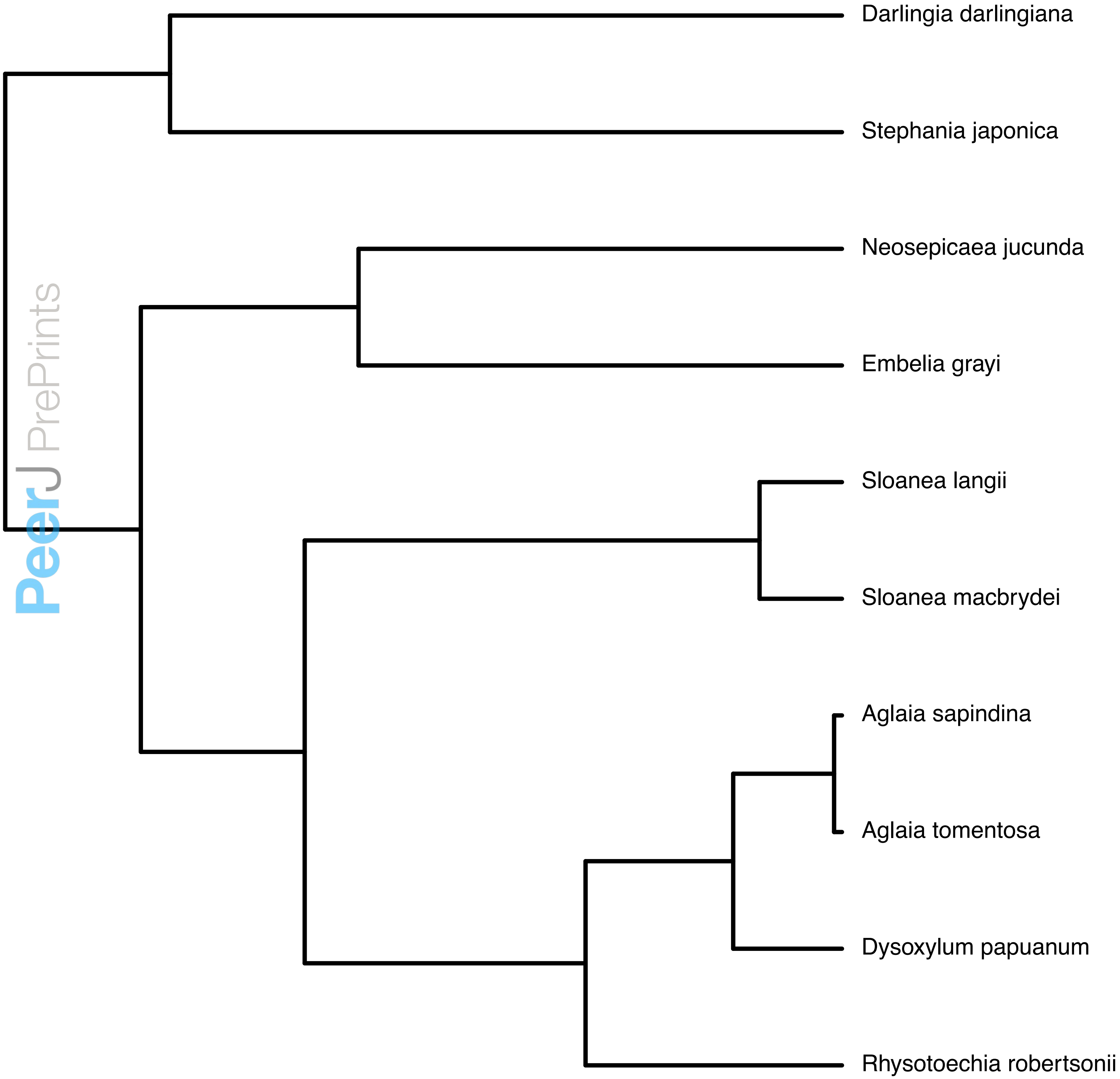
Figure S7. Complementary of the coverage (1 - coverage) for 95% confidence/credible intervals for random tree sets excluding the true tree, against different tree sample size (T), different simulated standard deviation (σ) and different sample sizes (N=10, blue; N=50, red). The number of replicates is 10,000.



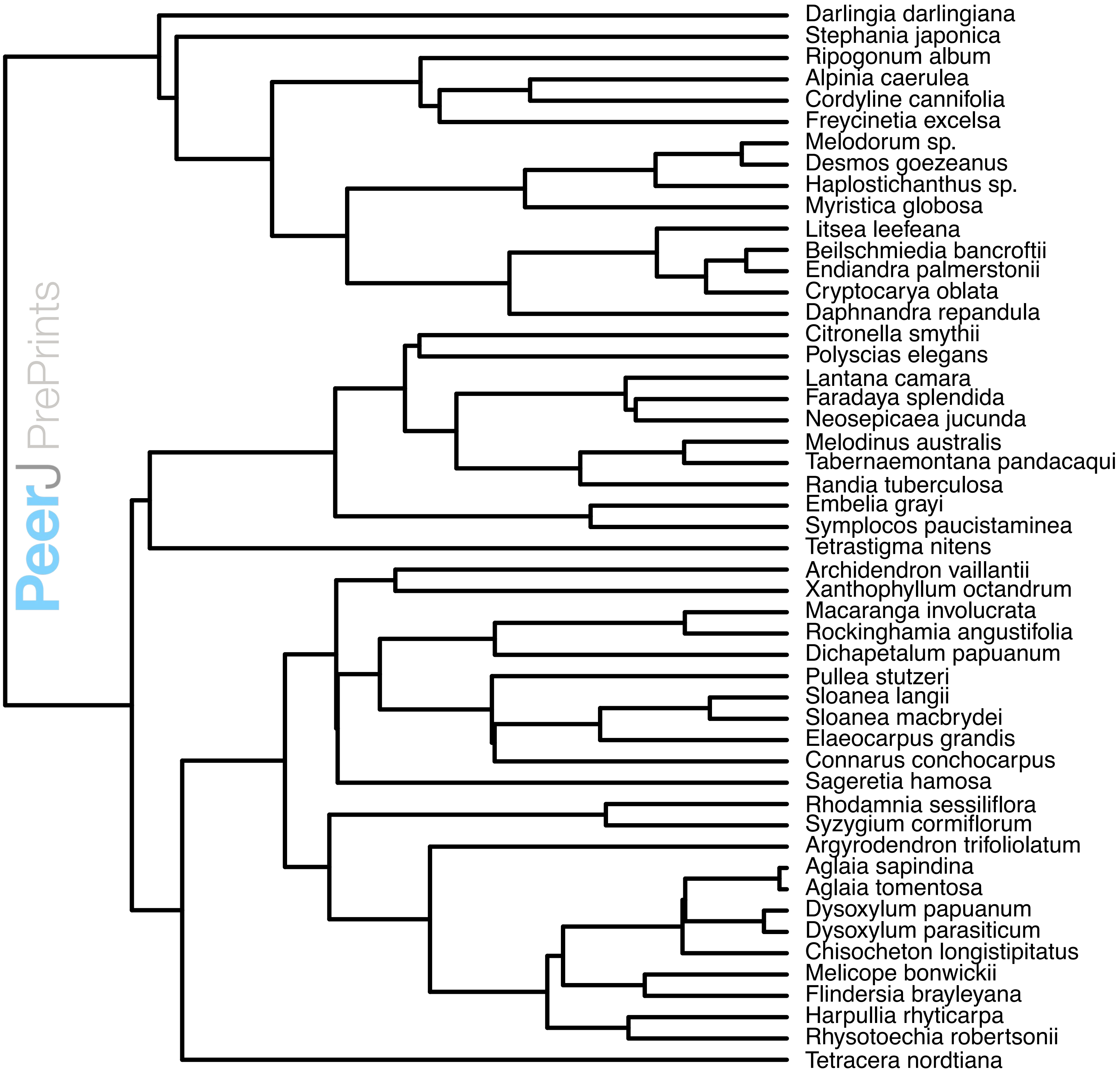
PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1216v1 | CC-BY 4.0 Open Access | rec: 6 Jul 2015, publ: 6 Jul 2015

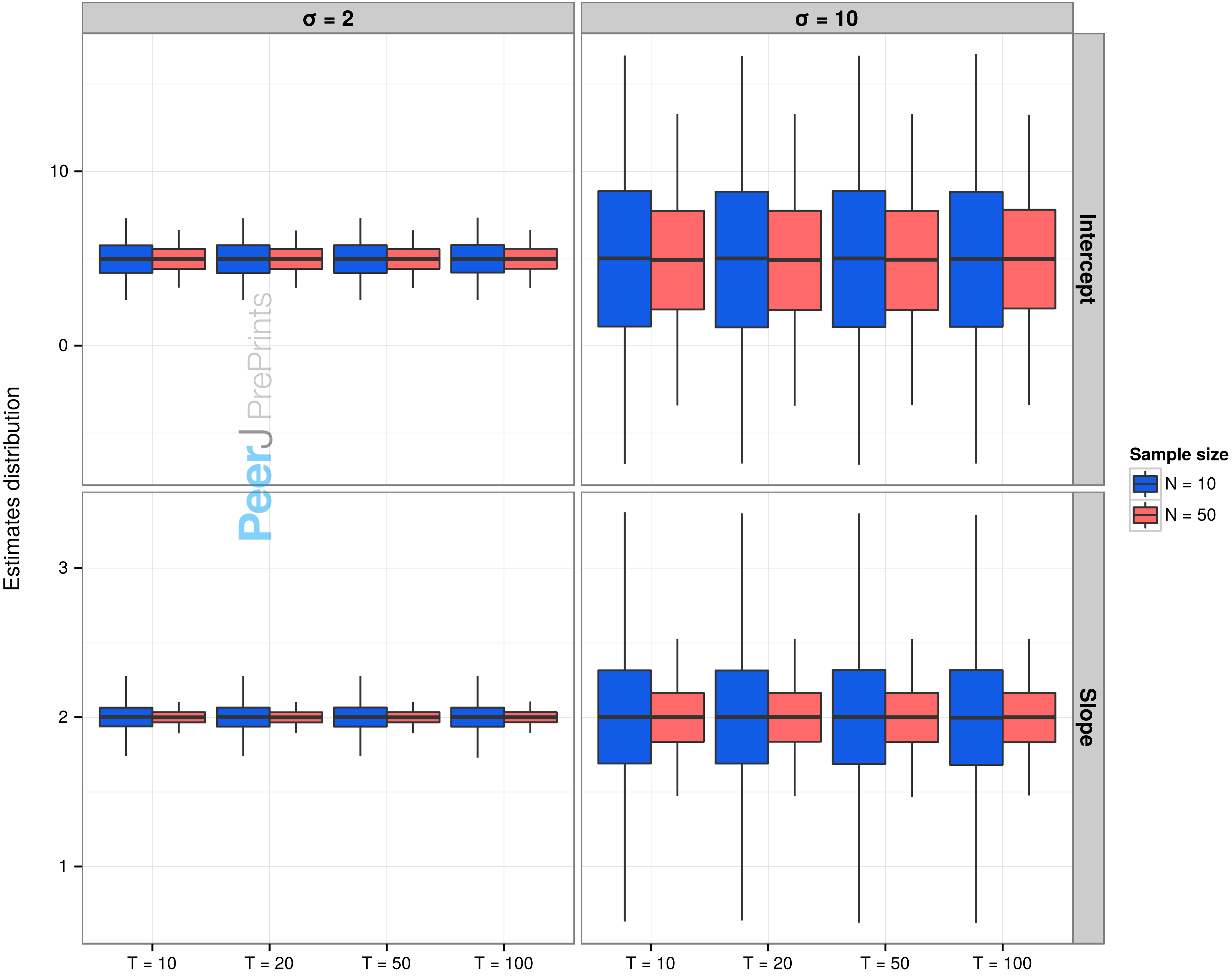

Intercept
Slope
σ = 2
σ = 5
σ = 10
σ = 15
1 – Coverage
0.00
0.05
0.10
True
Consensus (strict)
Consensus (majority)
GLS Rubin's rule (corrected df)
GLS Rubin's rule (original df)
Model averaging (AIC)
Bayesian
Sample size
N = 10
N = 50
*
★

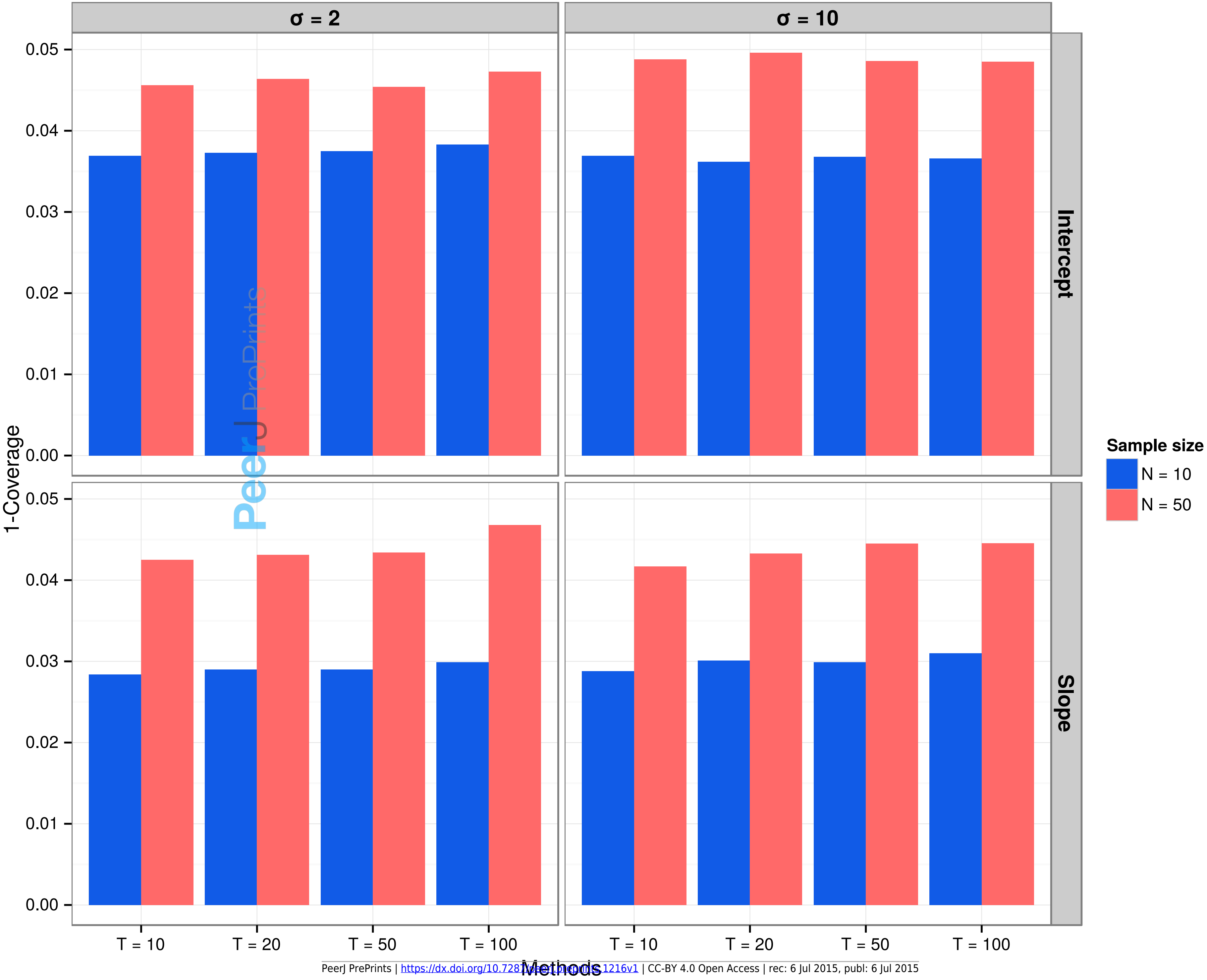

N = 10
N = 50
σ = 2
σ = 10
Relative efficiency
1.00
0.99
0.98
0.97
0.96
0.95
0.94
T = 10
T = 20
T = 50
T = 100
Tree sample size

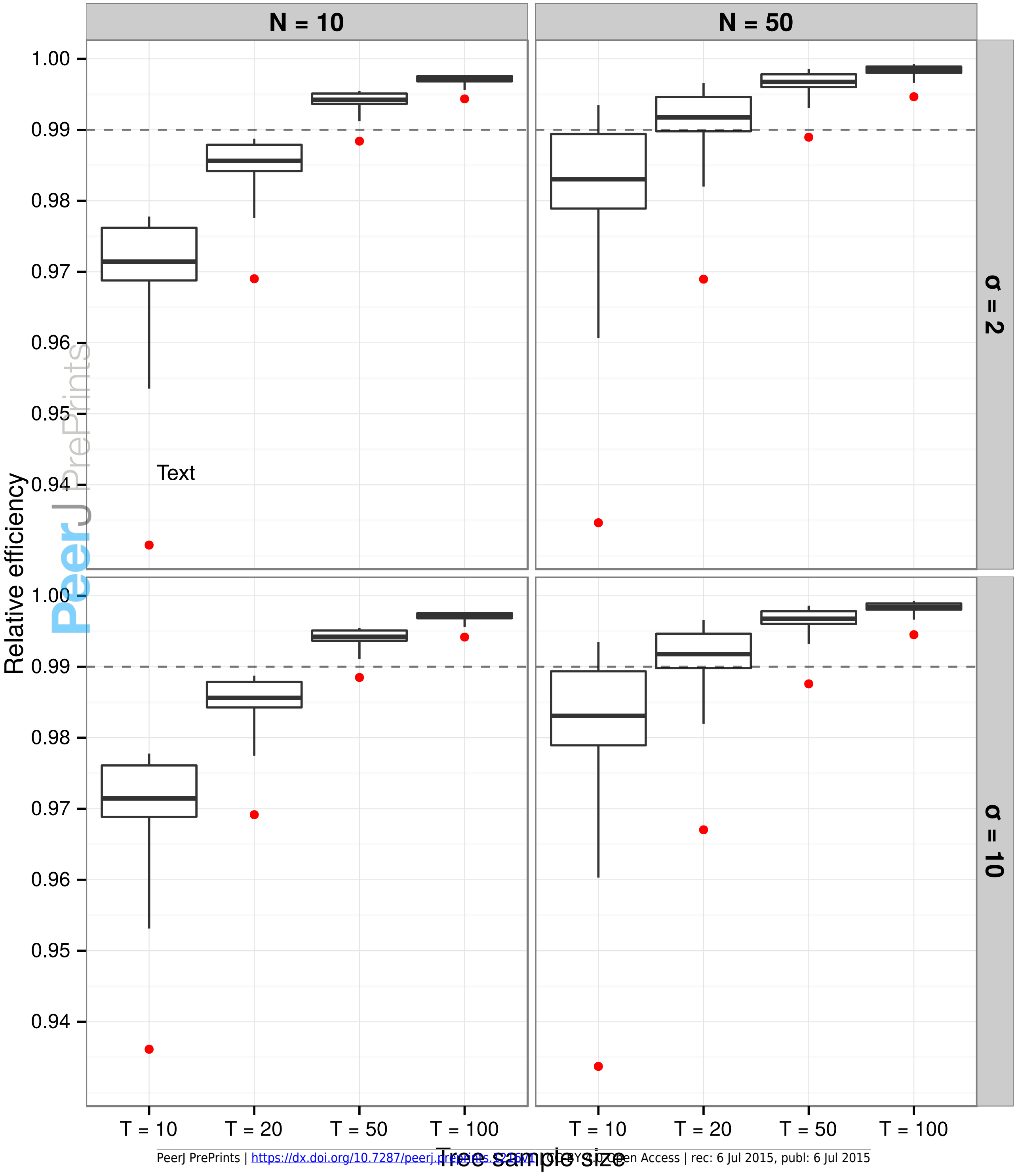


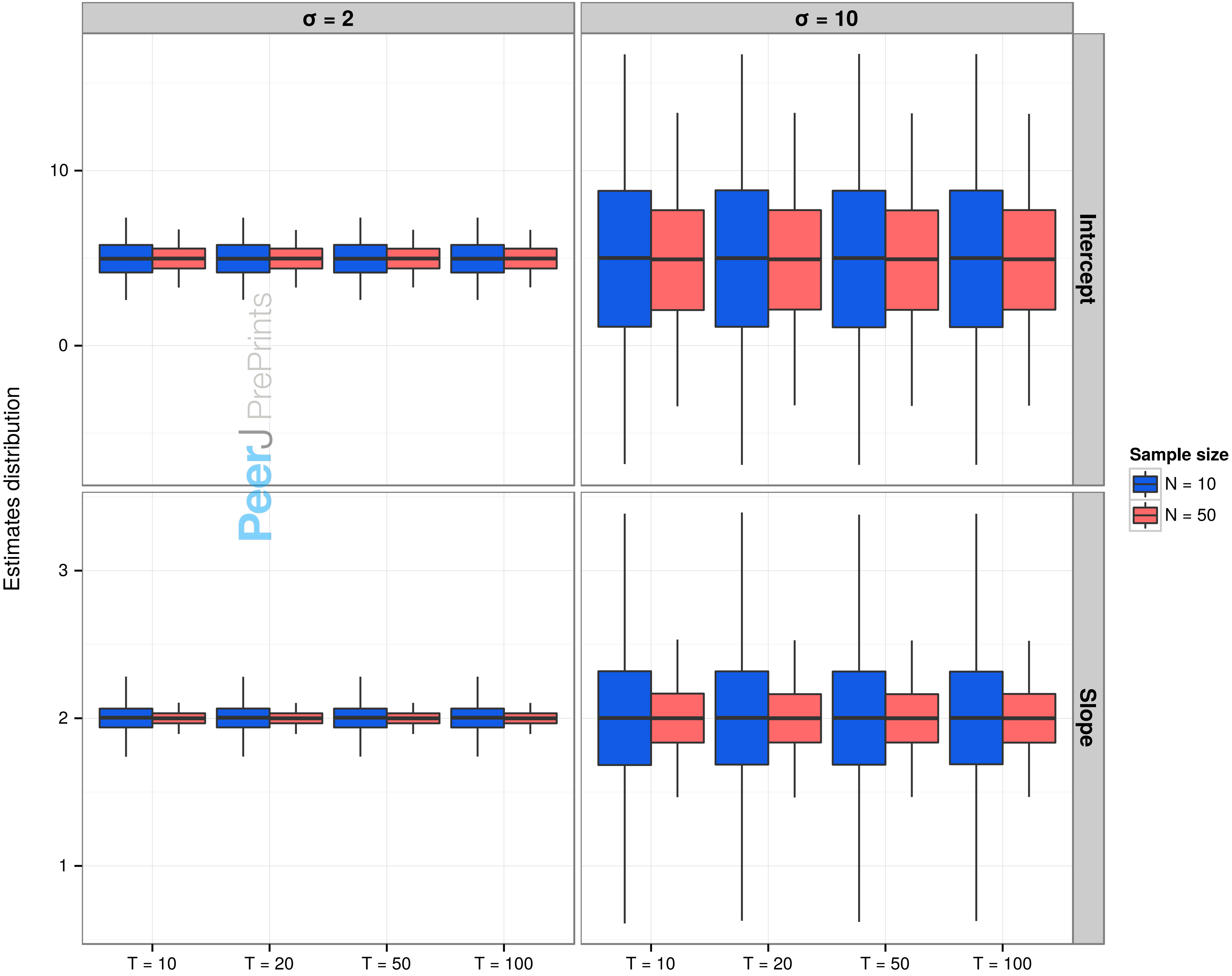
PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1216v1 | CC-BY 4.0 Open Access | rec: 6 Jul 2015, publ: 6 Jul 2015

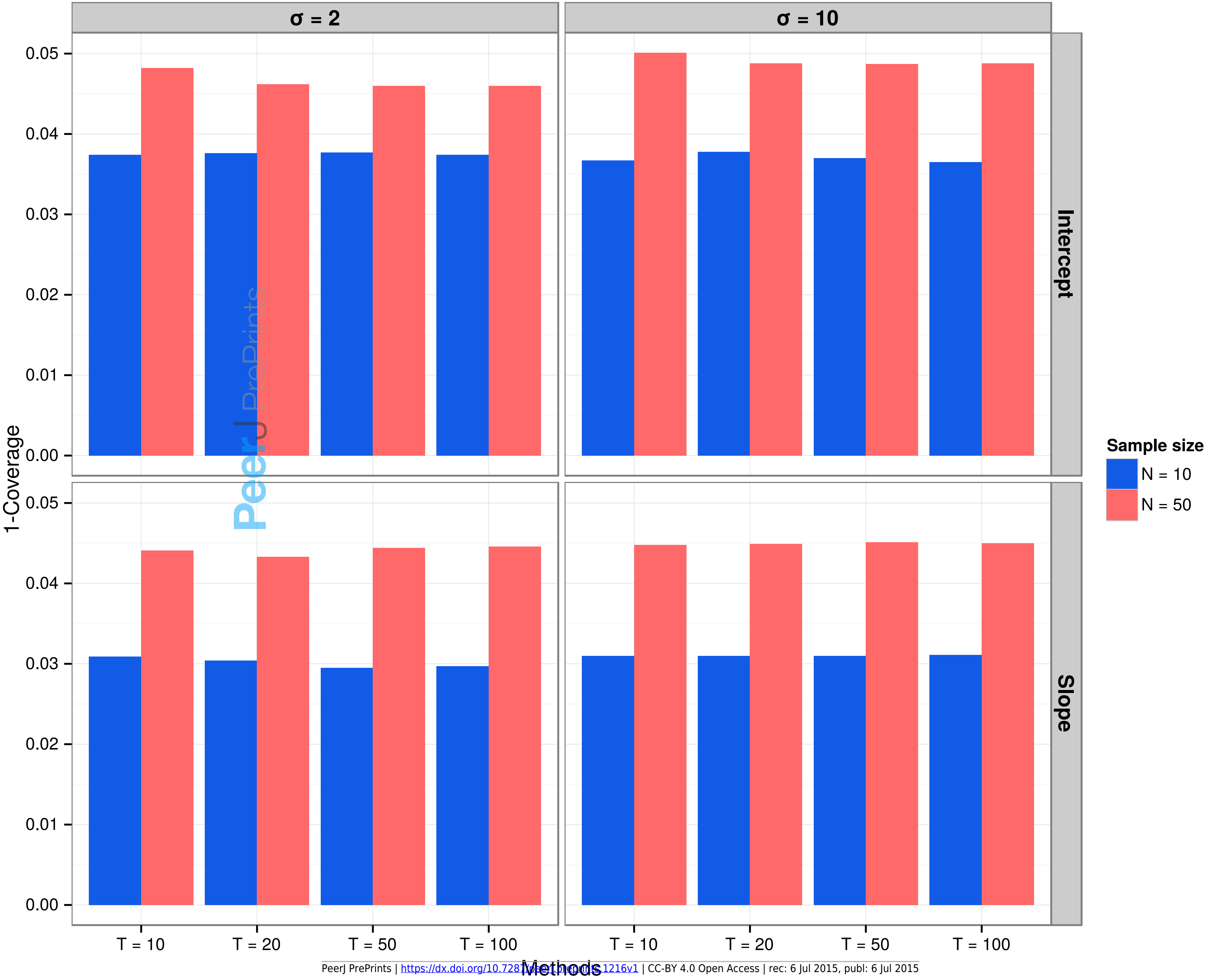

Darlingia darlingiana
Stephania japonica
Ripogonum album
Alpinia caerulea
Cordyline cannifolia
Freycinetia excelsa
Melodorum sp.
Desmos goezeanus
Haplostichanthus sp.
Myristica globosa
Litsea leefeana
Beilschmiedia bancroftii
Endiandra palmerstonii
Cryptocarya oblata
Daphnandra repandula
Citronella smythii
Polyscias elegans
Lantana camara
Faradaya splendida
Neosepicaea jucunda
Melodinus australis
Tabernaemontana pandacaqui
Randia tuberculosa
Embelia grayi
Symplocos paucistaminea
Tetrastigma nitens
Archidendron vaillantii
Xanthophyllum octandrum
Macaranga involucrata
Rockinghamia angustifolia
Dichapetalum papuanum
Pullea stutzeri
Sloanea langii
Sloanea macbrydei
Elaeocarpus grandis
Connarus conchocarpus
Sageretia hamosa
Rhodamnia sessiliflora
Syzygium cormiflorum
Argyrodendron trifoliolatum
Aglaia sapindina
Aglaia tomentosa
Dysoxylum papuanum
Dysoxylum parasiticum
Chisocheton longistipitatus
Melicope bonwickii
Flindersia brayleyana
Harpullia rhyticarpa
Rhysotoechia robertsonii
Tetracera nordtiana


σ = 2
σ = 10
Intercept
Slope
Estimates distribution
10
0
3
2
1
T = 10
T = 20
T = 50
T = 100
Sample size
N = 10
N = 50


σ = 2
σ = 10
Intercept
Slope
1-Coverage
0.05
0.04
0.03
0.02
0.01
0.00
T = 10
T = 20
T = 50
T = 100
Methods
Sample size
N = 10
N = 50


N = 10
N = 50
σ = 2
σ = 10
1.00
0.99
0.98
0.97
0.96
0.95
0.94
Text
Relative efficiency
T = 10
T = 20
T = 50
T = 100
Tree sample size


σ = 2
σ = 10
Intercept
Slope
Estimates distribution
10
0
3
2
1
T = 10
T = 20
T = 50
T = 100
T = 10
T = 20
T = 50
T = 100
Sample size
N = 10
N = 50


σ = 2
σ = 10
Intercept
Slope
1-Coverage
0.05
0.04
0.03
0.02
0.01
0.00
T = 10
T = 20
T = 50
T = 100
Methods
Sample size
N = 10
N = 50