Received: 27 March 2023 | Revised: 11 September 2023 | Accepted: 17 September 2023

DOI: 10.1111/jeb.14234

JOURNAL OF Evolutionary Biology

METHODS ARTICLE

phylosem: A fast and simple R package for phylogenetic inference and trait imputation using phylogenetic structural equation models

James T. Thorson[1] | Wouter van der Bijl[2]

[1]Resource Ecology and Fisheries Management, Alaska Fisheries Science Center, Seattle, Washington, USA

[2]Department of Zoology and Biodiversity Research Centre, University of British Columbia, Vancouver, British Columbia, Canada

Correspondence
James T. Thorson, Resource Ecology and Fisheries Management, Alaska Fisheries Science Center, Seattle, WA, USA.
Email: james.thorson@noaa.gov

Abstract

Phylogenetic comparative methods (PCMs) can be used to study evolutionary relationships and trade-offs among species traits. Analysts using PCM may want to (1) include latent variables, (2) estimate complex trait interdependencies, (3) predict missing trait values, (4) condition predicted traits upon phylogenetic correlations and (5) estimate relationships as slope parameters that can be compared with alternative regression methods. The Comprehensive R Archive Network (CRAN) includes well-documented software for phylogenetic linear models (*phylolm*), phylogenetic path analysis (*phylopath*), phylogenetic trait imputation (*Rphylopars*) and structural equation models (*sem*), but none of these can simultaneously accomplish all five analytical goals. We therefore introduce a new package *phylosem* for phylogenetic structural equation models (PSEM) and summarize features and interface. We also describe new analytical options, where users can specify any combination of Ornstein-Uhlenbeck, Pagel's-δ and Pagel's-λ transformations for species covariance. For the first time, we show that PSEM exactly reproduces estimates (and standard errors) for simplified cases that are feasible in *sem*, *phylopath*, *phylolm* and *Rphylopars* and demonstrate the approach by replicating a well-known case study involving trade-offs in plant energy budgets.

KEYWORDS
Ornstein-Uhlenbeck, phylogenetic comparative methods, phylogenetic trait imputation, structural equation models

1 | INTRODUCTION

Phylogenetic comparative methods (PCMs) are widely used to analyse data sets of multiple species while accounting for covariance that arises from evolutionary relatedness (Felsenstein, 1985; Harvey & Pagel, 1991; Pagel, 1999). The goal of these methods is usually to test evolutionary hypotheses of trait evolution, by analysing evolutionary associations between different traits or between a trait and a selective regime (e.g. environmental variables) (Garamszegi, 2014; Martins, 2000). In such data sets, variables may often have complex relationships with each other, and their influence on the trait of interest can be direct or mediated by other variables. As macro-evolutionary processes are typically not amenable to experimentation, it is necessary to disentangle these associations statistically (Pearl, 2000; Pearl et al., 2016; Shipley, 1997, 2002; von Hardenberg & Gonzalez-Voyer, 2013).

© 2023 European Society for Evolutionary Biology. This article has been contributed to by U.S. Government employees and their work is in the public domain in the USA.

While phylogenetic regression can test for direct associations, more complex relationships require the modelling of a casual graph using methods such as path analysis (Wright, 1934) or structural equation models (Kaplan, 2001).

PCMs involve a wide range of models and software including phylogenetic extensions of linear models (Tung Ho & Ané, 2014) and path analysis (van der Bijl, 2018; von Hardenberg & Gonzalez-Voyer, 2013). These methods require specifying evolutionary relatedness, and phylogenetic information is usefully represented in the R statistical environment using R-package *ape* (Paradis & Schliep, 2019). However, PCM software often requires complete data (i.e. all modelled variables are measured for each modelled taxon).

Ecologists often lack measurements for some combination of taxa and traits. In these cases, phylogenetic trait imputation (PTI) is widely used to infer missing values as well as their predictive variance (Hadfield & Nakagawa, 2010; Nakagawa & Freckleton, 2011). One widespread implementation is *Rphylopars* (Goolsby et al., 2017), which involves estimating a full-rank or diagonal covariance among traits. Alternatively, PTI can be implemented using phylogenetic factor analysis, which estimates a low-rank approximation to the full-rank covariance among traits (Hassler et al., 2022; Thorson et al., 2017; Tolkoff et al., 2018). Thorson et al. (2023) proposed a new phylogenetic structural equation model (PSEM) that merges structural equation models (SEM) and PCM. However, this approach has not been available with a simple, fast and easy-to-use package within the R statistical environment (R Core Team, 2021) and available via the Comprehensive R Archive Network (CRAN).

We here introduce and describe *phylosem*, the first general, simple and user-friendly package available on CRAN for PSEM (Thorson & van der Bijl, 2023). We specifically emphasize three novel contributions in this article:

1. We inventory the features available in alternative packages and discuss how PSEM can include these in a way that existing packages do not;
2. We demonstrate that PSEM using *phylosem* is equivalent to (and can replicate parameter estimates and standard errors from) existing packages in simplified instances;
3. We introduce new features available in *phylosem*, including functionality (A) to include any combination of several common transformations describing species covariance, (B) to estimate the intercepts in a phylogenetic path analysis and (C) estimating trait values for ancestral nodes in a phylogenetic tree.

2 | GENERALIZING EXISTING R PACKAGES

Analysts conducting PCM might want the following features:

1. estimating multiple dependencies among variables using a path diagram;

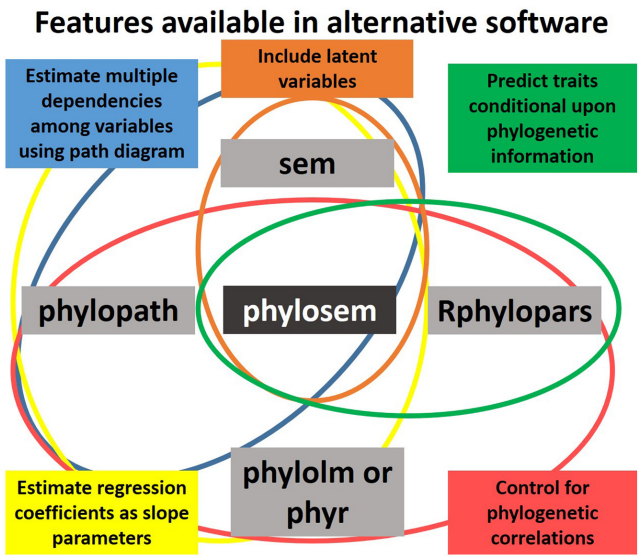

FIGURE 1 Visualizing the five desirable features for phylogenetic structural equation modelling (coloured ellipses and labelled using matching coloured boxes) and how five existing R-packages (grey boxes) include different combinations of these.

2. predict the value of unmeasured traits or latent traits that are impossible to directly measure;
3. predicting missing trait values conditional upon phylogenetic information;
4. estimating regression coefficients as slope parameters and
5. controlling for phylogenetic correlations.

These features are available in different combinations within the following packages:

- Phylogenetic linear models using *phylolm* (Tung Ho & Ané, 2014) and phylogenetic generalized linear models (PGLM) using *phyr* (Ives et al., 2020; Li et al., 2020) incorporate features 4 and 5;
- Phylogenetic path analysis using *phylopath* (van der Bijl, 2018; von Hardenberg & Gonzalez-Voyer, 2013) incorporates features 1, 4 and 5;
- Structural equation models using *sem* (Fox et al., 2020) incorporate features 1, 2 and 4 and
- Phylogenetic trait imputation using *Rphylopars* (Goolsby et al., 2017) incorporates features 3 and 5.

These desirable features and the capabilities of existing packages are summarized in Figure 1.

All these features are available in phylogenetic structural equation models (PSEM) using R-package *FishLife* (Thorson et al., 2023). However, *FishLife* includes additional features that are designed to fit a specific non-linear and data-integrated life-history model for fishes (Thorson, 2020) and therefore includes additional complexity in the user interface and logical code. For these reasons, *FishLife* is not available on CRAN and it does not include functionality that is common in widely used operational software.

Package *phylosem* generalizes features and estimates identical parameters, standard errors and predicted values to those packages. We specifically demonstrate that equivalence Appendices A–C:

- Appendix A shows the detailed comparison of syntax, runtime and parameter estimates with packages *phylolm*, *Rphylopars*, *phyr*, *sem* and applying a custom implementation of a phylogenetic generalized linear model using STAN (Stan Development Team, 2013) via package *brms* (Bürkner, 2017). It specifically confirms that runtimes are within an order of magnitude of those from these specialized packages with their varied estimation algorithms and that estimates are either identical or very similar;
- Appendix B shows a detailed comparison of *phylopath* and *phylosem* in terms of model selection and resulting parameter estimates.
- Appendix C shows that multiple transformations of the evolutionary covariance are identifiable. It also confirms that both *phylolm* and *phylosem* have an approximately linear increase in runtime with increasing tree size, as shown by testing across three orders of magnitude (i.e. trees with 10 to 10000 tips).

These appendices are also available as R-package vignettes, which allows these demonstrations to be updated if the syntax changes for these other packages.

3 | COVARIANCE AMONG TAXA IS NOT RESTRICTED TO BROWNIAN MOTION MODEL

As stated, package *FishLife* includes some of the same functionality as *phylosem* (with drawbacks noted previously). We here emphasize how *phylosem* extends the functionality in *FishLife*, including:

1. estimating the intercept equivalent to a linear model, so that the predicted value for an unmeasured trait can be written for a given taxa as a closed-form expression given the calculated intercept for that taxon and using path coefficients as slope parameters;
2. options for transforming the specified tree when estimating covariance among taxa and
3. specifying alternative distributions for each trait measurement.

We review these in the sequence below.

Package *phylosem* allows the user to estimate the equivalent of linear model 'intercepts' for a fitted model, and this option was not available in previous PSEM software (e.g. *FishLife*). PSEM involves estimating covariance among traits $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^t$, where the Cholesky of the covariance matrix is calculated as $\mathbf{L} = (\mathbf{I} - \mathbf{\Gamma})^{-1}\mathbf{S}$. This in turn involves estimating matrix $\mathbf{\Gamma}$, which includes estimated path coefficients, and matrix $\mathbf{S}$, which is the Cholesky of exogenous covariance (see Thorson et al., 2023 for details). Similarly, PSEM defines the correlation among taxa $\mathbf{R}$ from the user-specified tree and any estimated transformation parameters (α, λ and/or κ), as discussed in detail below. It then estimates a matrix of traits $\mathbf{B}$ composed of traits $\beta_{g,v}$ for each taxon g (including tips and ancestral nodes) and trait t, where $\mathbf{B}$ has a separable covariance constructed from the Kronecker product, $\mathrm{Var}(\mathbf{B}) = \mathbf{R} \otimes \mathbf{\Sigma}$. Given traits $\boldsymbol{\beta}_0$ for the root of the taxonomic tree, we calculate intercepts as $(\mathbf{I} - \mathbf{\Gamma})\boldsymbol{\beta}_0$, where these intercept estimates and resulting standard errors match those from *phylolm* when specifying a model that can be fitted by both packages.

Package *phylosem* defines the correlation among taxa $\mathbf{R}$ implicitly by specifying a series of conditional distributions for traits $\boldsymbol{\beta}_g$ for taxon g given traits $\boldsymbol{\beta}_{p_g}$ for ancestral node p_g separated by distance d_g in the user-specified tree. We define the following conditional distribution:

$$\boldsymbol{\beta}_g \sim \mathrm{MVN}\left(\boldsymbol{\mu} + \rho_g(\boldsymbol{\beta}_{p_g} - \boldsymbol{\mu}), \left(\lambda\tau_g + (1-\lambda)h_g\right)\mathbf{\Sigma}\right),$$

where $\boldsymbol{\mu}$ is the average trait value, λ is the additional variance for tips of the specified tree, h_g is the height of each tip and 0 for ancestral nodes and ρ_g is the correlation between two taxa. We first outline how to specify an Ornstein-Uhlenbeck model for correlation ρ_g:

$$\rho_g = e^{-\alpha d_g^{\kappa}},$$

where τ_g scales the conditional variance based on evolutionary distance d_g:

$$\tau_g = \frac{1}{2\alpha}\left(1 - e^{-2\alpha d_g^{\kappa}}\right),$$

and κ transforms the branch length. This model is completed by specifying a distribution for the root of the tree:

$$\boldsymbol{\beta}_0 \sim \mathrm{MVN}\left(\boldsymbol{\mu}, \frac{1}{2\alpha}\mathbf{\Sigma}\right),$$

where $\frac{1}{2\alpha}$ is the limit of τ_g as d_g approaches infinity under the Ornstein-Uhlenbeck model. This expression allows the user to estimate parameters that transform the correlation $\mathbf{R}$ among taxa in the specified tree. Potential parameters include

1. Ornstein-Uhlenbeck α: The user can estimate a first-order autoregressive process to represent covariance among taxa by estimating α, which has support $\alpha > 0$. We specifically use the Ornstein-Uhlenbeck parameterization (Tung Ho & Ané, 2014), so that the estimated parameter α is comparable to that from other software while specifying the stationary distribution for the probability distribution at the root. Alternatively, the user can simplify the model a priori by specifying a random walk process (i.e. eliminate α). This involves eliminating the conditional distribution for the root, specifying $\boldsymbol{\mu} = \mathbf{0}$, and replacing $\rho_g = 1$ and $\tau_g = d_g^{\kappa}$, which is (in a loose sense) the limit of ρ_g and τ_g as α approaches zero from above. The software specifically estimates $\log(\alpha)$ as parameter, and the user can back-calculate α which is then restricted to its natural support;

2. Pagel's-λ: Similarly, the user can specify a component of covariance that is independent for each tip of the specified tree (Pagel, 1999), where λ is the proportion explained by phylogeny and $1-\lambda$ is the proportion that is independent for each tip. Alternatively, the user can simplify the model by fixing $\lambda = 1$ a priori such that the tips have no additional variance relative to ancestral nodes. Parameter λ has support $0 < \lambda < 1$ and the software specifically estimates $\text{logit}(\lambda)$. The user can again back-calculate λ which is restricted to its natural support;
3. Pagel's-κ: Finally, the user can estimate a non-linear transformation κ of the edge length between any two nodes of the specified tree. Alternatively, the user can simplify the model by specifying $\kappa = 1$ a priori to eliminate this transformation. Parameter κ has support $\kappa > 0$ and the software specifically estimates $\log(\kappa)$ such that κ is restricted to its natural support.

In each case, *phylosem* calculates standard errors for $\log(\alpha)$, $\text{logit}(\lambda)$ and/or $\log(\kappa)$ via the inverse-Hessian matrix of the marginal log-likelihood. However, the software also calculates standard errors for derived values (α, λ and/or κ) via a generalization of the delta method (Kass & Steffey, 1989). For some data sets, the estimated parameter may approach positive or negative infinity, corresponding to the derived parameter approaching a natural bound in its support (e.g. $\text{logit}(\lambda) \gg 5$ corresponding to $\lambda \to 1$). In these cases, the Hessian of the marginal log-likelihood will not be invertible. In these cases, the user should then reduce the model as appropriate (e.g. when $\text{logit}(\lambda) \gg 5$, by turning off the Pagel's λ transformation such that $\lambda = 1$).

Package *phylosem* allows the user to estimate these transformations individually or in any combination and estimates the values of α, λ and/or κ at the same time as other model parameters. These transformations define the correlation among taxa $\mathbf{R}$, and this same correlation operates for each modelled variable (i.e. each column of trait-matrix $\mathbf{B}$). Future research could extend *phylosem* to allow separate transformations (i.e. separate estimates of α, λ and/or κ) for each trait. However, this would result in a non-separable covariance and therefore requires developing statistical methods beyond those of Thorson et al. (2023). We emphasize that *phylosem* estimates traits $\boldsymbol{\beta}_g$ for each taxon g, where these taxa are represented using a user-specified tree. We envision that tips will often represent species but users could instead specify a tree where tips are populations or even individuals, for example, estimating a Pagel's λ to estimate the proportion of variance explained by individuals within a population.

Estimating these three parameters in any combination allows users to explore a $2 \times 2 \times 2$ factorial cross of either estimating α, λ and κ, or fixing $\alpha \to 0$, $\lambda = 1$ and $\kappa = 1$ a priori. This results in eight possible forms for covariance among taxa for a given user-specified tree, and the user can use the Akaike Information Criterion (Akaike, 1974) to select among these. Other software with similar features typically allows the user to specify only one of these transformations at a time, for example, as done in *phylolm::phylolm*, *Rphylopars::phylopars* or *geiger::fitContinuous* (Pennell et al., 2014). As one exception, *caper::pgls* (Orme et al., 2018) does allow simultaneously estimating λ and δ but does not allow estimating these simultaneously with Ornstein-Uhlenbeck α as *phylosem* allows. These options are illustrated further in Appendix A, where we include examples illustrating that estimated values for α, λ and κ (approximately) match those arising from comparable software.

Given that previous software has not allowed these transformations to be specified simultaneously, readers may wonder whether the corresponding parameters are jointly identifiable. To address this question, we first note that if a model is not identifiable, then it will not be estimable for any data set (Jacquez & Greif, 1985). Therefore (by the law of contraposition), if a model is estimable for any single data set, then it must be identifiable. We therefore use the 'rhino' data set from phylopath to show that all three transformations are estimable for these data (see Appendix C). We therefore conclude that these three transformations are jointly identifiable, although they may not be estimable for other data sets. By default, 'phylosem' calculates the Hessian of all fixed and random effects and will throw an error message when this Hessian is not positive definite. This has been recommended as a default check for estimability in other models fitted using automatic differentiation (Hunter & Caswell, 2009).

As the simultaneous estimation of these transformations has not been explored so far, the utility and interpretation of this approach is not well understood. Nevertheless, we suspect that estimating multiple transformations will be helpful for some real-world analyses and therefore offer it as an option. For example, the Pagel's-λ transformation allows analysts to determine whether a specified tree is strongly predictive of ($\lambda \to 1$) or unrelated to ($\lambda \to 0$) trait covariance and might be useful as a statistical measurement of phylogenetic associations (e.g. Freckleton et al., 2002). By contrast, the Ornstein-Uhlenbeck α measures whether evolutionary drift is counterbalanced by stabilizing selection towards some fitness optimum (Lande, 1976) and might be useful as an ecological measurement of stabilizing selection. We envision that researchers might want to test both statistical and ecological hypotheses for a given system, and this can be accomplished by estimating both λ and/or α and evaluating model fit for the resulting 2×2 cross of models. However, we also cannot find any simulation or case-study experiments testing whether estimating multiple transformations is likely to improve or degrade estimation performance in real-world contexts. We therefore encourage analysts to evaluate performance on a case-by-case basis using a self- and cross-test simulation experiment, that is, fit real-world data using alternative combinations of transformations, simulate new data conditional on estimated parameters, then refit each simulated data set using the same combination of transformations and finally, measure average performance for each combination of simulated and estimated transformations. Future research could perhaps use a minimax framework to identify whether some combination of transformations is robust to model misspecification and therefore likely to perform well on average (e.g. similar to Thorson et al., 2021). We hope that simplified software such as *phylosem* will enable such an experiment.

Finally, package *phylosem* fits latent trait $\beta_{g,v}$ to any available measurement, while also allowing missing values that are inputted as

NAs. The user can specify different distributions for available data, similar to previous generalized linear models used for PCM (Hadfield & Nakagawa, 2010). Options include

1. *No measurement error*: The user can specify that any specified trait is measured without error, where this corresponds to default behaviour in *phylolm*;
2. *Normal measurement error*: The user can specify that a given trait measurement follows a normal distribution with mean $\beta_{g,v}$ and an estimated variance for measurement errors;
3. *Bernoulli errors*: The user can specify a binary (0 or 1) measurement for a given trait and specify that the measurement follows a Bernoulli distribution with probability $\pi_{g,v} = \text{logit}^{-1}(\beta_{g,v})$;
4. *Poisson errors*: Finally, the user can specify a non-negative integer measurement and specify that the measurement follows a Poisson distribution with intensity $\lambda_{g,v} = e^{\beta_{g,v}}$.

The user can specify a separate distribution for each individual trait measurement. In Appendix A, we compare results when specifying Poisson errors using simulated data between *phylosem*, *phyr* and a custom implementation using *brms*. We also compare results when specifying Bernoulli errors using simulated data between *phylosem* and *phyr*. Estimates and standard errors are similar (but not quite identical) in these tested cases. We therefore recommend that future research provide a systematic cross-test for phylogenetic generalized linear models (PGLMs) using different software implementations. Future releases of *phylosem* could presumably add additional link functions and distributions (e.g. allowing users to specify a known variance for some observations).

Package *phylosem* is capable of fitting non-Gaussian distributions and missing values because it approximates the high-dimensional integral across random states $\beta_{g,v}$ for all taxa and traits while calculating the marginal likelihood of parameters, where this marginal likelihood is then optimized within an R session. Package *phylosem* specifically uses the Laplace approximation while automatically detecting sparsity, available using Template Model Builder (Kristensen, 2014). It then applies a generalization of the delta method (Tierney et al., 1989) to calculate the standard errors for predicted states $\beta_{g,v}$. The computational methods are explained in detail in Thorson et al. (2023), and *phylosem* provides a simplified and generalized interface for this functionality.

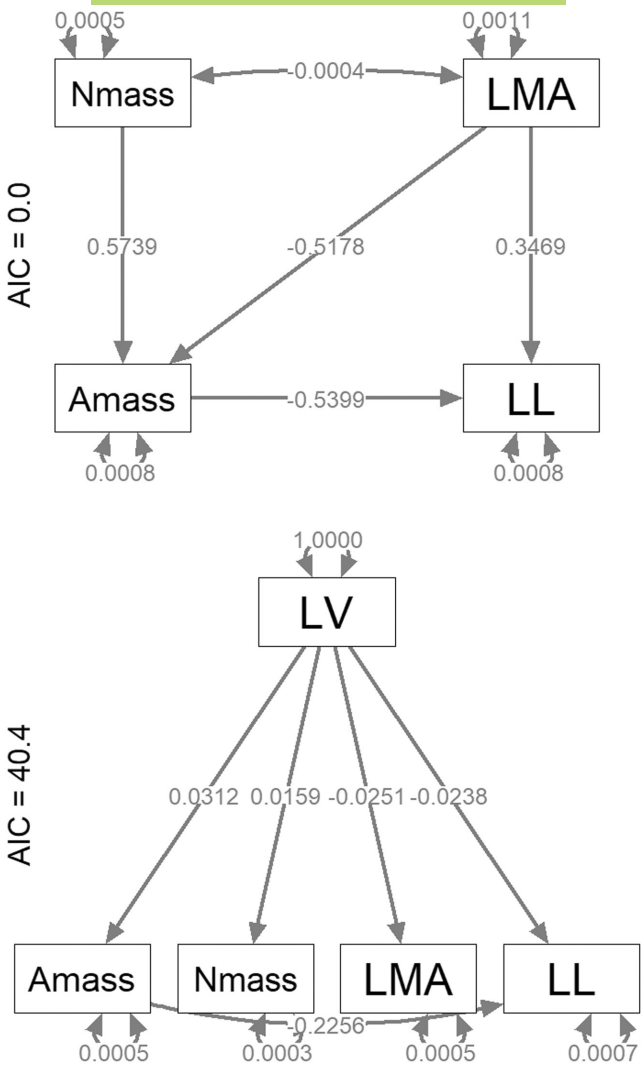

FIGURE 2 Comparison of path diagrams estimated using phylogenetic structural equation models (PSEM) and visualized using R-package semPlot (Epskamp, 2022) using two relationships hypothesized by Shipley et al. (2006) among four measured leaf traits (Amass, photosynthetic rate; LL, leaf lifespan; LMA, leaf mass per area; Nmass, leaf nitrogen content) and one latent variable (LV) and subsequently analysed by Mason et al. (2016). We show results using the full 'GLOPNET' database from Mason et al. (2016). Path diagrams show variance/covariance parameters (double-headed arrows) as well as path coefficients (single-headed arrows) and listing the ΔAIC on the left-hand side of each panel.

4 | CASE STUDY DEMONSTRATION: MODEL SELECTION FOR LEAF ENERGY BUDGETS

To demonstrate the importance of combining these features, we reanalyse a data set that has been used to test the leaf economics spectrum hypothesis. Shipley et al. (2006) compared two models, estimating: (#1) the impact of leaf photosynthetic rate on leaf lifespan or (#2) a latent variable explaining the associations between leaf photosynthetic rate, leaf lifespan and other traits. Their analysis applied structural equation modelling (SEM) without phylogenetic correction and supported model #2. Subsequently, Mason et al. (2016) used Rphylopars to impute missing trait values and then applied SEM using these two models to the estimated trait covariance. This re-analysis supported model #1. However, their two-stage analytical approach did not report the parsimony of these alternative models using the Akaike Information Criterion (AIC), perhaps because the two-stage approach does not propagate the precision of imputed values when computing an AIC value. By contrast, joint analysis

TABLE 1 List of user functions available in R-package *phylosem*, listing their primary arguments and intended usage.

Function	Arguments	Usage
Core model functionality		
phylosem(sem, tree, data, ...)	• sem: string describing SEM • tree: *phylo* object describing tree • data: a data frame of trait measurements, where missing values are NAs • ...: Additional options	Fitting PSEM to data
print(psem)	• psem: output from phylosem(.)	Print output
summary(psem)	• psem: output from phylosem(.)	Summarize slope and intercept estimates and standard errors
coef(psem)	• psem: output from phylosem(.)	Extract coefficients
AIC(psem)	• psem: output from phylosem(.)	Extract Akaike Information Criterion
Model comparison		
compare_phylosem(sem_set, ...)	• sem: list of strings describing SEMs • ...: Additional options matching those in phylosem(.)	Fit multiple PSEMs to data
best(psem_list)	• psem_list: output from compare_phylosem (.)	Extract best PSEM
choice(psem_list)	• psem_list: output from compare_phylosem (.)	Extract selected PSEM
average(psem_list)	• psem_list: output from compare_phylosem (.)	Average coefficients from multiple PSEMs
Coerce methods for summary and plotting		
as_fitted_DAG(psem)	• psem: output from phylosem(.)	Convert to fitted output from *phylopath* (van der Bijl, 2018)
as_sem(psem)	• psem: output from phylosem(.)	Convert to fitted output from *sem* (Fox et al., 2020)
as_phylo4d(psem)	• psem: output from phylosem(.)	Convert to an object that includes phylogenetic tree and estimated trait values (Bolker et al., 2015)

using *phylosem* involves fitting both models while simultaneously imputing missing trait values, which allows us to calculate a single likelihood (and resulting AIC value) for each model.

We jointly impute missing values and fit a structural equation model using *phylosem* while specifying a Brownian motion model and assuming that measurements are without error. Joint analysis using *phylosem* confirms that model #1 is much more parsimonious (ΔAIC = 40.4, see Figure 2), in support of Mason et al. (2016). We note, however, that marginal AIC (as used here) will sometimes select a model with greater complexity than the alternative conditional AIC (Greven & Kneib, 2010), and that this behaviour has been observed in previous PCM studies (Ho & Ané, 2014). We therefore recommend experimental evidence in support of hypothesized evolutionary trade-offs, for example, from long-term evolutionary experiments.

5 | DISCUSSION

Package *phylosem* is designed to be the first user-friendly implementation of PSEM. It accomplishes this by including a minimum of new logical code, such that developers and users can readily inspect all user-interface and logical code. It is designed with options to coerce the fitted output to object classes used by package *sem*, *phylopath* and *phylobase* (Bolker et al., 2015). These common packages can then be used to plot estimates of path coefficients, estimated traits, the estimated structural equation model and to compute direct and indirect effects. It uses the same interface as *phylopath* to allow users to automatically compare parsimony across a specified set of model structures (Table 1).

We believe that PTI will be improved by using SEM for several reasons:

1. improved parsimony (and decreased predictive error) when replacing an unstructured covariance among traits with ecologically meaningful trait associations and
2. increased scope of model use, for example, when fitting a model to a large number of traits such that an unstructured covariance among traits would become infeasible to estimate.

Similarly, we believe that PCMs will be improved by simultaneously implementing PTI because data from taxa with some but not all trait data can still be fitted, and the sampling fraction of the tips can be increased. Finally, PSEM improves upon existing phylogenetic path analysis by allowing inclusion of latent variables and exogenous covariance.

AUTHOR CONTRIBUTIONS

James T. Thorson: Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); project administration (equal); resources (equal); software (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Wouter van der Bijl:** Data curation (equal); formal analysis (equal); investigation (equal);

methodology (equal); software (equal); validation (equal); writing – review and editing (equal).

ACKNOWLEDGEMENTS

We thank S. Friedman and T. Buser for comments on an earlier draft, and K. Kristensen for developing Template Model Builder and making this package possible. We also thank E. Goolsby and C. Mason for discussion of the case study comparison in Figure 2, and for providing the file 'time_tree.tre' used in that comparison. Finally, we thank two anonymous reviewers for helpful comments.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/jeb.14234.

DATA AVAILABILITY STATEMENT

We present a re-analysis of a case study (Mason et al., 2016; Shipley et al., 2006) involving data-measuring leaf traits (Mason et al., 2016) and a tree phylogeny (Smith et al., 2011). The leaf traits are publicly available (ele12542-sup-0003-datas1.xls in Supporting Information at https://onlinelibrary.wiley.com/doi/10.1111/ele.12542) and we obtained the tree phylogeny 'time_tree.tre' by contacting those authors. Analysis was done R-package *phylosem*, which is currently on CRAN. We specifically publish results obtained using release 1.0.2, which is available on Zenodo via https://zenodo.org/record/8371246.

ORCID

James T. Thorson https://orcid.org/0000-0001-7415-1010

REFERENCES

Akaike, H. (1974). New look at statistical-model identification. *IEEE Transactions on Automatic Control, AC*, *19*(6), 716–723.

Bolker, B., Butler, M., Cowan, P., de Vienne, D., Eddelbuettel, D., Holder, M., Jombart, T., Kembel, S., Michonneau, F., & Orme, B. (2015). *phylobase: Base package for phylogenetic structures and comparative data*. R Package Version 0.8.0.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Epskamp, S. (2022). *semPlot: Path diagrams and visual analysis of various SEM packages (1.1.6)* [Computer software]. https://CRAN.R-project.org/package=semPlot

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, *125*(1), 1–15. https://doi.org/10.1086/284325

Fox, J., Nie, Z., & Byrnes, J. (2020). *Sem: Structural equation models. R package version 3.1-11* [Computer software]. https://CRAN.R-project.org/package=sem

Freckleton, R. P., Harvey, P. H., & Pagel, M. (2002). Phylogenetic analysis and comparative data: A test and review of evidence. *The American Naturalist*, *160*(6), 712–726. https://doi.org/10.1086/343873

Garamszegi, L. Z. (Ed.). (2014). *Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice* (2014th ed.). Springer.

Goolsby, E. W., Bruggeman, J., & Ané, C. (2017). Rphylopars: Fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution*, *8*(1), 22–27. https://doi.org/10.1111/2041-210X.12612

Greven, S., & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, *97*(4), 773–789. https://doi.org/10.1093/biomet/asq042

Hadfield, J. D., & Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, *23*(3), 494–508. https://doi.org/10.1111/j.1420-9101.2009.01915.x

Harvey, P. H., & Pagel, M. D. (1991). *The comparative method in evolutionary biology*. Oxford University Press.

Hassler, G. W., Gallone, B., Aristide, L., Allen, W. L., Tolkoff, M. R., Holbrook, A. J., Baele, G., Lemey, P., & Suchard, M. A. (2022). Principled, practical, flexible, fast: A new approach to phylogenetic factor analysis. *Methods in Ecology and Evolution*, *13*(10), 2181–2197. https://doi.org/10.1111/2041-210X.13920

Ho, L. S. T., & Ané, C. (2014). Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution*, *5*(11), 1133–1146. https://doi.org/10.1111/2041-210X.12285

Hunter, C. M., & Caswell, H. (2009). Rank and redundancy of multistate mark-recapture models for seabird populations with unobservable states. In *Modeling demographic processes in marked populations* (pp. 797–825). Springer. https://doi.org/10.1007/978-0-387-78151-8_37

Ives, A., Dinnage, R., Nell, L. A., Helmus, M., & Li, D. (2020). *phyr: Model based phylogenetic analysis*. R package version 1.1.0.

Jacquez, J. A., & Greif, P. (1985). Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Mathematical Biosciences*, *77*(1–2), 201–227.

Kaplan, D. (2001). Structural equation modeling. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 15215–15222). Pergamon. https://doi.org/10.1016/B0-08-043076-7/00776-2

Kass, R. E., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). *Journal of the American Statistical Association*, *84*(407), 717–726. https://doi.org/10.2307/2289653

Kristensen, K. (2014). *TMB: General random effect model builder tool inspired by ADMB*. https://github.com/kaskr/adcomp

Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. *Evolution*, *30*, 314–334.

Li, D., Dinnage, R., Nell, L. A., Helmus, M. R., & Ives, A. R. (2020). phyr: An r package for phylogenetic species-distribution modelling in ecological communities. *Methods in Ecology and Evolution*, *11*(11), 1455–1463. https://doi.org/10.1111/2041-210X.13471

Martins, E. P. (2000). Adaptation and the comparative method. *Trends in Ecology & Evolution*, *15*(7), 296–299. https://doi.org/10.1016/S0169-5347(00)01880-2

Mason, C. M., Goolsby, E. W., Humphreys, D. P., & Donovan, L. A. (2016). Phylogenetic structural equation modelling reveals no need for an 'origin' of the leaf economics spectrum. *Ecology Letters*, *19*(1), 54–61. https://doi.org/10.1111/ele.12542

Nakagawa, S., & Freckleton, R. P. (2011). Model averaging, missing data and multiple imputation: A case study for behavioural ecology. *Behavioral Ecology and Sociobiology*, *65*(1), 103–116. https://doi.org/10.1007/s00265-010-1044-7

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., & Pearse, W. (2018). *Caper: Comparative analyses of phylogenetics and evolution in R* [Computer software]. https://CRAN.R-project.org/package=caper

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*(6756), 877–884. https://doi.org/10.1038/44766

Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pearl, J. (2000). *Causality: Models, reasoning, and inference* (p. xvi, 384). Cambridge University Press.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., FitzJohn, R. G., Alfaro, M. E., & Harmon, L. J. (2014). Geiger v2. 0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, *30*(15), 2216–2218.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Shipley, B. (1997). Exploratory path analysis with applications in ecology and evolution. *The American Naturalist*, *149*(6), 1113–1138. https://doi.org/10.1086/286041

Shipley, B. (2002). *Cause and correlation in biology: A user's guide to path analysis, structural equations and causal inference*. Cambridge University Press.

Shipley, B., Lechowicz, M. J., Wright, I., & Reich, P. B. (2006). Fundamental trade-offs generating the worldwide leaf economics spectrum. *Ecology*, *87*(3), 535–541. https://doi.org/10.1890/05-1051

Smith, S. A., Beaulieu, J. M., Stamatakis, A., & Donoghue, M. J. (2011). Understanding angiosperm diversification using small and large phylogenetic trees. *American Journal of Botany*, *98*(3), 404–414. https://doi.org/10.3732/ajb.1000481

Stan Development Team. (2013). *Stan: A C++ library for probability and sampling, Version 1.3* [Computer software]. http://mc-stan.org/

Thorson, J. T. (2020). Predicting recruitment density dependence and intrinsic growth rate for all fishes worldwide using a data-integrated life-history model. *Fish and Fisheries*, *21*(2), 237–251. https://doi.org/10.1111/faf.12427

Thorson, J. T., Cunningham, C. J., Jorgensen, E., Havron, A., Hulson, P.-J. F., Monnahan, C. C., & von Szalay, P. (2021). The surprising sensitivity of index scale to delta-model assumptions: Recommendations for model-based index standardization. *Fisheries Research*, *233*, 105745. https://doi.org/10.1016/j.fishres.2020.105745

Thorson, J. T., Maureaud, A. A., Frelat, R., Mérigot, B., Bigman, J. S., Friedman, S. T., Palomares, M. L. D., Pinsky, M. L., Price, S. A., & Wainwright, P. (2023). Identifying direct and indirect associations among traits by merging phylogenetic comparative methods and structural equation models. *Methods in Ecology and Evolution*, *14*(5), 1259–1275. https://doi.org/10.1111/2041-210X.14076

Thorson, J. T., Munch, S. B., Cope, J. M., & Gao, J. (2017). Predicting life history parameters for all fishes worldwide. *Ecological Applications*, *27*(8), 2262–2276. https://doi.org/10.1002/eap.1606

Thorson, J. T., & van der Bijl, W. (2023). phylosem: Phylogenetic structural equation model. [Computer software] https://CRAN.R-project.org/package=phylosem

Tierney, L., Kass, R. E., & Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, *84*(407), 710–716.

Tolkoff, M. R., Alfaro, M. E., Baele, G., Lemey, P., & Suchard, M. A. (2018). Phylogenetic factor analysis. *Systematic Biology*, *67*(3), 384–399. https://doi.org/10.1093/sysbio/syx066

Tung Ho, L. S., & Ané, C. (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology*, *63*(3), 397–408.

van der Bijl, W. (2018). phylopath: Easy phylogenetic path analysis in R. *PeerJ*, *6*, e4718. https://doi.org/10.7717/peerj.4718

von Hardenberg, A., & Gonzalez-Voyer, A. (2013). Disentangling evolutionary cause-effect relationships with phylogenetic confirmatory path analysis. *Evolution*, *67*(2), 378–387. https://doi.org/10.1111/j.1558-5646.2012.01790.x

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, *5*(3), 161–215.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Thorson, J. T., & van der Bijl, W. (2023). phylosem: A fast and simple R package for phylogenetic inference and trait imputation using phylogenetic structural equation models. *Journal of Evolutionary Biology*, *36*, 1357–1364. https://doi.org/10.1111/jeb.14234