Ecological Archives **M081-018-A2**

Ives, A. R., and M. R. Helmus. 2011. Generalized linear mixed models for phylogenetic analyses of community structure. ***Ecological Monographs*** **81:511-525.**

Appendix B. Statistical estimation and validation of PGLMMs.

Description of the estimation method

To fit the PGLMM models to data, we used a combination of penalized quasi-likelihood (PQL) (Schall 1991, Breslow and Clayton 1993) and restricted maximum likelihood (REML) (Harville 1974, Zeger and Liang 1985, Liang and Zeger 1986) for estimating the fixed and random components of the models, respectively. Let

$$\mathbf{C} = \sum_{l=1}^{p} \sigma_l^2 \Sigma_l \qquad \text{(B.1)}$$

denote the working covariance matrix consisting of the sum of covariance matrices for p random effects (e.g., kron($\mathbf{I}_m$, $\sigma^2_{\text{spp}}\boldsymbol{\Sigma}_{\text{spp}}$) and kron($\sigma^2_{\text{site}}\mathbf{I}_m$, $\mathbf{I}_n$) for b_i and $c_{\text{site}[i]}$, respectively in Eq. 1), and let $\mathbf{V} = \mathbf{C} + \mathbf{W}^{-1}$ where $\mathbf{W}$ is the $nm \times nm$ diagonal matrix with diagonal elements $\boldsymbol{\mu}^{\bullet}(1 - \boldsymbol{\mu})$. Here,

$$\boldsymbol{\mu} = \frac{\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{b})}{1 + \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{b})} \qquad \text{(B.2)}$$

where $\mathbf{X}$ is the $nm \times q$ matrix containing independent variables (including categorical variables) corresponding to the fixed effects whose working coefficients are contained within the $q \times 1$ vector $\boldsymbol{\beta}$, and $\mathbf{B}$ is the $nm \times 1$ vector containing working estimates of the sum of coefficients of the random effects for each data point. Letting $\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B} + (\mathbf{Y} - \boldsymbol{\mu})^{\bullet}(\boldsymbol{\mu}^{\bullet}(1 - \boldsymbol{\mu}))^{-1}$, the updated values of the working estimates of $\boldsymbol{\beta}$ are

$$\boldsymbol{\beta} = \frac{\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}}{\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}} \qquad \text{(B.3)}$$

and the updated working estimates of **B** are

$$\mathbf{B} = \mathbf{C}\mathbf{V}^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}) \tag{B.4}$$

both of which are conditional on the working estimate of **C** that contains the estimates of the variances $\sigma^2{}_l$ of the random effects component of the model.

Conditional on the working estimates of **β** and **B**, we estimated the variance components of the random effects $\sigma^2{}_l$ by minimizing the negative log restricted likelihood function

$$L = \mathrm{sum}\left(\log\left(\mathrm{diag}\left(\mathrm{chol}(\mathbf{V})\right)\right)\right) + \frac{1}{2}\mathbf{H}'\mathbf{V}^{-1}\mathbf{H} + \frac{1}{2}\log\left(\det\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)\right) + \frac{\text{constant}}{2} \tag{B.5}$$

where $\mathbf{V} = \mathbf{C} + \mathbf{W}^{-1}$ (see Eq. B.1), $\mathrm{diag}\left(\mathrm{chol}(\mathbf{V})\right)$ is the vector of diagonal elements of the Cholesky decomposition of **V**, $\mathbf{H} = \mathbf{Z} - \mathbf{X}\boldsymbol{\beta}$, det() denotes the determinant, $\text{constant} = (nm - p)\log \boldsymbol{\pi} - \log\left(\det\left(\mathbf{X}'\mathbf{V}\mathbf{X}\right)\right)$, and nm is the number of data points. We minimized the log restricted likelihood function numerically using a simplex search routine (fminsearch.m in matlab); although slower than other methods, it proved to be very stable and always led to convergence.

For selection among models differing in both fixed and random effects using, for example, Akaike’s Information Criterion (Burnham and Anderson 2002), it is necessary to use maximum likelihood (ML) estimation rather than REML. This can be done by replacing the negative restricted likelihood function in equation B5 with the negative likelihood function

$$L = \mathrm{sum}\left(\log\left(\mathrm{diag}\left(\mathrm{chol}(\mathbf{V})\right)\right)\right) + \frac{1}{2}\mathbf{H}'\mathbf{V}^{-1}\mathbf{H} + \frac{nm}{2}\log \boldsymbol{\pi}. \tag{B.6}$$

For overall estimation of model parameters, PQL estimation of coefficients **β** and **B** (Eqs. B.1–B.4) is alternated with REML (or ML) estimation of the random effects variances $\sigma^2{}_l$ (Eq. B.5 or B.6) until convergence is achieved. Profile likelihood estimates of the confidence intervals of $\sigma^2{}_l$ are computed directly from Eqs. B.5 or B.6.

Investigation of the statistical properties of the PGLMMs

To investigate the statistical properties of our estimation of PGLMMs, we performed simulations for models I–IV. In contrast to the simulations used in the main text, the simulations we perform here use the statistical model that we then fit to the simulated data. This procedure is analogous to performing parametric bootstrapping to obtain inference about the estimators of coefficients in a model, but here we assign values to coefficients rather than first estimating them from a data set. We perform these simulations under two scenarios: when the true values of the variance parameters of interest, namely σ_{spp}, σ_{phylo}, σ_{repulse}, and σ_{trait}, were greater than zero, and when they were zero (Table B1); the second scenario corresponds to the case in which there is no phylogenetic signal or trait-based pattern of community structure. In both scenarios, the true value of σ_{site} was greater than zero.

Information about the bias and precision of the estimators of σ is given by the mean and range, respectively, of estimated values for the simulated data sets. In the scenario with phylogenetic signal, estimates of σ tended to be downwards biased, and this bias was greatest for the target variance parameters (Table B1, Fig. B1). There also tended to be negative correlations between these estimates and other variance components of the model.

Information about the accuracy and power of hypothesis testing and confidence intervals computed using profile likelihoods is given by counting the proportion of simulations in which the null hypothesis was rejected and the proportion of simulations for which the 95% confidence intervals excluded the true values of σ used to simulate the data. By definition of 95% confidence intervals, 2.5% of the data sets should have had the lower confidence bound above the true value of σ, and 2.5% should have had the upper confidence bound below the true value of σ. In the scenarios without phylogenetic signal for which the true values of $\sigma = 0$ (σ_{spp}, σ_{phlyo}, σ_{repulse} and σ_{trait} under the "No phylogenetic signal" heading), the expected number of rejections (false positive, or type I errors) of the null hypothesis H0:$\sigma = 0$ at an α-level of 0.025 should be 2.5% of the data sets. For some models the rejection rates were lower than this, suggesting that the hypothesis test of H0:$\sigma = 0$ will generate false negatives (type II errors).

Although downwards bias and overly wide confidence intervals will in general decrease the rejection rate of the null hypothesis of no phylogenetic signal, the negative correlation between estimates of the variance parameters may lead to inflated rejection rates. This negative correlation implies that the estimation procedure has a difficult time attributing variance between the two possible sources. If there is site-to-site variance in species occurrences ($\sigma_{\text{site}} > 0$) and correlation in estimates, then part of the site-to-site variance may be attributed to σ_{spp} or σ_{repulse} in models I and III. Similarly, in model II the covariance between σ_{phlyo} and σ_{slope} indicates difficulty in assigning phylogenetic signal to variation in slopes among species. Given this uncertainty, we suggest performing bootstrapping when applying PGLMMs to real data sets.

Comparison with lmer/Laplace

We tested our approach against the R package lmer (Bates et al. 2008) that uses the Laplace approximation to obtain the maximum likelihood estimates. We did this because the literature shows that standard implementation of PQL estimation for binary processes does not perform well (Breslow and Clayton 1993, Austin 2010). For these tests, we used a statistical model for binary processes that has simple blocked random effects (as opposed phylogenetic covariance matrices which lmer cannot fit):

$$\begin{aligned}
&\Pr(Y_i=1) = \mu_i \\
&\mu_i = \text{logit}^{-1}(\beta_0 + b_1 + b_2) && \text{(B.7)} \\
&b_1 \sim \text{N}(0,\sigma_1^2) \\
&b_2 \sim \text{N}(0,\sigma_2^2)
\end{aligned}$$

with $n = 60$, 10 categories in group 1 (i.e., b_1 takes 10 values), 20 categories in group 2, group 2 completely nested within group 1, $\beta_0 = 0$, $\sigma_1 = 0.71$, and $\sigma_2 = 0.32$. From 1000 bootstrap simulations, the mean estimates from lmer were 0.50 ± 0.50 and 0.28 ± 0.44 (±SD) for σ_1 and σ_2 respectively (Fig. B2). For our estimation methods the corresponding values were 0.54 ± 0.43 and 0.22 ± 0.31. Thus, our estimate of σ_1 was a little less biased, and our estimate of σ_2 was a little more biased than lmer, and the precision of our estimates was slightly higher.

LITERATURE CITED

Austin, P. C. 2010. Estimating multilevel logistic regression models with the number of clusters is low: a comparison of different statistical software procedures. International Journal of Biostatistics 6:Article 16.

Bates, D., M. Maechler, and B. Dai. 2008. lme4: Linear mixed-effects models using S4 classes.

Breslow, N. E., and D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88:9–25.

Burnham, K. T., and D. R. Anderson. 2002. Model selection and inference: a practical information-theoretic approach. Second edition. Springer, New York, New York, USA.

Harville, D. A. 1974. Bayesian inference for variance components using only error contrasts. Biometrika 61:383–385.

Liang, K. Y., and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73:13–22.

Schall, R. 1991. Estimation in generalized linear models with random effects. Biometrika 78:719–727.

Zeger, S. L., and K. Y. Liang. 1985. Longitudinal data analysis with generalized linear models. Biometrics 41:582–583.

TABLE B1. Parametric bootstrap exploration of the properties of the PGLMM models including the bootstrap mean and 95% inclusion interval of the distribution of estimates of σ, and the correlation between estimates; this provides information about the bias and precision of the estimators of σ. For each simulated data set, hypothesis tests (H0:$\sigma = 0$ at the $\alpha = 0.025$ level) and 95% confidence intervals were computed using profile likelihoods. The table reports the proportion of simulations in which the null hypothesis H0:$\sigma = 0$ was rejected, and the proportion of simulations in which the true estimate of σ was either below or above the 95% confidence interval. In all cases communities consisted of 32 species on a balanced phylogeny and 20 sites.

Model	True value	Bootstrap Mean	95% inclusion interval	Correlation	Reject H0:$\sigma = 0$	Below 95% conf. interval	Above 95% conf. interval
Model I (Eq. 1)							
σ_{spp}	0.71	0.55	(0.06, 0.84)	cor(σ_{phylo},σ_{slope})=–0.21	0.40	0.000	0.057
σ_{site}	0.71	0.68	(0.32, 1.01)		0.92	0.013	0.038
No phylogenetic signal							
σ_{spp}	0	0.19	(0, 0.60)	cor(σ_{phylo},σ_{slope})=–0.21	0.028	0.028	0
σ_{site}	0.71	0.71	(0.40, 1.03)		0.98	0.026	0.024
Model II (Eq. 2)							
σ_{phylo}	0.71	0.59	(0, 1.03)	cor(σ_{phylo},σ_{slope})=–0.44	0.76	0.016	0.109
σ_{slope}	0	0.19	(0, 0.57)	cor(σ_{slope},σ_{site})=0.01	0.10	0.102	0
σ_{site}	0.71	0.70	(0.38, 0.98)	cor(σ_{phylo},σ_{site})=0.10	0.98	0.013	0.035
No phylogenetic signal							
σ_{phylo}	0	0.19	(0, 0.73)	cor(σ_{phylo},σ_{slope})=–0.40	0.089	0.089	0
σ_{slope}	0.71	0.63	(0.13, 0.96)	cor(σ_{slope},σ_{site})=0.04	0.84	0.009	0.086
σ_{site}	0.71	0.70	(0.39, 1.00)	cor(σ_{phylo},σ_{site})=0.03	0.99	0.015	0.028
Model III (Eq. 3)							
$\sigma_{repulse}$	0.71	0.66	(0.15, 0.87)	cor($\sigma_{repulse}$,σ_{site})=–0.10	0.49	0.001	0.005
σ_{site}	0.5	0.48	(0.18, 0.77)		0.79	0.025	0.038
No phylogenetic signal							
$\sigma_{repulse}$	0	0.04	(0.01, 0.05)	cor($\sigma_{repulse}$,σ_{site})=0.13	0.004	0.004	0
σ_{site}	0.5	0.55	(0.27, 0.84)		0.90	0.064	0.014
Model IV (Eq. 7)							
σ_{trait}	0.71	0.66	(0.05, 1.26)	cor(σ_{trait},σ_{site})=–0.01	0.36	0.020	0.056
σ_{site}	0.5	0.50	(0.23, 0.77)		0.50	0.029	0.014
No phylogenetic signal							
σ_{trait}	0	0.003	(0.002, 0.004)	cor(σ_{trait},σ_{site})=–0.43	0.000	0.000	0
σ_{site}	0.5	0.50	(0.23, 0.75)		0.87	0.022	0.022

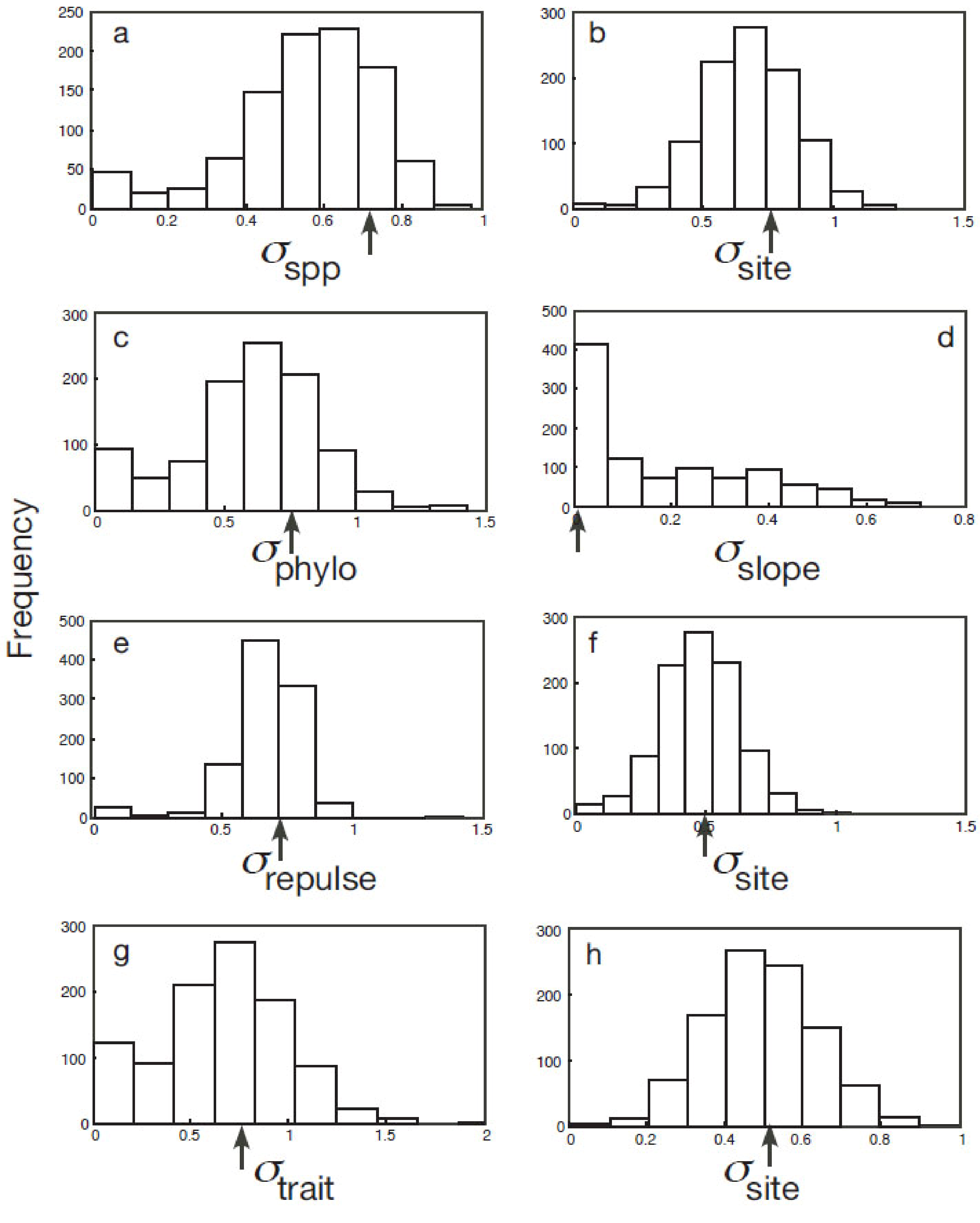

FIG. B1. Bootstrap distributions of parameters σ for (a,b) model I for phylogenetic patterns in species co-occurrences (Eq. 1), (c,d) model II for phylogenetic patterns in the sensitivities of species to and environmental gradient (Eq. 2), (e,f) model III for phylogenetic repulsion of related species (Eq. 3), and (g,h) model IV for trait values. True values of σ used to simulate the data are given by arrows and correspond to entries in Table B1 that contain phylogenetic signal.

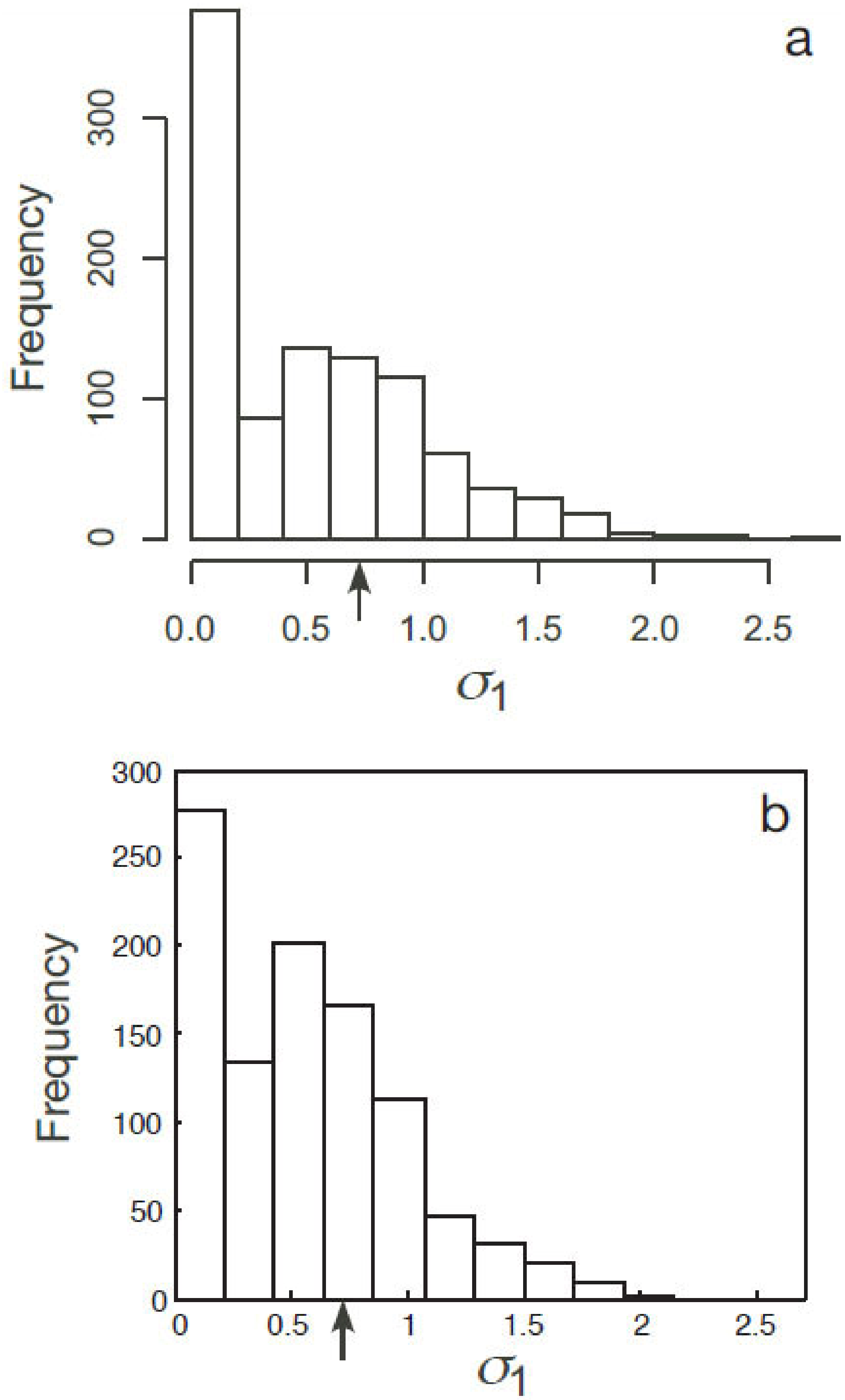

FIG. B2. Bootstrap distributions of parameter σ_1 for (a) lmer, and (b) the estimation procedure we used for PGLMMs (PQL and REML). The true value of $\sigma_1 = 0.71$ used to simulate the data is given by arrows.