Analysis of relative abundances on environmental gradients: supplemental information

Fiona Chong and Matthew Spencer

April 23, 2018

S1 Estimation methods

We fitted the model described in the main text using Bayesian estimation, via the NUTS algorithm (Hoffman and Gelman, 2014) in Stan 2.16.0, with the `rstan` package (version 2.15.1) and R version 3.4.1 (R Core Team, 2017). In order to speed up convergence by reducing posterior correlations, we actually used the orthogonal polynomial of order 2 for depth as the explanatory variable, and then obtained the coefficients for depth and squared depth by transformation, as described in Kennedy and Gentle (1980, pp. 342-345). We specified weak independent Cauchy$(0, 2.5)$ priors on β_0 and the coefficients of the orthogonal polynomial of depth, and weak independent half-Cauchy$(0, 2.5)$ priors on the standard deviations of ε. We chose an LKJ prior with scale parameter $\eta = 2$ on the correlation matrix Ω of ε (Lewandowski et al., 2009), for which the prior density is proportional to $\det \Omega$. This means that in the prior, the modal correlation matrix is the identity matrix, but the density is not very concentrated around this mode (Stan Development Team, 2017, section 60.1). We parameterized the correlation matrix using its Cholesky factor. We ran four independent chains, with 5000 warmup and 5000 sampling iterations. We checked that the chains had approximately converged by inspection of trace plots and potential scale reduction statistics (which were less than 1.0015 for all parameters). Effective sample size was at least 1950 for all parameters.

S2 Simulation study

We checked the performance of the estimation method on 100 simulated data sets, generated under the model using the posterior mean estimates from the real data (Tables S1 and S2). We fitted the model to each data set as described above. For almost all parameters (Tables S3 and S4), nominal 95% credible intervals contained the true parameter values 90 to 100 times out of 100. The exceptions were element 3 of the coefficient for squared depth effect, and elements $(6,6)$, $(7,7)$, $(8,1)$, $(8,6)$ and $(8,8)$ of the Cholesky factor of the covariance matrix, where the nominal 95% credible intervals contained the true parameters less often than required.

Posterior distributions of parameters from simulated data were generally centred not too far from the true values, although there was more variation in location for intercept and depth coefficients (Figure S1) than for elements of the Cholesky factor of the error covariance matrix (Figure S2). However, the posterior distributions of parameters for ilr components 7 and 8 had very long tails (Figures S1 and S2, bottom two rows).

The long-tailed distributions and poorer performance of credible intervals for parameters associated with components 7 and 8 is probably a consequence of the number of observations for the taxa whose distributions they reflect. Component 7 is proportional to the log of the ratio of *Aurelia aurita* to the geometric mean of all taxa other than *Aurelia aurita* and colonial ascidians, and component 8 to the log of the ratio of colonial ascidians to all other taxa. Since *Aurelia aurita* and colonial ascidians were the least abundant taxa, with non-zero point counts in only 7 and 15 out of 125 stills respectively, it is not surprising that estimation of parameters describing their distributions is more difficult than for other taxa.

References

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1351–1381.

Kennedy, Jr., W. J. and Gentle, J. E. (1980). *Statistical computing*. Marcel Dekker, New York.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100:1989–2001.

R Core Team (2017). R: A language and environment for statistical computing. `https://www.R-project.org/`.

Stan Development Team (2017). *Stan Modeling Language Users Guide and Reference Manual, Version 2.16.0*.

Table S1: Coefficients $\boldsymbol{\beta}_0$ (intercept), $\boldsymbol{\beta}_1$ (depth), $\boldsymbol{\beta}_2$ (squared depth) from the regression model described in the main text. Values are posterior means and 95% highest posterior density credible intervals in isometric logratio coordinates, with the default basis from the R package compositions.

$\boldsymbol{\beta}_0$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$
-1.11 (-1.30, -0.94)	-0.01 (-0.14, 0.11)	-0.62 (-0.79, -0.45)
0.50 (0.35, 0.65)	-0.10 (-0.20, 0.01)	-0.39 (-0.52, -0.24)
-3.00 (-3.45, -2.56)	-2.69 (-3.17, -2.24)	0.72 (0.29, 1.11)
1.04 (0.86, 1.23)	0.15 (-0.00, 0.31)	-0.24 (-0.40, -0.08)
-3.06 (-3.52, -2.64)	0.59 (0.30, 0.89)	0.39 (0.06, 0.71)
-2.14 (-2.64, -1.66)	1.11 (0.71, 1.50)	-0.17 (-0.56, 0.19)
-7.91 (-11.56, -5.17)	2.46 (0.05, 4.94)	0.56 (-0.56, 2.14)
-3.16 (-4.54, -1.90)	0.07 (-0.38, 0.58)	-0.04 (-0.60, 0.49)

Table S2: Cholesky factor of covariance matrix Σ from the regression model described in the main text. Values are posterior means and 95% highest posterior density credible intervals in isometric logratio coordinates, with the default basis from the R package compositions.

0.57 (0.46, 0.69)							
-0.15 (-0.26, -0.04)	0.40 (0.32, 0.49)						
-0.22 (-0.53, 0.08)	-0.43 (-0.73, -0.12)	1.04 (0.80, 1.32)					
-0.02 (-0.16, 0.13)	-0.00 (-0.15, 0.14)	0.07 (-0.09, 0.23)	0.55 (0.43, 0.67)				
-0.01 (-0.23, 0.18)	0.01 (-0.19, 0.23)	-0.08 (-0.36, 0.11)	-0.01 (-0.21, 0.19)	0.30 (0.00, 0.69)			
-0.34 (-0.67, -0.01)	0.25 (-0.09, 0.57)	-0.06 (-0.45, 0.31)	-0.05 (-0.38, 0.29)	0.14 (-0.50, 0.72)	0.84 (0.47, 1.21)		
-0.46 (-1.44, 0.50)	-0.43 (-1.46, 0.57)	-0.09 (-1.37, 1.10)	0.60 (-0.43, 1.77)	-0.08 (-1.44, 1.32)	-0.23 (-1.42, 1.01)	1.51 (0.47, 2.87)	
0.49 (-0.11, 1.17)	0.11 (-0.51, 0.75)	-0.22 (-0.90, 0.44)	-0.08 (-0.73, 0.57)	0.06 (-0.98, 1.07)	0.45 (-0.36, 1.32)	0.24 (-0.77, 1.19)	1.01 (0.31, 1.79)

Table S3: Percentage of simulated data sets for which nominal 95% credible intervals contained true parameter values for coefficients β_0 (intercept), β'_1 (linear term in orthogonal polynomial for depth) and β'_2 (quadratic term in orthogonal polynomial for depth). Rows are ilr components.

β_0	β'_1	β'_2
98	96	94
94	99	94
98	94	83
97	98	94
94	93	93
95	97	97
98	90	100
96	100	100

Table S4: Percentage of simulated data sets for which nominal 95% credible intervals contained true parameter values for Cholesky factor of error covariance matrix $\boldsymbol{\Sigma}$. Rows and columns are ilr components.

	1	2	3	4	5	6	7	8
1	98	0	0	0	0	0	0	0
2	95	94	0	0	0	0	0	0
3	96	93	99	0	0	0	0	0
4	97	96	99	93	0	0	0	0
5	100	100	100	100	99	0	0	0
6	97	91	99	99	100	86	0	0
7	92	100	100	92	100	100	84	0
8	80	100	99	100	100	89	100	78

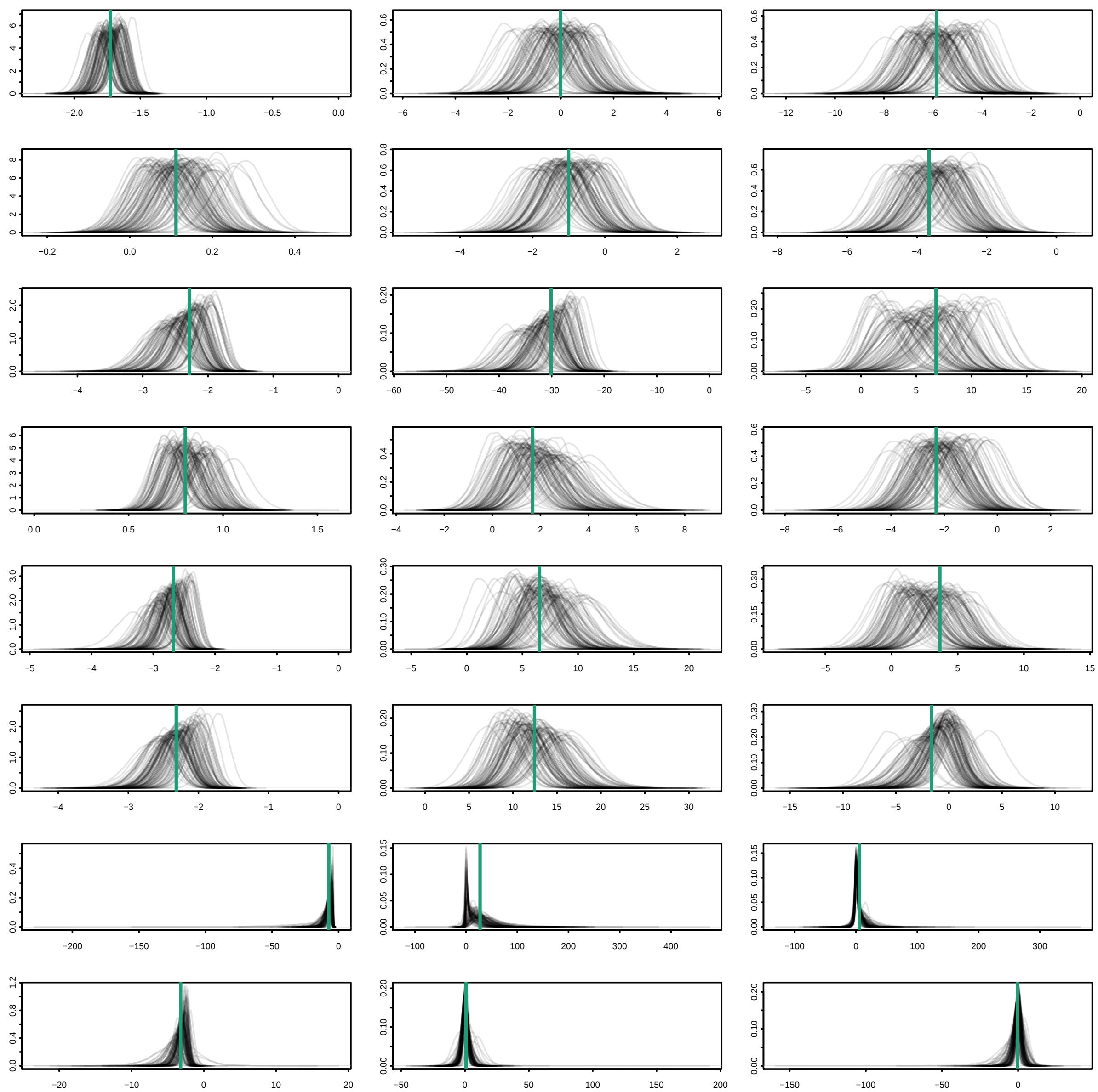

Figure S1: Posterior distributions (black lines) from 100 simulated data sets of estimated coefficients β_0 (intercept, first column), β_1' (linear term in orthogonal polynomial for depth, second column) and β_2' (quadratic term in orthogonal polynomial for depth, third column). Vertical green lines: true parameter values. Rows are ilr components.

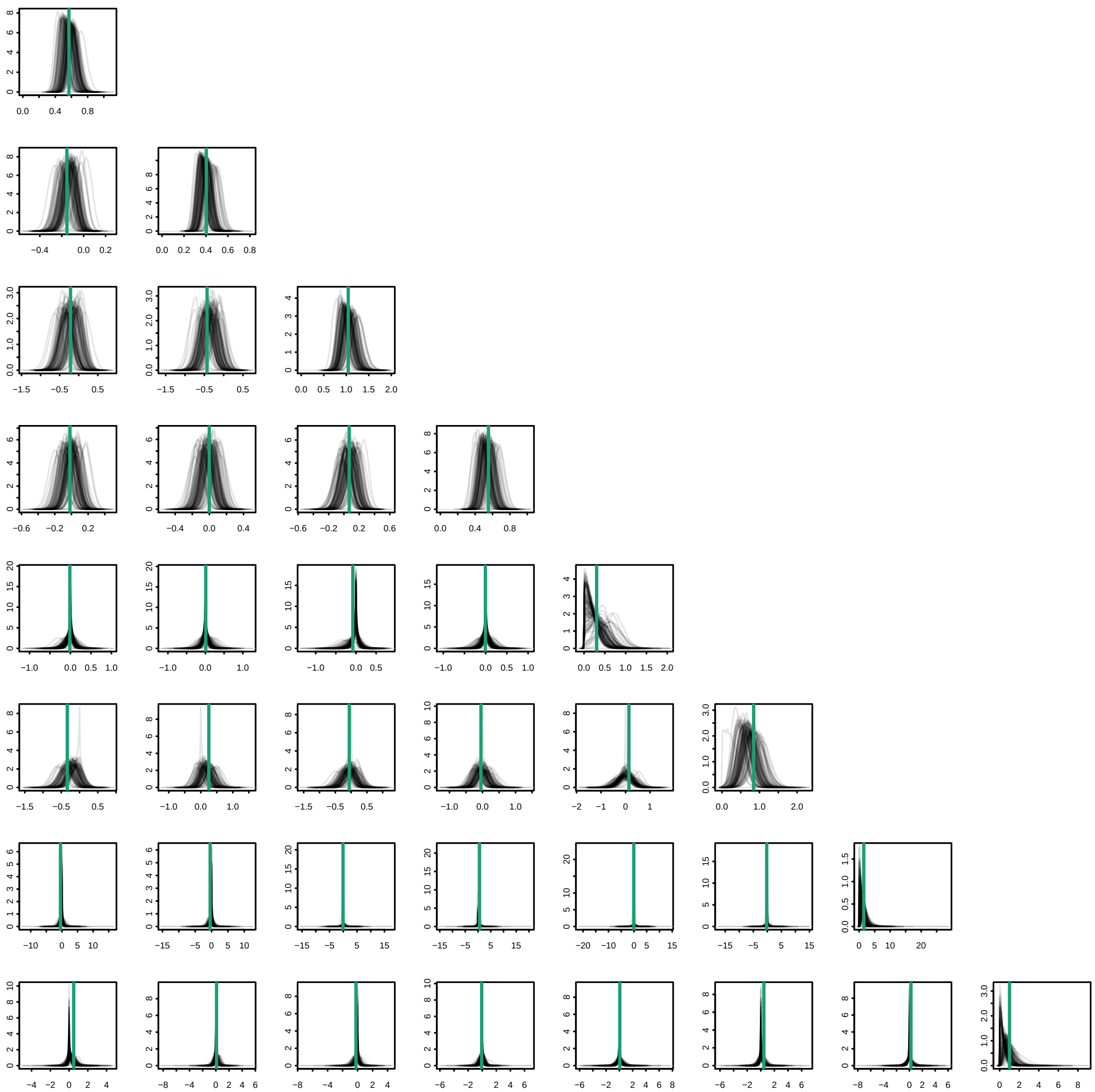

Figure S2: Posterior distributions (black lines) from 100 simulated data sets of estimated elements of the Cholesky factor of the error covariance matrix Σ. Vertical green lines: true parameter values. Rows and columns are ilr components.