Ecological Archives E095-290-A2

J. Wilson White, Steven G. Morgan, and Jennifer L. Fisher. 2014. Planktonic larval mortality rates are lower than widely expected. ***Ecology 95:3344-3353.***

Appendix B. Computational details and method validation.

Markov chain Monte Carlo optimization

Overview of Markov chain - Monte Carlo (MCMC) methods can be found elsewhere (Gilks et al. 1996, Lele et al. 2007, White et al. 2008). In short, MCMC is a family of approaches for simulating a posterior distribution of interest when calculating the posterior directly is not feasible. In particular, applying Bayes' rule to calculate the posterior for a high-dimension problem (i.e., many parameters) requires calculating an integral over that high-dimension parameter space, and MCMC provides an efficient shortcut around that integration step. Because it simulates the posterior directly, it can be more flexible than maximum likelihood approaches that rely on particular assumptions about the shape of a likelihood surface.

We used the Metropolis algorithm (Gilks et al. 1996) to create a Markov chain of parameter values that approximate the posterior distribution $\Pr(\theta, k \mid n_{i,t}, n_{i+1,t})$. Briefly, the Metropolis algorithm operates as follows: given data **D** and an estimate of parameter values $\Xi^m = (k, \theta)$ representing the mth entry in the Markov chain, sample new candidate parameter values Ξ' from the proposal distribution $Q(\mathbf{Y})$. Then calculate the likelihoods $L(\mathbf{D} \mid \Xi')$ and $L(\mathbf{D} \mid \Xi^m)$ and the priors $\pi(\Xi')$ and $\pi(\Xi^m)$. Assuming a symmetrical proposal distribution (which we used, see next paragraph), the candidate values Ξ' are accepted with probability p that is equal to the ratio of the posteriors: $p = \min[1, \pi(\mathbf{D} \mid \Xi') / \pi(\mathbf{D} \mid \Xi^m)]$ (this is the Metropolis rule). If Ξ' is accepted, it becomes entry Ξ^{m+1} in the Markov chain; if it is not accepted, the chain does not move and

$\Xi^{m+1} = \Xi^m$. Given an arbitrary set of starting parameter values, the algorithm will eventually converge on a solution in which it consistently draws new values of Ξ from the posterior distribution of Ξ. After M iterations, the first b iterations are discarded (the 'burn-in' containing the approach to convergence) and every c^{th} element of the remaining chain is saved (c is the thinning interval); those values of Ξ are used as an estimate of the posterior.

Both model parameters (k and θ) are biologically constrained to be positive real numbers, and could in theory take on values that span several orders of magnitude. Therefore we found it more efficient to estimate the posterior distributions of $\log_{10} k$ and $\log_{10} \theta$, exponentiating the parameter values as necessary to perform the VLT calculations during MCMC. Lacking reliable prior information on the distribution of $\log_{10} k$ and $\log_{10} \theta$ (in part because of possible deficiencies in earlier mortality estimates), we used noninformative normal priors for both parameters: $N(0, 1\times10^4)$ and $N(-1, 1\times10^4)$, for $\log_{10} k$ and $\log_{10} \theta$, respectively.

We used a one-at-a-time Metropolis algorithm in which only one element of Ξ (i.e., one parameter value) was altered in each iteration. The proposal distribution $Q(\Xi^m)$ was a symmetric normal distribution with mean Ξ^m and a coefficient of variation (cv) ranging from 2 to 0.1. We used a delayed-rejection algorithm to improve mixing, so if candidate value Ξ' was not accepted, a new candidate value was generated from a proposal distribution with a smaller cv. Each round of delayed rejection began with cv = 2 and stepped down exponentially from 2 to 0.1; if a candidate value was still not accepted then the chain did not move ($\Xi^{m+1} = \Xi^m$) and the algorithm moved to the next step.

We ran the algorithm for $M = 5\times10^4$ iterations, discarded the first $b = 10^3$ as the burn-in, and used a thinning interval of $c = 10$ to avoid autocorrelation in the chain. For each species we ran 3 duplicate chains using different starting values to ensure convergence on a common

solution; starting values were generated as random draws from a normal distribution with standard deviation 0.5 and mean of 0.5 (for mortality rates θ) or 10 (for clumping parameter k). Chains were visually inspected for convergence; those that converged successfully were pooled to obtain estimate the posterior. Trial and error suggested that these options produced desirable levels of mixing in the chains and that results were not sensitive to initial starting conditions.

Likelihood calculation

The MCMC algorithm requires calculation of the likelihood $L(\mathbf{D}|\Xi)$. As mentioned above, we were unable to derive an analytical expression for $L(n_{i,t}, n_{i+1,t} \mid \theta, k)$. In such cases, it is possible to use the method of implicit likelihood, in which one simulates the stochastic process assumed to underlie the data and uses the resulting distribution of random values as an approximation of the desired probability distribution (Diggle and Gratton 1984, Marjoram et al. 2003). In this case, we independently simulated random values for the abundance of each stage, $n_{i,t}^*$ and $n_{i+1,t}^*$ (where the asterisk indicates simulated data) for particular values of k and θ, then used the ratio of those simulated values to approximate the distribution of the ratio $n_{i,t}^*/n_{i+1,t}^*$. To do this, we generated 10^4 random values of $n_{i,t}^*$ from a negative binomial distribution with overdispersion parameter k and an expected value (parameter m in Eq. 6) given by Eq. (1). We generated the same number of values of $n_{i+1,t}^*$ in the same way, but using the expected value given by Eq. (3). We then took the ratio of each pair of $n_{i,t}^*$ and $n_{i+1,t}^*$ values to obtain 10^4 estimates of $n_{i,t}^*/n_{i+1,t}^*$. In simulating these values we assumed that $\square_i = 1$; because we take the ratio $n_{i,t}^*/n_{i+1,t}^*$, the result is not sensitive to this choice of dummy value. The standard approach to calculating implicit likelihoods involves using nonparametric kernel density estimation to approximate a smooth probability distribution from the distribution produced by stochastic simulation (Diggle and Gratton 1984). We found that kernel density estimation was too

computationally expensive to be feasible for the hundreds of thousands of likelihood calculations required by MCMC runs. Instead, we grouped the $n_{i,t}^{*}/n_{i+1,t}^{*}$ into 1000 equally spaced bins ranging from zero to the maximum observed value (in either the actual or simulated data) and used the proportion of observations in each bin divided by the bin width as the estimate of the probability density for that bin. This procedure allowed us to estimate the likelihood of an observed ratio $n_{i,t}/n_{i+1,t}$ given the approximate distribution $n_{i,t}^{*}/n_{i+1,t}^{*}$ corresponding to particular estimates of the parameters k and θ (and known values for $\square_{i}$ and $\square_{i+1}$). To obtain consistent results we performed 5 replicate simulations of the likelihood and took the average of those values as the likelihood of a particular value of θ and k.

Validation using simulated data

To check the validity of our estimation method, we tested it on a suite of simulated datasets with known values of k and θ. We simulated four different larval life histories with two larval stages and either the same or different mortality rate in each stage (Table B1). Each stage had a 50-day duration, and recruitment to the first stage was 5×10^{4} larvae day^{-1}. We simulated ten replicate datasets for each life history, then applied our estimation method to each dataset. The results revealed that our method accurately estimated both k and θ for cases in which θ was constant across larval stages. When θ varied, the method estimated an value in between that of the two stages (Fig. B1). These estimates consistently outperformed estimates obtained using the original VLT method (Table B2), which were rarely close to the true value and had very large standard deviations, indicating broad confidence regions.

Uncertainty in stage duration

One possible uncertainty in our analysis is incorrect estimation of stage duration, α_i. Although stage duration depends strongly on water temperature and we chose values based on studies from environments with similar temperatures to our field site, it is possible that the actual durations were a few days shorter or long. To evaluate the potential impact of this uncertainty, we examine Eq [4] in the main text, which shows the relationship between θ, α, and the ratio of abundances of stages (the data used for calculation).

$$n_{i,t}/n_{i+1,t} = [\exp(\theta_i\alpha_i)-1]/[1 - \exp(-\theta_{i+1}\alpha_{i+1})].$$

Assuming that adjacent stages have equal duration (which is true for most cases in our dataset), we can calculate the value of the stage abundance ratio for different values of θ□□□α (Fig. B2). For relatively low values of q (e.g., ≤ 0.01 day^{-1}) there is very little variation in the stage abundance ratio for a wide range of stage durations. However, for higher mortality rates (e.g., > 0.1 day^{-1}), the abundance ratio is quite sensitive to the value of the stage duration. Therefore the mortality rate estimates may be more sensitive to inaccurate duration data. To gauge this effect, we re-estimated the mortality rate for the scenario 2 simulated dataset (Table B1; mortality of 0.8 day^{-1}) with an actual stage duration of 5 days, but specifying an incorrect duration of either 3 days or 7 days in the estimation procedure. When the stage duration was under-estimated, the method still accurately estimated the mortality rate; when the stage duration was over-estimated, the mortality rates were slightly under-estimates (Fig. B3). Thus even for a very high mortality rate, corresponding to highest sensitivity of the abundance ratio to inaccurate stage duration, mortality estimates were not wildly inaccurate. Fortunately most of the estimated mortality rates in our study were ≤ 0.1 day^{-1}, so we have higher confidence in their robustness to incorrect estimates of stage duration.

LITERATURE CITED

Diggle, P. J., and R. J. Gratton. 1984. Monte carlo methods of inference for implicit statistical models. Journal of the Royal Statistical Society B 46:193–227.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. Markov chain Monte Carlo in practice. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Lele, S. R., B. Dennis, and F. Lutscher. 2007. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecology Letters 10:551–563.

Marjoram, P., J. Molitor, V. Plagnol, S. Tavaré. 2003. Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences USA 100:15324–15328.

White, J. W., J. D. Standish, S. R. Thorrold, and R. R. Warner. 2008. Markov chain Monte Carlo methods for assigning larvae to natal sites using natural geochemical tags. Ecological Applications 18:1901–1913.

TABLE B1. Larval life histories used in simulated data.

Scenario	**No. stages**	$\boldsymbol{k}$	$\boldsymbol{\theta}$
1	2	0.4	0.3, 0.3
2	2	0.4	0.8, 0.8
3	2	0.4	0.8, 0.2
4	2	0.4	0.2, 0.8

TABLE B2. Estimated mortality rates (including standard deviation and number of useable observations [out of 50]) for simulated datasets using original VLT method.

Scenario	**Replicate**	**Actual θ □□ays□**	**Estimated θ (SD)**	***n***
1	1	0.3, 0.3	1.67 (1.84)	49
1	2	0.3, 0.3	1.25 (1.86)	50
1	3	0.3, 0.3	1.42 (2.07)	50
1	4	0.3, 0.3	0.91 (1.05)	48
1	5	0.3, 0.3	1.28 (1.89)	50
1	6	0.3, 0.3	2.02 (1.98)	49
1	7	0.3, 0.3	1.42 (1.93)	48
1	8	0.3, 0.3	1.82 (2.42)	49
1	9	0.3, 0.3	1.20 (1.38)	50
1	10	0.3, 0.3	1.29 (1.67)	50
2	1	0.8, 0.8	0.34 (0.87)	45
2	2	0.8, 0.8	0.26 (0.70)	48
2	3	0.8, 0.8	0.24 (0.64)	46
2	4	0.8, 0.8	0.61 (1.21)	49
2	5	0.8, 0.8	0.57 (1.23)	44
2	6	0.8, 0.8	0.48 (1.19)	47
2	7	0.8, 0.8	0.19 (0.37)	49
2	8	0.8, 0.8	0.21 (0.65)	48
2	9	0.8, 0.8	0.30 (0.99)	49
2	10	0.8, 0.8	0.46 (1.39)	47
3	1	0.8, 0.2	0.34 (0.59)	46
3	2	0.8, 0.2	0.60 (1.06)	50
3	3	0.8, 0.2	0.48 (1.13)	49
3	4	0.8, 0.2	0.61 (1.23)	49
3	5	0.8, 0.2	0.24 (0.72)	49
3	6	0.8, 0.2	0.28 (0.73)	48
3	7	0.8, 0.2	0.30 (0.51)	48
3	8	0.8, 0.2	0.70 (1.42)	47
3	9	0.8, 0.2	0.26 (0.56)	49
3	10	0.8, 0.2	0.72 (1.45)	49
4	1	0.2, 0.8	0.61 (1.29)	48
4	2	0.2, 0.8	0.56 (0.99)	50
4	3	0.2, 0.8	0.48 (0.93)	48
4	4	0.2, 0.8	0.61 (1.22)	49
4	5	0.2, 0.8	1.27 (1.93)	49
4	6	0.2, 0.8	0.75 (1.67)	50
4	7	0.2, 0.8	0.77 (1.29)	50
4	8	0.2, 0.8	0.99 (1.53)	49
4	9	0.2, 0.8	0.65 (0.96)	50
4	10	0.2, 0.8	0.61 (1.17)	48

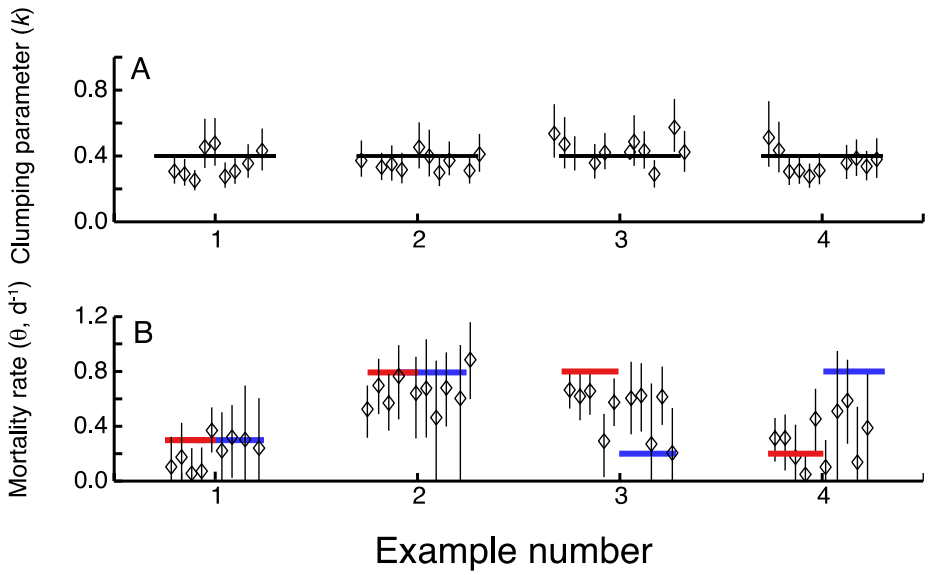

FIG. B1. Posterior estimates of (A) the clumping parameter k and (B) mortality rate θ for four simulated datasets. Diamonds indicate posterior mean values estimated for each of the ten replicate datasets (error bars show 95% HPD intervals). Horizontal bars show true estimates of each parameter. In (B), true mortality rates for stages are indicated as red (early) and blue (late).

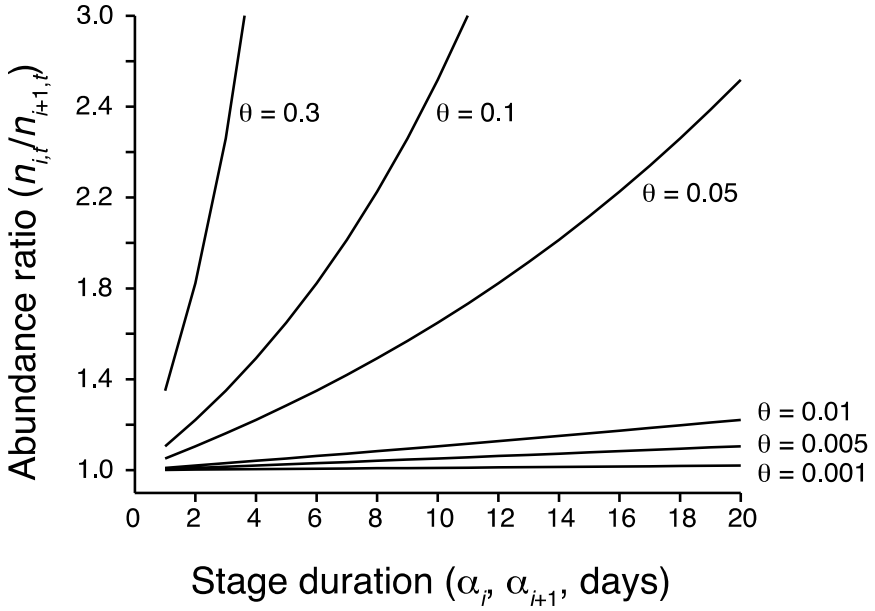

FIG. B2. Variation in the abundance ratio for a range of values of the mortality rate (θ, days^{-1}) and the stage duration of adjacent stages i and i+1, based on the relationship in Eq. [4].

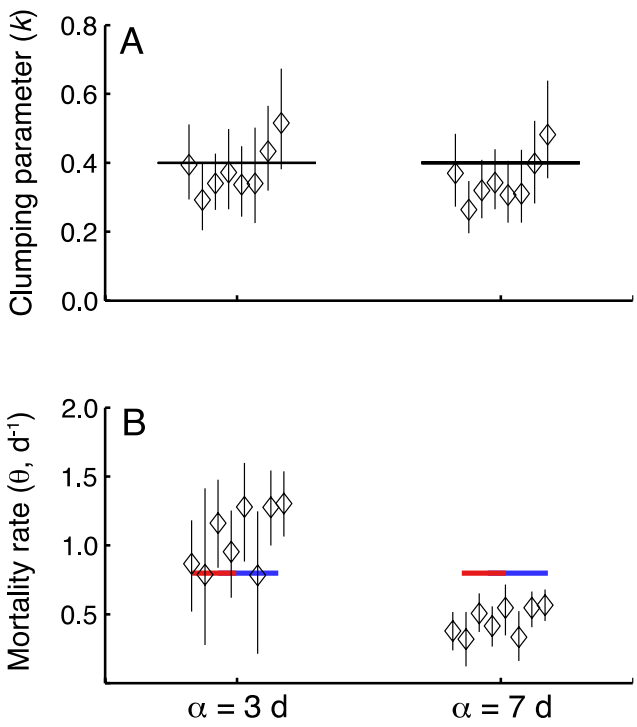

FIG. B3. Posterior estimates of the (A) clumping parameter *k* and (B) mortality rate θ for the scenario 2 simulated dataset, when the correct stage duration of 5 days was mis-parameterized as either 3 days or 7 days. Diamonds indicate posterior mean values estimated for each of eight replicate datasets (error bars show 95% HPD intervals). Horizontal bars show true estimates of each parameter. In (B), true mortality rates for stages are indicated as red (early) and blue (late).