Ecography

E4899

Forster, M. A. and Warton, D. I. 2007. A metacommunity-scale comparison of species-abundance distribution models for plant communities of eastern Australia. – Ecography 30: 449–458.
.

Appendix 1. Community distances and locations

Table S1. The shortest distance between each local community (in kilometres) at which abundance was measured for the five metacommunities used in this study.

Closed forest	1	2	3	4	5
1	0				
2	28	0			
3	33	6	0		
4	33	8	5	0	
5	123	103	95	98	0

Open forest	1	2	3	4	5
1	0				
2	10	0			
3	10	1	0		
4	10	1	1	0	
5	15	5	5	5	0

Open forest/ woodland	1	2	3	4	5
1	0				
2	4	0			
3	8	9	0		
4	10	7	2	0	
5	13	10	8	5	0

Woodland	1	2	3	4	5
1	0				
2	6	0			
3	20	14	0		
4	23	15	2	0	
5	20	15	10	10	0

Ridge top woodland	1	2	3	4	5
1	0				
2	5	0			
3	5	10	0		
4	5	8	5	0	
5	4	8	2	3	0

Table S2. Metacommunity locations in latitude (South) and longitude (East) based on Geocentric Datum of Australia 1994.

Metacommunity	Local community	Latitude	Longitude
Closed forest	1	33° 29′ 892″	151° 23′ 516″
	2	33° 44′ 351″	150° 18′ 006″
	3	33° 39′ 676″	150° 17′ 403″
	4	33° 43′ 486″	150° 19′ 161″
	5	34° 59′ 442″	150° 59′ 575″
Open forest	1	33° 15′ 367″	151° 14′ 455″
	2	33° 16′ 522″	151° 17′ 683″
	3	33° 16′ 257″	151° 17′ 731″
	4	33° 16′ 691″	151° 18′ 087″
	5	33° 16′ 976″	151° 18′ 444″
Open forest/ woodland	1	33° 38′ 623″	150° 18′ 159″
	2	33° 35′ 499″	150° 15′ 977″
	3	33° 32′ 664″	150° 20′ 612″
	4	33° 35′ 516″	150° 15′ 989″
	5	33° 36′ 642″	150° 16′ 575″
Woodland	1	34° 27′ 694″	150° 24′ 672″
	2	34° 23′ 288″	150° 19′ 165″
	3	34° 20′ 582″	150° 13′ 745″
	4	34° 32′ 519″	150° 28′ 662″
	5	34° 22′ 664″	150° 27′ 181″
Ridge top woodland	1	33° 40′ 212″	151° 10′ 650″
	2	33° 39′ 258″	151° 13′ 583″
	3	33° 40′ 750″	151° 07′ 509″
	4	33° 42′ 609″	151° 10′ 028″
	5	33° 41′ 286″	151° 08′ 943″

Appendix 2. Technical details

This appendix describes technical details in using maximum likelihood estimation to estimate parameters for the three distributions which were computationally difficult:

- the zero-sum multinomial model (ZSM)
- the compound Poisson-lognormal model (PLN)
- the compound negative binomial-lognormal model (NBLN)

To use maximum likelihood estimation involves several steps:

- Derive the probability distribution function for the given model
- Calculate the likelihood function, given a set of values for parameters. Because of the form of the probability function, this involved numerical integration
- Use a generic optimisation routine to find the set of values for parameters that maximises the likelihood function

Each of these three steps is discussed in a separate section, below.

Probability distribution functions

For each of the three models considered here, the probability distribution function does not have a closed form. Instead, each probability distribution is an integral over all possible values of some measure of expected abundance M.

For the pooled data model, we observe an abundance x that is a realisation of some random variable X, whose probability distribution function has the form:

$$P(X=x)=\int_0^a P(M=m)\ P(X=x \mid M=m)\ dm \qquad (1)$$

For the metacommunity model, we observed a cluster of five abundance measurements of each species (one at each of five local communities), which we can write as a vector $\mathbf{x}=(x^{(1)},x^{(2)},x^{(3)},x^{(4)},x^{(5)})$, and as a realisation of the 5-variate distribution $\mathbf{X}=(X^{(1)},X^{(2)},X^{(3)},X^{(4)},X^{(5)})$. The joint probability distribution involves a product of five terms of the form $P(X^{(i)}|M)$:

$$P(X=x)=\int_0^a P(M=m)\prod_{j=1}^{5} P(X^{(i)}=x^{(i)} \mid M=m)\ dm \qquad (2)$$

Table S3 summarises the values of a, P(M) and P(X|M) for the three models under consideration. For the ZSM model fitted as a metacommunity model, the number of individuals in the local community varied from one local community to another. Consequently, the total number of individuals in the ith local community will be written as $J^{(i)}$, and similarly for other parameters that vary with local community.

In the case of the zero-sum multinomial model, the terms given in Table 6 give expected counts of species with abundance x, rather than a probability. Consequently, P(X=x) is only determined up to some proportionality constant k. The constant can be found by summing kP(X=x) over all possible non-zero values of x (x=1,...,J), and noting that the sum is equal to one, as described in Alonso and McKane (2004). Alternatively, a much more efficient approach to finding the proportionality constant is to estimate k[1–P(X=0)]=kP(X>0) and set it to one. This is more efficient because it only involves evaluation of P(X=x) at one point (x=0) rather than at J values, achieving the same outcome but with substantial reductions in computation time.

The parameters of the zero-sum multinomial model require some explanation:

- θ is the "fundamental biodiversity parameter", a measure of how species rich the metacommunity is.
- γ is an immigration parameter. This can be expressed in terms of a parameter λ representing the probability of immigration from the metacommunity, $\gamma = \frac{(J-1)\gamma}{1-\lambda}$.
- J is the total number of individuals in the local community. This was taken to be the total number of observations in the sample to which the model was fitted.

Table S3. The values of a, P(M) and P(X|M) for equations (1) and (2) that give the probability distribution function for the three models considered here. Note that k is a constant that is estimated as described in the text, $\phi(x;\mu,\sigma^2)$ is the probability function for the normal distribution with mean μ and variance σ^2, and C_x^n ("n choose x") is the number of possible combinations of x objects chosen from n distinct objects. The terms for the zero-sum multinomial model are taken from Vallade and Houchmandzadeh (2003), when J_m is infinitely large.

model	parameters	a	P(M=m)	P(X=x|M=m)
ZSM	θ,γ,J,k	1	$k\frac{\theta}{m}(1-m)^{\theta-1}$	$C_x^J \frac{\Gamma(\gamma m+x)}{\Gamma(\gamma m)} \frac{\Gamma\left(\gamma(1-m)+J-x\right)}{\Gamma\left(\gamma(1-m)\right)} \frac{\Gamma(\gamma)}{\Gamma(\gamma+J)}$
PLN	μ,σ	∞	$\frac{1}{m}\phi(\log(m);\mu,\sigma^2)$	$\frac{e^{-m}m^x}{x!}$
NBLN	μ,σ,κ	∞	$\frac{1}{m}\phi(\log(m);\mu,\sigma^2)$	$C_x^{x+1/\kappa-1} \frac{\kappa^x m^x}{(1+\kappa m)^{x+1/\kappa}}$

The term P(M=m) was estimated for the zero-sum multinomial model using the asymptotic limit as the total number of individuals in the metacommunity (J_m) approaches infinity (as in Alonso and McKane 2004). This limit provides a good approximation to probabilities when J_m is reasonably large (>1000, say). Because J>1000 for all local communities in the study, clearly J_m is very large.

Zero-correction and conditioning on total abundance

A species will only appear in the species list for a metacommunity if it is present in at least one of the local communities. The log-likelihood was corrected to incorporate this information, i.e. it was calculated conditional on abundance being non-zero (for pooled data analyses) and conditional on at least one local abundance being non-zero (for metacommunity analyses).

In addition, log-likelihoods were also conditioned on total abundance, so that results would be comparable across all models fitted. For the ZSM model, the probabilities are calculated given knowledge of (i.e. conditional on) the total abundance in the local community. The log-likelihood calculations for other models were corrected so that they also condition on total abundance in the local community.

So in summary, models were zero-corrected and conditioned on total abundance as follows:

- For pooled data analyses, we used $P(X = x \mid 0 < X \leq J) = \frac{P(X = x)}{P(0 < X \leq J)}$ in log-likelihood calculations, where J is the total of observed abundances in the pooled dataset.
- For the metacommunity analyses, we used $P(\mathbf{X} = \mathbf{x} \mid 0 < X^{(i)} \leq J^{(i)},\ \forall i) = \frac{P(\mathbf{X} = \mathbf{x})}{\prod_{i=1}^{5} P(0 < X^{(i)} \leq J^{(i)})}$ in log-likelihood calculations, where $J^{(i)}$ is the total abundance in the ith local community.

Note that for the zero sum multinomial model, in the previous section we described methods for calculating the proportionality constant k in order to translate expected counts into probabilities. The method we used to estimate this correction term involved conditioning on non-zero abundance, so no additional corrections are required here for the zero-sum multinomial (as they have already been applied).

Avoiding floating point errors

In several instances for which m was small, calculating the probability distribution function involved calculating the ratios of very small numbers, which sometimes led to problems with floating point errors. To avoid such computational problems, we calculated the limit as $m \to 0$, as described in the following.

For the zero-sum multinomial estimated from pooled data, we used the following limit when $\gamma m < 10^{-8}$:

$$\lim_{m \to 0} \frac{P(X = x \mid M = m)}{P(0 < X \le J \mid M = m)} = \frac{1}{\sum_{j=0}^{J-1} \frac{1}{\gamma + j}} \frac{\Gamma(J+1)\Gamma(J+\gamma-x)}{\Gamma(J-x+1)\Gamma(J+\gamma)}$$

For the zero-sum multinomial estimated for the metacommunity model, we used the following limit when $\gamma m < 10^{-8}$:

$$\lim_{m \to 0} \frac{P(\mathbf{X} = \mathbf{x} \mid M = m)}{P(0 < X^{(i)} \le J^{(i)},\ \forall i \mid M = m)} = \frac{m^4}{\prod_{i=1}^{5} \gamma^{(i)} \left(\sum_{j=0}^{J^{(i)}-1} \frac{1}{\gamma^{(i)} + j} \right)} \prod_{\{i: x^{(i)} > 0\}} \frac{\gamma^{(i)}}{x^{(i)}} \frac{\Gamma(J^{(i)}+1)\Gamma(J^{(i)}+\gamma^{(i)}-x^{(i)})}{\Gamma(J^{(i)}-x^{(i)}+1)\Gamma(J^{(i)}+\gamma^{(i)})}$$

For the lognormal distributions estimated from pooled data, we use the following limit when $m<10^{-8}$:

$$\lim_{m \to 0} P(0 < X \le J) = m$$

For the lognormal distributions estimated for the metacommunity model, we use the following limit when $5m<10^{-8}$:

$$\lim_{m \to 0} P(0 < X^{(i)} \le J^{(i)},\ \forall i) = 5m$$

An additional problem described in McGill et al. (2006) is that the numbers on the numerator and denominator can get very large, causing computational issues. We resolved using the standard approach of doing calculations on the log-scale (as in McGill et al. 2006).

Numerical integration to calculate the likelihood function

Given a set of possible values for parameters, the likelihood function can be calculated as the product of probabilities. If the observed species abundances are $(x_1,x_2,...,x_n)$, then the likelihood function for the pooled data model is

$$\prod_{j=1}^{n} P(X_j = x_j \mid 0 < X_j \le J)$$

For the metacommunity model, the observed species abundances are $(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n)$ and the likelihood function is

$$\prod_{i=1}^{5} \prod_{j=1}^{n} P(X_j = \mathbf{x}_j \mid O < X_j^{(i)} \le J^{(i)},\ \forall i)$$

For computation, this is more conveniently written as a log-likelihood, which is a sum of log probabilities.

Because each probability was an integral that could not be written in closed form, numerical integration was required to estimate the value of the likelihood function. We used Matlab's quad function for numerical integration for compound lognormal distributions, but resorted to our own numerical integration function in the zero-sum multinomial case.

As in McGill et al. (2006), the integrand for the zero sum multinomial distribution often approaches a delta function (i.e. a function with all its probability mass at one point). This causes problems for numerical integration, as standard algorithms such as Matlab's quad can miss the peak altogether and grossly underestimate probabilities. Contrary to McGill et al. (2006), we found that Matlab's quad8 and quadl functions did not resolve the problem, indeed they sometimes perform worse that quad. We resolved this problem by "brute force": designing our own numerical integration functions that evaluated the function of interest at many points (specifically, at 100 points initially, increasing this number by a factor of 20 for each required iteration). Simpson's rule was applied in order to estimate the integral. This approach is not very computationally efficient, due to the large number of function evaluations, however it was a simple approach which addressed the problem and returned solutions within a short timeframe (most parameter estimation routines took less than 10 min to complete on a pentium 4 processor). Our two numerical integration functions, quadDave and quadDave2 are available from David Warton's website (http://web.maths.unsw.edu.au/~dwarton/).

We note that for our datasets, use of specialised numerical integration functions in place of the quad function made negligible difference to results. However, in other instances (such as an analysis of the Barro Colorado Island tree dataset) this can affect results. Also, we note that choice of numerical integration function was not an issue at all when fitting metacommunity models, because the multiple abundance observations tended to spread the probability mass more evenly over the range of possible values for M. Further, our software converged to a solution very quickly for metacommunity models (usually within one minute), presumably because the likelihood and integrand functions were more well-behaved when data from multiple communities was available for model estimation.

For the compound lognormal distributions, the integrals with respect to m that appear in the log-likelihood function were unsuitable for use with a standard numerical integration algorithm, again due to behaviour like a delta-function for some parameter sets. A simple solution available to us in this case was transformation: to integrate with respect to y, where y is a transformation of m for which the function being integrated is more suitable for numerical integration – a relatively smooth function that is defined over a finite range of values of y. We found a suitable transformation which gives good and reliable estimates of the log-likelihood function under a wide range of test conditions:

$$\log\left(\frac{1+y}{1-y}\right) = \frac{\log(m) - \alpha}{\varpi}$$

The log-likelihood function is estimated by writing integrals with respect to y and integrating from –1 to 1, and applying a numerical integration algorithm.

The success of this transformation depends heavily on the choices of α and ω. Good choices will involve ensuring that the integrand always has non-trivial values only when $\frac{\log(m)-\alpha}{\varpi}$ is near zero – this ensures that the integrand with respect to y has a regular shape with a peak near zero, rather than it rising to a maximum near one of the limits of integration. This can be achieved by applying likelihood theory to the integrand: find the maximum likelihood estimator of log(m) in the integrand, and find the observed information for log(m) (i.e. the negative of the second derivative with respect to log(m)). We can then choose α to be the maximum likelihood estimator of log(m) (which centres values of $\frac{\log(m)-\alpha}{\varpi}$ about zero), and choose ω to be the square root of the inverse of the observed information (which rescales $\frac{\log(m)-\alpha}{\varpi}$ so that the density is spread out over the neighbourhood of zero). Doing this, α and ω solve the following equations:

$$\frac{\alpha-\mu}{\sigma^2}+e^{\alpha}-x_j=0, \quad \text{if X|M is Poisson}$$

$$\frac{\alpha-\mu}{\sigma^2}+\frac{e^{\alpha}-x_j}{1+\kappa e^{\alpha}}=0, \quad \text{if X|M is negative binomial}$$

$$\frac{1}{\varpi^2}=\begin{cases}\frac{1}{\sigma^2}+e^{\alpha}, & \text{if X|M is Poisson}\\ \frac{1}{\sigma^2}+e^{\alpha}\frac{(1+\kappa x_j)}{(1+\kappa e^{\alpha})^2}, & \text{if X|M is negative binomial}\end{cases}$$

Notice that α and ω are functions of x_j – so we used a different transformation for species with different abundances. When $x_j=0$, we used $\alpha=\mu$ and $\omega=\sigma$.

Maximisation of the log-likelihood

Maximisation of the log-likelihood used the generic Matlab function "fminsearch". This function required all parameters to be unbounded, and an initial estimate of parameters, as described below.

The fminsearch function required parameters to be unbounded, i.e. defined for any real number. However, some parameters were only defined over a bounded region, so the model needed to be reparameterised before using fminsearch. The reparameterisations were as follows:

- λ (defined for $0 \le \lambda \le 1$) was reparameterised as $\eta = \log\left(\frac{\lambda}{1-\lambda}\right)$

- σ (defined for $\sigma \geq 0$) was reparameterised as $\sigma_{new} = |\sigma|$. Similarly for κ, τ.

A set of initial estimates for parameters was required, so that fminsearch can find a solution by iteration from this starting point. From trials using different starting values, and from general exploration of the parameter space, it was found that models converged to a unique solution in all cases except for NBLN fitted to pooled data. Careful choice of starting values only affected time to convergence rather than the solution itself, except for NBLN fitted to pooled data.

For the zero-sum multinomial model, the chosen starting values were θ=4, λ=0.5.

For the compound lognormal distributions, obvious starting values for m and s are the sample mean and variance of log-transformed data. For NBLN fitted to pooled data, several initial estimates were used for κ (0, 0.01, 0.1, 1, 10, 100), and when fitted as a metacommunity model, we used κ=0 (noting that in extensive testing, results converged to a unique solution irrespective of choice of κ).

Appendix 3. Parameter estimates

Table S4. Parameters for corrected values from Table 2. OFW = Open Forest/Woodland; RTW = Ridge Top Woodland.

	ZSM		PLN		NBLN			NB		LS
	θ	m	μ	σ	log(μ)	σ	ϕ	μ	ϕ	θ
Closed forest	2.9	0.35	2.8	2.3	4.2	0.0	7.2	66.8	7.2	0.999
Open forest	10.3	0.06	3.2	1.5	3.5	1.3	0.5	59.8	2.7	0.998
OFW	25.9	0.003	2.7	2.4	4.1	0.0	6.7	62.2	6.7	0.999
Woodland	31.8	0.005	2.2	2.4	3.4	0.0	8.1	31.2	8.1	0.997
RTW	19.1	0.01	2.9	1.8	3.8	0.0	2.8	46.9	2.8	0.997

Table S5. Parameter estimates for values in Table 3. OFW = Open Forest/Woodland; RTW = Ridge Top Woodland.

	ZSM		PLN		NBLN		
	θ	m	μ	σ	μ	σ	ϕ
Closed forest	4.4	0.020	1.2	2.3	3.5	4.3	4.4
Open forest	10.6	0.011	1.6	1.5	2.1	1.1	5.8
OFW	7.7	0.014	1.1	2.4	2.5	5.1	5.1
Woodland	8.6	0.016	0.6	2.4	1.4	5.7	5.4
RTW	10.0	0.017	1.3	1.8	1.7	1.7	3.4