vol. 175, no. 4 the american naturalist april 2010

E-Article

Entropy Maximization and the Spatial Distribution of Species

Bart Haegeman[1] and Rampal S. Etienne[2,*]

1. INRIA (Institut National de Recherche en Informatique et en Automatique) Sophia Antipolis-Mediterranée, Research Team MERE (Modélisation et Ressources en Eau), Unité Mixte de Recherche Systems Analysis and Biometrics, 2 place Pierre Viala, 34060 Montpellier, France; 2. Community and Conservation Ecology Group, Centre for Ecological and Evolutionary Studies, University of Groningen, P.O. Box 14, 9750 AA Haren, The Netherlands

Submitted November 27, 2008; Accepted October 3, 2009; Electronically published February 18, 2010

abstract: Entropy maximization (EM, also known as MaxEnt) is a general inference procedure that originated in statistical mechanics. It has been applied recently to predict ecological patterns, such as species abundance distributions and species-area relationships. It is well known in physics that the EM result strongly depends on how elementary configurations are described. Here we argue that the same issue is also of crucial importance for EM applications in ecology. To illustrate this, we focus on the EM prediction of species-level spatial abundance distributions. We show that the EM outcome depends on (1) the choice of configuration set, (2) the way constraints are imposed, and (3) the scale on which the EM procedure is applied. By varying these choices in the EM model, we obtain a large range of EM predictions. Interestingly, they correspond to spatial abundance distributions that have been derived previously from mechanistic models. We argue that the appropriate choice of the EM model assumptions is nontrivial and can be determined only by comparison with empirical data.

Keywords: spatial abundance distribution, scale transformation, prior distribution, random-placement model, broken stick model, HEAP model.

Introduction

Entropy maximization (EM) is an inference technique that originated in statistical mechanics (Jaynes 1957, 2003). The philosophy behind EM inference is to provide the probability distribution (which we denote $P(\mathbf{x})$) over system configurations (which we denote $\mathbf{x}$) that corresponds best to the available information. Because a probability distribution with higher entropy encodes less information, the probability distribution that corresponds best to the available information, formulated in terms of constraints, can be found by maximizing the entropy subject to these constraints.

Recently, several studies have explored EM inference in ecological problems (Shipley et al. 2006; Banavar and Maritan 2007; Pueyo et al. 2007; Dewar and Porté 2008; Haegeman and Loreau 2008; Harte et al. 2008). Most attention has been paid to EM inference of species abundance distributions (Banavar and Maritan 2007; Pueyo et al. 2007; Dewar and Porté 2008), but Harte et al. (2008) provide an exception: they apply EM to simultaneously predict, with a minimal number of assumptions (constraints), several macroecological patterns, such as species abundance distributions and species-level spatial abundance distributions, that together give species-area relationships. In this article, we zoom in on EM inference of spatial abundance distributions.

We treat the spatial abundance distribution of a given species in a simple, spatially implicit manner. We divide a spatial region in M cells and then describe the arrangement of N individuals over these cells. This description is spatially implicit because we do not take into account the correlations that might exist between neighboring cells. As a consequence, all abundance distributions that we consider are unchanged with random permutation of the cells. Note that this is also the approach taken in different mechanistic models (Coleman 1981; Harte et al. 2005, 2008; Conlisk et al. 2007).

However, this description of spatial abundance distribution by itself does not suffice to apply the EM algorithm. The complete specification of an EM problem requires a number of additional assumptions. These assumptions might appear incidental on first sight, but we show here that they have a major effect on the EM prediction: it turns out that there is not a single EM prediction for the spatial abundance distribution but a myriad of distributions that are obtained under various assumptions, none of which seems to stand out as the most plausible. Interestingly, most of these distributions were obtained as the outcome of process-based models, including the discrete broken stick model (MacArthur 1960), the random-placement model (Coleman 1981), and the single-division model (Conlisk et al. 2007).

To introduce the general framework of the EM approach, we distinguish the formulation of the EM problem

* Corresponding author; e-mail: r.s.etienne@rug.nl.

Am. Nat. 2010. Vol. 175, pp. E74–E90.
DOI: 10.1086/650718

from its solution. Whereas the solution of an EM problem can be found with a purely technical recipe, the formulation of the EM problem requires a number of assumptions that can have an important effect on the solution. We first present different EM assumptions for the prediction of the abundance distribution over spatial cells. The corresponding EM problems are solved systematically in the subsequent section.

Formulating the EM Problem for Spatial Distributions

Formulating the EM problem consists of three steps (Haegeman and Loreau 2009): specifying (1) system configurations, (2) the prior distribution over the system configurations, and (3) the constraints on the system configurations. We will discuss these three steps in order.

Specifying the System Configurations

An EM problem formulation starts with specification of the system configurations. For the case of spatial abundance distribution, there are two obvious and simple ways to do so. The first is to specify for every individual to which cell it belongs. Such a system configuration is denoted by a vector $\mathbf{m}$: its nth component, m_n, gives the cell to which the nth individual belongs. The number of components of $\mathbf{m}$ gives the number of individuals,

$$N(\mathbf{m}) = \text{number of components of } \mathbf{m}. \tag{1}$$

The second way to specify the system configuration is to specify for every cell the number of individuals it contains. Such a system configurations is denoted by a vector $\mathbf{n}$: its mth component, n_m, gives the number of individuals belonging to cell m. The number of components of $\mathbf{n}$ equals the number of cells M. The number of individuals in configuration $\mathbf{n}$ is

$$N(\mathbf{n}) = \sum_{m=1}^{M} n_m. \tag{2}$$

We consider both configurations $\mathbf{m}$ and $\mathbf{n}$, and we will show that the seemingly innocent choice between them can lead to completely different EM predictions. We will call them labeled and unlabeled configurations, respectively, because configurations $\mathbf{m}$ presuppose that individuals are labeled (we know for each individual to which cell it belongs), whereas configurations $\mathbf{n}$ do not require labels for individuals (we merely count the number of individuals in each cell; their identities are lost).

We illustrate the difference between labeled and unlabeled system configurations for a region with $M = 2$ cells and $N = 3$ individuals. An example of a labeled configuration is the vector $\mathbf{m} = (1, 1, 2)$, meaning that the first individual belongs to cell 1, the second individual belongs to cell 1, and the third individual belongs to cell 2. There are eight labeled configurations with $M = 2$ and $N = 3$: (1, 1, 1), (1, 1, 2), (1, 2, 1), (2, 1, 1), (1, 2, 2), (2, 1, 2), (2, 2, 1), and (2, 2, 2).

An example of an unlabeled configuration is the vector $\mathbf{n} = (2, 1)$, meaning that the first cell contains two individuals, and the second cell contains one individual. There are four unlabeled configurations with $M = 2$ and $N = 3$: (3, 0), (2, 1), (1, 2), and (0, 3).

Clearly, there are more labeled than unlabeled configurations. In fact, every labeled configuration corresponds to exactly one unlabeled configuration, but a given unlabeled configuration can correspond to several labeled configurations. In this example, the link between the two types of configurations is

$$\underbrace{(1,1,1)}_{(3,0)} \quad \underbrace{(1,1,2) \quad (1,2,1) \quad (2,1,1)}_{(2,1)}$$

$$\underbrace{(1,2,2) \quad (2,1,2) \quad (2,2,1)}_{(1,2)} \quad \underbrace{(2,2,2)}_{(0,3)}.$$

Note that not all unlabeled configurations correspond to the same number of labeled configurations: vector $\mathbf{n} = (3, 0)$ has one labeled configuration, whereas vector $\mathbf{n} = (2, 1)$ has three. In general, the number $\mathcal{M}(\mathbf{n})$ of labeled configurations that correspond to a given unlabeled configuration $\mathbf{n}$ is given by a multinomial coefficient,

$$\mathcal{M}(\mathbf{n}) = \binom{N}{\mathbf{n}} = \frac{N!}{n_1! n_2! \ldots n_M!}. \tag{3}$$

Specifying the Prior Distribution over the System Configurations

Next, one must specify a prior distribution on the set of system configurations. This prior distribution corresponds to the probabilities one would attribute to the configurations if no constraints were imposed on the system. Typically, an uninformative prior is chosen, giving equal probability to all configurations. We denote the prior distribution $P_0(\mathbf{x})$, where $\mathbf{x}$ represents the configuration (either labeled, $\mathbf{x} = \mathbf{m}$, or unlabeled, $\mathbf{x} = \mathbf{n}$). For an uninformative prior, the distribution $P_0(\mathbf{x})$ is simply a constant.

We could, however, choose any distribution for the prior. A particular choice for the labeled configurations $\mathbf{m}$ could be

$$P_0(\mathbf{m}) = \frac{1}{\mathcal{M}(\mathbf{n}(\mathbf{m}))}, \tag{4}$$

where $\mathbf{n}(\mathbf{m})$ is the unlabeled configuration that corresponds to the labeled configuration $\mathbf{m}$. This prior, defined on system configurations $\mathbf{m}$, makes all unlabeled configurations $\mathbf{n}$ equally probable. We thus observe that specifying the system configuration and specifying the prior are in some sense interchangeable.

Specifying the Constraints on the System Configurations

Finally, we have to specify the constraints we want to take into account in the EM problem. We consider two types of constraints: hard and soft constraints. A hard constraint restricts the set of system configurations to a particular subset, thus ruling out configurations that fall outside this subset. In other words, all configurations not satisfying the constraint have zero probability. For spatial distributions, one could consider only configurations with a specified number of individuals N. This can be formulated as

$$N(\mathbf{x}) = N, \tag{5}$$

with the function $N(\mathbf{m})$ for labeled configurations given in equation (1) and the function $N(\mathbf{n})$ for unlabeled configurations given in equation (2).

A soft constraint does not restrict the system configurations but acts on statistics of the system configurations. For spatial distributions, one could impose that the mean number of individuals in the EM solution has a specified number of individuals N. This can be formulated as

$$\sum_{\mathbf{x}} P(\mathbf{x})N(\mathbf{x}) = N, \tag{6}$$

where $P(\mathbf{x})$ is the EM probability distribution we are trying to solve for and $N(\mathbf{x})$ is given by equation (1) for labeled configurations $\mathbf{m}$ and equation (2) for unlabeled configurations $\mathbf{n}$. Thus, a soft constraint does not completely rule out some configurations but effectively assigns differential nonzero probabilities to them.

General Recipe for Solving the EM Problem

Once the set of system configurations, the prior distribution, and the hard and/or soft constraints have been specified, the EM problem can readily be solved. The solution consists of finding the probability distribution $P(\mathbf{x})$ that maximizes the relative entropy $H(P|P_0)$ subject to the constraints. The relative entropy with respect to the prior distribution $P_0(\mathbf{x})$ is given by

$$H(P|P_0) = -\sum_{\mathbf{x}} P(\mathbf{x}) \ln \frac{P(\mathbf{x})}{P_0(\mathbf{x})}. \tag{7}$$

For an uninformative prior (i.e., $P_0(\mathbf{x})$ independent of $\mathbf{x}$), maximizing relative entropy $H(P|P_0)$ is equivalent to maximizing Shannon entropy $H(P)$,

$$H(P) = -\sum_{\mathbf{x}} P(\mathbf{x}) \ln P(\mathbf{x}). \tag{8}$$

Solution methods for this maximization problem are well known. If all constraints are of the hard type, maximizing (relative) entropy is particularly simple. Configurations that satisfy all constraints have a probability proportional to the prior distribution $P_0(\mathbf{x})$; configurations that do not satisfy all constraints have zero probability. Hence, the EM solution reads

$$P(\mathbf{x}) = \frac{1}{Z} P_0(\mathbf{x}),$$

if $\mathbf{x}$ satisfies all constraints, where Z is a normalization constant given by

$$Z = \sum_{\mathbf{x}} P_0(\mathbf{x}), \tag{9}$$

where the sum is over all vectors $\mathbf{x}$ that satisfy the hard constraint. This guarantees that

$$\sum_{\mathbf{x}} P(\mathbf{x}) = 1.$$

If there are soft constraints, one can use the technique of Lagrange multipliers. For the soft constraint (6), the EM solution can be written in terms of a corresponding Lagrange multiplier α,

$$P(\mathbf{x}) = \frac{1}{Z} P_0(\mathbf{x}) e^{-\alpha N(\mathbf{x})},$$

if $\mathbf{x}$ satisfies all hard constraints, with the normalization constant given by

$$Z = \sum_{\mathbf{x}} P_0(\mathbf{x}) e^{-\alpha N(\mathbf{x})},$$

where the sum runs over vectors $\mathbf{x}$ that satisfy all hard constraints. The Lagrange multiplier α must be determined by imposing the soft constraint (6). The latter constraint can be rewritten as

$$N = -\frac{\partial}{\partial \alpha} \ln Z,$$

which often allows one to solve explicitly for α.

It should be noted that the same EM problem can be written in different ways: different combinations of system configurations, prior distribution, and constraints can yield equivalent EM problems. A first example concerns exchangeability of constraints and system configurations: instead of imposing a hard constraint on the set of system configurations, one could equivalently start out with a smaller set of system configurations. A second, less trivial, example concerns the exchangeability of the prior distribution P_0 and the specification of the system configurations, to which we already alluded in the previous section: the EM problem formulated in terms of labeled configurations $\mathbf{m}$ with an uninformative prior $P_0(\mathbf{m}) \propto 1$ is equivalent to the EM problem formulated in terms of unlabeled configurations $\mathbf{n}$ with an informative prior $P_0(\mathbf{n}) \propto \mathcal{M}(\mathbf{n})$. Thus, an alternative definition of the set of system configurations can be mimicked by introducing an appropriate prior distribution.

The latter fact is particularly important for the next section. It implies that the EM problem with labeled configurations and an uninformative prior is not equivalent to the EM problem with unlabeled configurations and an uninformative prior. The ratio between the two EM solutions is given by the multiplicity factor (eq. [3]). To see that this factor drastically modifies the EM distribution, consider the example of $N = 6$ individuals distributed over $M = 3$ cells. The multiplicity factors for the most and least even distributions, $\mathbf{n}_1 = (2, 2, 2)$ and $\mathbf{n}_2 = (6, 0, 0)$, respectively, are $\mathcal{M}(\mathbf{n}_1) = 90$ and $\mathcal{M}(\mathbf{n}_2) = 1$. In other words, whereas the EM problem in terms of unlabeled configurations assigns the same prior probability to $\mathbf{n}_1$ and $\mathbf{n}_2$, the EM problem in terms of labeled configurations assumes that the even configuration $\mathbf{n}_1$ is a priori 90 times as probable as the clustered configuration $\mathbf{n}_2$. Therefore, these two EM problems will lead to very different predictions (the EM prediction for labeled configurations will give relatively more weight to evenly distributed configurations than will the EM prediction for unlabeled configurations). Obviously, these differences become even more pronounced for larger M and N.

EM Solutions for Spatial Distributions

In this section, we study all four EM problems for spatial distributions that result from different combinations of (1) working with either labeled or unlabeled configurations and (2) using either hard or soft constraints to impose a total number of individuals. In all cases we assume an uninformative prior on the system configuration. Here we summarize the results and discuss similarities and differences between different EM solutions; we refer to the appendixes for the formal derivations.

Labeled Configurations with a Hard Constraint

With an uninformative prior on labeled configurations, all vectors $\mathbf{m}$ have equal probability. The EM problem for labeled configurations with a hard constraint is solved in appendix A. The resulting probability distribution for unlabeled configurations $\mathbf{n}$ is (see eq. [A3])

$$P_{M,N}^{(\text{lab,hard})}(\mathbf{n}) = \binom{N}{\mathbf{n}} \frac{1}{M^N}, \tag{10}$$

where $P^{(\text{lab, hard})}$ denotes the probability distribution that results from applying the EM procedure for labeled configurations with a hard constraint.

Equation (10) is a joint distribution for the abundances of all cells, which we call a "multicell abundance distribution." For this EM problem, the multicell abundance distribution is multinomial: all individuals are placed independently in one of the M cells, and every cell has the same probability $1/M$ that a given individual is assigned to that cell. This is the spatial abundance distribution for the random-placement (RP) model (Coleman 1981).

From equation (10) we can compute the marginal distribution for the abundance of any one cell, which we call the "one-cell abundance distribution." It is given by

$$P_{M,N}^{(\text{lab,hard})}(n_1) = \binom{N}{n_1} \frac{1}{M^{n_1}} \left(1 - \frac{1}{M}\right)^{N-n_1} \tag{11}$$

for $n_1 \leq N$, which is a binomial distribution.

Labeled Configurations with a Soft Constraint

The EM problem for labeled configurations with a soft constraint is solved in appendix A. The resulting probability distribution for unlabeled configurations $\mathbf{n}$ is (see eq. [A6])

$$P_{M,N}^{(\text{lab,soft})}(\mathbf{n}) = \binom{N(\mathbf{n})}{\mathbf{n}} \frac{1}{N+1} \left[\frac{N}{M(N+1)}\right]^{N(\mathbf{n})}, \tag{12}$$

where $P^{(\text{lab, soft})}$ denotes the probability distribution that results from applying the EM procedure for labeled configurations with a soft constraint.

The one-cell abundance distribution is the marginal of the multicell abundance distribution (12) and is given by

$$P_{M,N}^{(\text{lab,soft})}(n_1) = \frac{M}{N+M} \left(\frac{N}{N+M}\right)^{n_1}, \tag{13}$$

which is a geometric distribution with mean N/M. The

distribution for the total number of individuals $N(\mathbf{n})$ is given by

$$P_{M,N}^{(\text{lab, soft})}(N(\mathbf{n})) = \frac{1}{N+1}\left(\frac{N}{N+1}\right)^{N(\mathbf{n})}, \tag{14}$$

which is a geometric distribution with mean N (note that the soft constraint requires that the mean equals N). The link with the hard-constraint solution (10) can be made by conditioning on the total number of individuals,

$$P_{M,N}^{(\text{lab, soft})}(\mathbf{n}|N(\mathbf{n}) = K) = P_{M,K}^{(\text{lab, hard})}(\mathbf{n}). \tag{15}$$

Unlabeled Configurations with a Hard Constraint

With an uninformative prior on unlabeled configurations, all vectors $\mathbf{n}$ have equal probability. The EM problem for unlabeled configurations with hard constraint is solved in appendix B. The resulting probability distribution for unlabeled configurations $\mathbf{n}$ is (see eq. [B2])

$$P_{M,N}^{(\text{unl, hard})}(\mathbf{n}) = 1\Big/\binom{N+M-1}{N}, \tag{16}$$

where $P^{(\text{unl, hard})}$ denotes the probability distribution that results from applying the EM procedure for unlabeled configurations with a hard constraint.

The multicell abundance distribution (16) gives, by construction, equal weight to all unlabeled configurations. Specifying an unlabeled configuration for M cells and N individuals is equivalent to splitting a community of N individuals into M parts. The idea that all such splits are equally probable is reminiscent of the discrete broken-stick (DBS) model (MacArthur 1960; Etienne and Olff 2005) for the distribution of species' abundances. Distribution (16) can be interpreted as the spatial counterpart of the DBS species abundance distribution, with one main difference: whereas in the species abundance distribution each species has at least one individual, in the spatial abundance distribution cells may be empty. From distribution (16) we can compute the one-cell abundance distribution (see eq. [B3])

$$P_{M,N}^{(\text{unl, hard})}(n_1) = \binom{N-n_1+M-2}{N-n_1}\Big/\binom{N+M-1}{N} \tag{17}$$

for $n_1 \leq N$.

Unlabeled Configurations with a Soft Constraint

The EM problem for labeled configurations with a soft constraint is solved in appendix B. The resulting probability distribution for unlabeled configurations $\mathbf{n}$ is (see eq. [B5])

$$P_{M,N}^{(\text{unl, soft})}(\mathbf{n}) = \prod_{m=1}^{M}\left(\frac{M}{N+M}\left(\frac{N}{N+M}\right)^{n_m}\right), \tag{18}$$

where $P^{(\text{unl, soft})}$ denotes the probability distribution that results from applying the EM procedure for unlabeled configurations with soft constraint.

The multicell abundance distribution (18) has a simple structure: it is the product of independent one-cell abundance distributions, each of which is given by

$$P_{M,N}^{(\text{unl, soft})}(n_1) = \frac{M}{N+M}\left(\frac{N}{N+M}\right)^{n_1}. \tag{19}$$

This is a geometric distribution with mean N/M. The distribution for the total number of individuals $N(\mathbf{n})$ is given by

$$P_{M,N}^{(\text{unl, soft})}(N(\mathbf{n})) = \binom{N(\mathbf{n})+M-1}{N(\mathbf{n})}\left(\frac{M}{N+M}\right)^{M}\left(\frac{N}{N+M}\right)^{N(\mathbf{n})}, \tag{20}$$

which is a negative binomial distribution. The link with the hard-constraint solution (16) can be made by conditioning on the total number of individuals,

$$P_{M,N}^{(\text{unl, soft})}(\mathbf{n}|N(\mathbf{n}) = K) = P_{M,K}^{(\text{unl, hard})}(\mathbf{n}). \tag{21}$$

Link between Hard- and Soft-Constraint Solutions

For both labeled and unlabeled configurations, the EM problems with hard and soft constraints on the number of individuals yield related results. The relationship is given in equations (15) and (21): the soft-constraint solution conditioned on the total number of individuals $N(\mathbf{n}) = K$ equals the hard-constraint solution for the total number of individuals K. This conditioning property is generally valid for EM solutions.

The difference between hard- and soft-constraint EM solutions resides in their distribution for the total number of individuals $N(\mathbf{n})$. For the hard constraint, the distribution for $N(\mathbf{n})$ is concentrated at the constraint N. For the soft constraint, we know that the distribution for $N(\mathbf{n})$ has its mean at the constraint N, by construction. If the

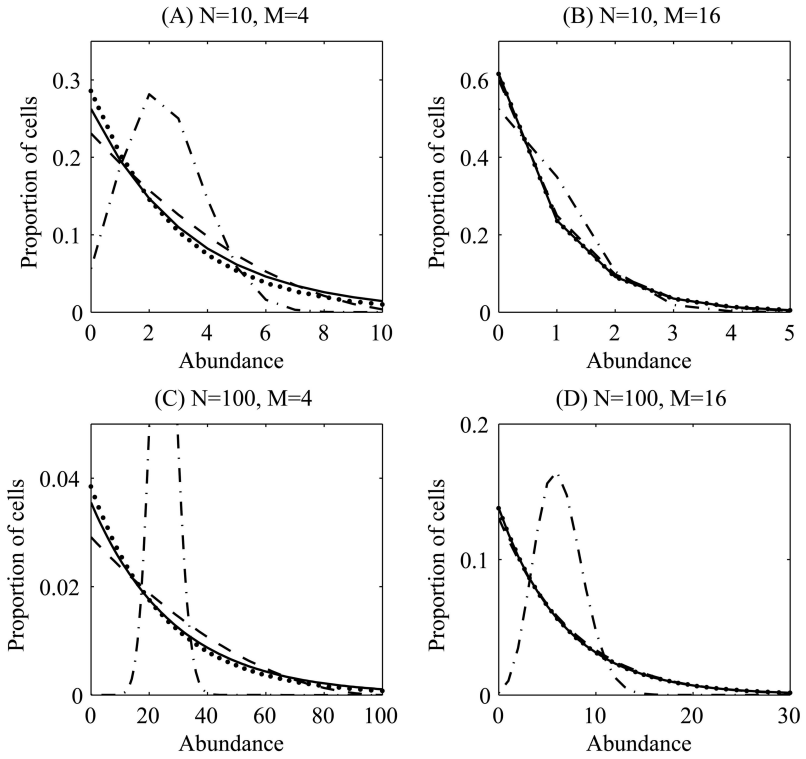

Figure 1: Comparison between different entropy maximization (EM) predictions for the one-cell abundance distribution: the EM solution (eq. [11]) for labeled configurations with hard constraint (*dash-dotted line*), the EM solution (eq. [13]) for labeled configurations with soft constraint (*dotted line*), the EM solution (eq. [17]) for unlabeled configurations with hard constraint (*dashed line*), and the EM solution (eq. [19]) for unlabeled configurations with soft constraint (*dotted line*; one-cell abundance distributions $P^{(\text{lab, soft})}$ and $P^{(\text{unl, soft})}$ coincide). Also plotted is the EM solution computed by Harte et al. (2008; *solid line*). The panels show these distributions for different values of M and N. *A*, $M = 4$ and $N = 10$; *B*, $M = 16$ and $N = 10$; *C*, $M = 4$ and $N = 100$; *D*, $M = 16$ and $N = 100$. The solutions for unlabeled configurations are very close for all values of M and N and are visually indistinguishable for $M = 16$. The solution for labeled configurations with hard constraint is very different.

variation around the mean is small, that is, if the soft-constraint distribution for $N(\mathbf{n})$ is sharply peaked at N, then the EM solutions for hard and soft constraints are practically equivalent.

The EM solutions for labeled and unlabeled configurations behave quite differently in this respect. For labeled configurations, the distribution $P_{M,N}^{(\text{lab,soft})}(\mathbf{n})$ is geometric (see eq. [14]), with a large variation around the mean N. This can be verified by computing the coefficient of variation,

$$\text{CV} = \sqrt{\frac{N+1}{N}},$$

which is greater than 1 for all N and M. The EM distributions $P^{(\text{lab, hard})}$ and $P^{(\text{lab, soft})}$ are therefore quite different.

For unlabeled configurations, the distribution $P_{M,N}^{(\text{unl,soft})}(N(\mathbf{n}))$ is a negative binomial (see eq. [20]) and sharply peaked at N. Indeed,

$$\text{CV} = \sqrt{\frac{N+M}{NM}}.$$

In most cases of interest, M and N are large (say, $M > 10$ and $N > 10$), and the coefficient of variation is much less than 1. Hence, the EM distributions $P^{(\text{unl, hard})}$ and $P^{(\text{unl, soft})}$ can be considered equivalent.

These conclusions are illustrated in figure 1, which compares the one-cell abundance distributions for the four EM distributions we have analyzed: $P^{(\text{lab,hard})}$, $P^{(\text{lab,soft})}$, $P^{(\text{unl,hard})}$, and $P^{(\text{unl,soft})}$. Note that the one-cell abundance distributions $P^{(\text{lab,soft})}$ and $P^{(\text{unl,soft})}$ are mathematically identical; see equations (13) and (19). Formula (17) for $P^{(\text{unl, hard})}$ is not identical to these, but its curve almost always coincides with the two soft-constraint solutions. However, distribution $P^{(\text{lab, hard})}$ has a completely different one-cell abundance distribution; see equation (11).

Scale Dependence of EM Solutions

We have shown that several spatial distributions can be obtained from the EM algorithm with different assumptions in the formulation of the EM problem. In this section, we consider the scale on which the EM algorithm is applied. Indeed, the scale of the EM problem, measured by the number of cells M, requires close scrutiny. We investigate whether the outcome of an EM computation depends on the scale on which the problem was formulated. This is particularly important when combining EM distributions on different scales, for example, to compute species-area relationships (Harte et al. 2008). Distributions on different scales should be combined only if they are consistent. We demonstrate that this condition is not necessarily satisfied by EM distributions.

For a region of fixed size, the number of cells into which the region is partitioned determines the scale of the problem. The larger the number of cells, the finer the scale. We consider two different scales, M_1 and M_2, and we assume that M_1 is the finer scale and is related to the coarser scale M_2 by an integer factor $\ell = M_1/M_2$. In other words, a cell on scale M_2 consists of ℓ cells on scale M_1.

We introduce a scale transformation from the fine scale M_1 to the coarse scale M_2. For any configuration on scale M_1, there is a corresponding configuration on scale M_2. However, there are several configurations on scale M_2 that correspond to a given configuration on scale M_1. To find the probability of a configuration on scale M_2, we sum the probabilities of all configurations on scale M_1 compatible with the configuration on scale M_2.

To illustrate the scale transformation, consider unlabeled configurations with $N = 2$ individuals. The fine scale has $M_1 = 4$ cells; the coarse scale has $M_2 = 2$ cells. The scale transformation regroups the first two cells on scale $M_1 = 4$ in the first cell on scale $M_2 = 2$, and the last two cells on scale $M_1 = 4$ in the second cell on scale $M_1 = 4$. This leads to the following correspondence:

$$\underbrace{(2,0,0,0)\quad (1,1,0,0)\quad (0,2,0,0)}_{(2,0)}$$

$$\underbrace{(0,0,0,2)\quad (0,0,1,1)\quad (0,0,2,0)}_{(0,2)}$$

$$\underbrace{(1,0,1,0)\quad (0,1,1,0)\quad (1,0,0,1)\quad (0,1,0,1).}_{(1,1)}$$

Scale consistency can then be defined as follows. Apply the EM algorithm on fine scale M_1, and use the scale transformation from M_1 to M_2 to obtain a spatial distribution on coarse scale M_2. If the latter distribution corresponds to the distribution obtained by applying the EM algorithm directly on scale M_2 then the EM distributions are called "scale consistent."

In appendix C, we show that for EM problems stated in terms of labeled configurations, the resulting spatial abundance distributions are scale consistent. However, for EM problems stated in terms of unlabeled configurations, the distributions are not scale consistent. As a consequence, a new set of EM distributions can be obtained by, first, applying the EM procedure for unlabeled configurations on scale M_1 and, second, computing averages of the EM solution to obtain a consistent probability distribution for configurations on a coarser scale M_2. We again distinguish hard and soft constraints for the number of individuals.

Averaged Solution for Unlabeled Configurations with Hard Constraint

The EM problem with averaging and hard constraint is solved in appendix C. The resulting distribution for unlabeled configurations $\mathbf{n}$ is (see eq. [C2])

$$P^{(\text{avg, hard})}_{M,N,\ell}(\mathbf{n}) = \left[1\Big/\binom{N+\ell M-1}{N}\right]\prod_{m=1}^{M}\binom{n_m+\ell-1}{n_m}, \quad (22)$$

where $P^{(\text{avg, hard})}$ denotes the probability distribution that results (on scale M) from applying the EM procedure (on scale ℓM) with averaging (scale factor ℓ) and hard constraint.

The multicell abundance distribution (22) has been used previously to model spatial abundance distribution (Conlisk et al. 2007). It arises from the so-called single-division (SD) model based on certain colonization rules of individuals into cells. On a more abstract level, it is related to the Pólya-Eggenberger urn scheme (Johnson et al. 1997). Distribution (22) has the marginal one-cell abundance distribution

$$P^{(\text{avg, hard})}_{M,N,\ell}(n_1) = \left[\binom{n_1+\ell-1}{n_1}\binom{N-n_1+\ell M-\ell-1}{N-n_1}\Big/\binom{N+\ell M-1}{N}\right], \quad (23)$$

which is a negative hypergeometric distribution.

Note that EM distributions $P^{(\text{lab, hard})}$, $P^{(\text{lab, soft})}$, $P^{(\text{unl, hard})}$, and $P^{(\text{unl, soft})}$ are uniquely determined by the number of individuals N and the number of cells M. In contrast, the distribution $P^{(\text{avg, hard})}$ has one additional parameter, namely, the factor ℓ of the scale transformation.

Averaged Solution for Unlabeled Configurations with Soft Constraint

The EM problem with averaging and soft constraint is solved in appendix C. The resulting distribution for unlabeled configurations **n** is (see eq. [C4])

$$P^{(\mathrm{avg,\,soft})}_{M,N,\ell}(\mathbf{n}) = \prod_{m=1}^{M} \binom{n_m + \ell - 1}{n_m} \left(\frac{\ell M}{N + \ell M}\right)^{\ell} \left(\frac{N}{N + \ell M}\right)^{n_m}, \tag{24}$$

where $P^{(\mathrm{avg,\,soft})}$ denotes the probability distribution that results (on scale M) from applying the EM procedure (on scale ℓM) with averaging (scale factor ℓ), and with soft constraint.

The multicell abundance distribution (24) has a simple structure: cell abundances are independent, and all have the same one-cell abundance distribution,

$$P^{(\mathrm{avg,\,soft})}_{M,N,\ell}(n_1) = \binom{n_1 + \ell - 1}{n_1} \left(\frac{\ell M}{N + \ell M}\right)^{\ell} \left(\frac{N}{N + \ell M}\right)^{n_1}, \tag{25}$$

which is a negative binomial distribution. The link with the hard-constraint solution (22) can be made by conditioning on the total number of individuals,

$$P^{(\mathrm{avg,\,soft})}_{M,N,\ell}(\mathbf{n} | N(\mathbf{n}) = K) = P^{(\mathrm{avg,\,hard})}_{M,K,\ell}(\mathbf{n}). \tag{26}$$

Hence, the SD model, given by equation (22), can be interpreted as a product of negative binomial distributions conditioned on the total number of individuals (Conlisk et al. 2007).

Link between Averaged Solutions with Hard and Soft Constraints

Using an argument analogous to that for unlabeled configurations, one can show that the averaged EM solutions with hard and soft constraints are practically equivalent. First, we note that distributions (22) and (24) have the same distribution, conditional on the number of individuals; see equation (26). Second, the soft-constraint distribution for the number of individuals is sharply peaked at the constraint N. Indeed, the coefficient of variation,

$$\mathrm{CV} = \sqrt{\frac{N + \ell M}{N \ell M}},$$

is much less than 1 if ℓM and N are large, a condition that is satisfied in most cases of interest.

Link between Solution for Labeled Configurations and Averaged Solution for Unlabeled Configurations

The averaged EM solutions are constructed from the solution of the EM problems with unlabeled configurations. One can verify that EM distributions $P^{(\mathrm{unl,\,hard})}$ and $P^{(\mathrm{unl,\,soft})}$ are recovered from $P^{(\mathrm{avg,\,hard})}$ and $P^{(\mathrm{avg,\,soft})}$ by setting $\ell = 1$. Here we establish a link between the averaged EM solutions and the solution of the EM problems with labeled configurations.

To do so, we consider the limit $\ell \to \infty$. It is shown in appendix C that (see eq. [C7])

$$\lim_{\ell\to\infty} P^{(\mathrm{avg,hard})}_{M,N,\ell}(\mathbf{n}) = P^{(\mathrm{lab,hard})}_{M,N}(\mathbf{n}).$$

However, we also find that (see eq. [C8]),

$$\lim_{\ell\to\infty} P^{(\mathrm{avg,soft})}_{M,N,\ell}(\mathbf{n}) \neq P^{(\mathrm{lab,soft})}_{M,N}(\mathbf{n}).$$

The reason for this asymmetry is that the distributions are the same, depending on the number of individuals (see eq. [C10]), but their distribution for the number of individuals is different. For the first, $\lim_{\ell\to\infty} P^{(\mathrm{avg,\,soft})}_{M,N,\ell}$, the number of individuals has a Poisson distribution (eq. [C9]); for the second, $P^{(\mathrm{lab,\,soft})}_{M,N}(\mathbf{n})$, the number of individuals has a geometric distribution (eq. [13]).

We conclude that the family of averaged EM distributions ($P^{(\mathrm{avg,hard})}_{M,N,\ell}(\mathbf{n})$ and $P^{(\mathrm{avg,soft})}_{M,N,\ell}(\mathbf{n})$), parameterized by the scale factor ℓ, comprises many of the other EM solutions. For $\ell = 1$, we recover the EM solution for labeled configurations with both hard and soft constraints. For $\ell \to \infty$, we recover the EM solution for unlabeled configurations with hard constraint but not that with soft constraint. For intermediate values of ℓ, we find interpolating spatial abundance distributions; see figure 2.

Discussion

Entropy maximization (EM) is a mathematical framework that can be used to infer a probability distribution on the set of system configurations, given partial information about the system configuration. Several EM applications in ecology can be imagined and have been studied recently. Here we studied the EM problem for spatial abundance distributions. More precisely, we considered a region divided into a number of cells and derived probability distributions for the arrangement of individuals over the cells without taking into account the spatial location of cells.

We showed that an EM problem formulation requires several assumptions or choices and that the outcome of the EM algorithm depends strongly on these choices. There is not a unique EM prediction for the spatial abundance distribution. On the contrary, we obtained a variety of EM

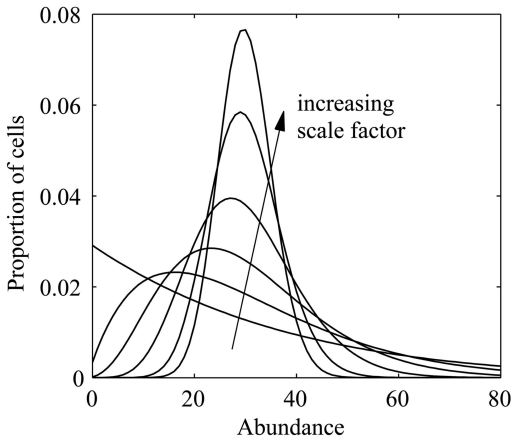

Figure 2: One-cell abundance distributions for different entropy maximization solutions with hard constraint. We consider a species with $N = 300$ individuals and a spatial domain with $M = 10$ cells. The distribution $P^{(\text{avg, hard})}$ is plotted for different scale factors: $\ell = 1$, 2, 4, 10, and 40 and $\ell \to \infty$. Note that $P^{(\text{avg, hard})}$ with $\ell = 1$ corresponds to $P^{(\text{lab, hard})}$ and that $P^{(\text{avg, hard})}$ for $\ell \to \infty$ corresponds to $P^{(\text{lab, hard})}$.

predictions, depending on what might look like details in the EM problem: do we formulate the EM problem in terms of labeled or unlabeled configurations; what prior distribution do we assume; do we impose the number of individuals as a hard or a soft constraint; and on what scale is the EM problem formulated?

The fact that EM allows for a wide range of spatial abundance distributions should not come as a surprise. The EM procedure is an inference technique that depends crucially on the information used in the inference. This information is contained not only in the constraints but also in the way we define system configurations and in the prior distribution over the configurations. Our study indicates that these implicit assumptions should be made explicit in any application of EM procedure, because they can radically change the predicted probability distributions.

An analogous situation exists in physics. Consider a system of N noninteracting particles, each occupying one of M energy levels. This physical system is comparable to our ecological example of distributing N individuals over M spatial cells. Labeled (distinguishable) particles give rise to the classical Maxwell-Boltzmann (MB) distribution, whereas unlabeled (indistinguishable) particles give rise to the quantum mechanical Bose-Einstein (BE) distribution. A third distribution exists, the Fermi-Dirac (FD) distribution, which is also quantum mechanical but has the additional constraint that no more than one particle can be in any one state. In our ecological example, this would mean that one cell cannot contain more than one individual. It is well known that a coarse-grained description of both BE and FD distributions (i.e., by taking together many quantum mechanical energy levels) tends toward the MB distribution. Similarly, our EM solution for unlabeled species at scale M_1 averaged at scale M_2 becomes the EM solution for labeled individuals when $M_1 \to \infty$ (fig. 1).

The equivalence of the EM solutions with hard and soft constraints is a property that is generally satisfied in statistical mechanics (except in phase transitions). The lack of equivalence between $P^{(\text{lab, hard})}$ and $P^{(\text{lab, soft})}$ seems to be pathological and related to a similar problem in statistical mechanics for classical systems. To fix this problem, one must introduce an appropriate prior distribution, the so-called Boltzmann counting. In appendix D, we present an alternative computation for the EM problem in terms of labeled configurations, using as a prior distribution the analog of Boltzmann counting. This yields distributions, $P^{(\text{lab, alt, hard})}$ and $P^{(\text{lab, alt, soft})}$, that are practically equivalent. If we accept the replacement of the pathological distribution $P^{(\text{lab, soft})}$ with $P^{(\text{lab, alt, soft})}$, then all EM distributions derived in this article are part of the family of averaged EM distributions $P^{(\text{avg, hard})}$ and $P^{(\text{avg, soft})}$.

Harte et al. (2008)'s EM application for spatial abundance distributions is different from ours. Their EM problem is written directly in terms of one-cell abundance distributions: their system configuration is simply the abundance in a single cell. They implicitly assume a prior that assigns equal probability to each abundance. Their constraints are (1) that the number of individuals in a cell is smaller than the total number of individuals N in the entire region (a hard constraint because it rules out any configuration with abundance greater than N) and (2) that the mean number of individuals in a cell equals N/M (a soft constraint). The solution for the one-cell abundance distribution is different from but close to our solutions $P^{(\text{lab, soft})}$, $P^{(\text{unl, hard})}$, and $P^{(\text{unl, soft})}$; see figure 1. One may wonder what multicell abundance distribution underlies the one-cell abundance distribution of Harte et al. (2008). In appendix E, we solve the EM problem for the multicell abundance distribution under their constraints 1 and 2, and we show that the corresponding one-cell abundance distribution is identical to that of Harte et al. (2008) but not scale consistent. However, this does not mean that Harte et al. (2008)'s one-cell abundance cannot be embedded in a scale-consistent multicell abundance distribution. Marginal probability distributions do not, in general, completely determine the joint probability distributions, so it is possible that a scale-consistent multicell abundance distribution exists that yields the same one-cell abundance distribution. The unlabeled configurations that we have studied are all invariant under permutations of the cells, thanks to the fact that spatial location is not taken into account. A multicell abundance distribution that is permutation invariant and scale consistent and has

Harte et al. (2008)'s one-cell abundance distribution does not seem to exist. One must incorporate space to find such a distribution. How such a scale-consistent multicell abundance distribution should result from a properly formulated spatial EM problem remains an open problem.

Note that the discussion of multicell versus one-cell abundance distribution has an analogy in neutral theory's predictions for species abundance distributions (Chave et al. 2006), where sampling formulas have been derived that are multispecies abundance distributions (Etienne 2005, 2007), in contrast to one-species abundance distributions (Volkov et al. 2003). Sampling formulas are required for a detailed comparison between theory and observation, because there may be several sampling formulas that are compatible with a single one-species abundance distribution (Chave et al. 2006). Here we have also seen that the same one-cell abundance distribution can correspond to several multicell abundance distributions: compare the joint distributions (18) and (12), which give rise to the identical marginals (19) and (13). Given the central role of data comparison in the EM modeling approach, multicell abundance distributions are therefore required for more powerful EM applications.

It is remarkable that the simple EM applications we have considered yield spatial abundance distributions that have been obtained previously by studying more detailed and often dynamical mechanistic models. We encountered the RP model as the solution of the EM problem with labeled configurations and hard constraint. The DBS model was found as the solution of the EM problem with unlabeled configurations and soft constraint, while scale transforming the latter distribution yields the SD model. The fact that these distributions can be obtained from simple EM applications might indicate a certain robustness. For example, one can expect that a (weakly) perturbed mechanistic model would lead to the same EM distribution.

In fact, even more previously studied spatial abundance distributions can be written as EM solutions. For example, the model based on the hypothesis of equal allocation probabilities (HEAP; Harte et al. 2005) is related to the unlabeled-configurations solution $P^{(\mathrm{unl,\ hard})}$. We have shown that this distribution is not scale consistent, which naturally led us to the averaged distribution $P^{(\mathrm{avg,\ hard})}$. Similarly, the HEAP model can be interpreted as an iterative averaging approach, applying a scale factor $\ell = 2$ in every iteration step. We remark that this iterative approach is not equivalent to the one-step scale transformation on which our distribution $P^{(\mathrm{avg,\ hard})}$ is based.

This suggests that almost any reasonable spatial abundance distribution can be written as the solution of an EM problem. The choice of assumptions in the EM problem formulation is indeed wide. One could consider other ways to define system configurations (different from labeled and unlabeled configurations), one could work with informative prior distributions, or other consistency requirements could be imposed. We believe that the present understanding of the problem of allocating individuals to spatial cells does not allow us to decide which EM problem formulation is most appropriate. This should caution ecologists that applying the EM method does not automatically yield useful results; the accuracy of EM predictions can be determined only by comparison with empirical data.

This illustrates both a strength and a weakness of the EM procedure. Entropy maximization applications are based on a minimal number of assumptions (e.g., labeled or unlabeled individuals, scale consistency), yielding an efficient formalism to generate predictions that can be compared with empirical data. However, the EM problem formulation does not directly establish a link with an underlying mechanistic model. In fact, different process-based models will typically yield similar ecological patterns. Although translating ecological processes into an EM problem formulation (i.e., a set of configurations, a prior distribution, and constraints) can be a nontrivial problem, the EM procedure might develop into a valuable tool to extract from a detailed mechanistic model a minimal set of assumptions that determine the model predictions.

We considered only a subproblem of the spatial distribution of a species in an ecological community. First, we considered only one species at a time, neglecting the effects of other species and their spatial distribution. Second, we considered only the abundance distribution over cells, without taking into account the spatial location of these cells. Stronger correlations can be expected with abundance distributions for nearby cells than with those for distant cells (see Maddux 2004 and Ostling et al. 2004 for a consistency problem related to the spatial structure). In turn, this might influence the predicted species abundance distribution. Whether the EM approach can be usefully applied for several species at once and/or for spatially structured communities is an interesting set of topics for future research.

Acknowledgments

We thank J. Harte, M. Loreau, A. Ostling, T. Zillio, and an anonymous reviewer for fruitful discussions and comments. Financial support for R.S.E. was provided by the Netherlands Organisation for Scientific Research (NWO).

APPENDIX A

EM for Labeled Configurations

We apply the EM algorithm under the assumption that labeled configurations **m** are a priori equally probable. We

maximize entropy subject to a constraint on the total number of individuals. There are two ways to impose this constraint.

Hard Constraint on N

The first possibility is to restrict the set of configurations to vectors **m** with the exact number of individuals N,

$$N(\mathbf{m}) = N. \tag{A1}$$

There are no further constraints to impose, so the EM computation is trivial. Configurations **m** with $N(\mathbf{m}) \neq N$ have zero probability; configurations **m** with $N(\mathbf{m}) = N$ all have equal probability. As there are M^N such configurations, the EM distribution is

$$P_{M,N}^{(\text{lab, hard})}(\mathbf{m}) = \frac{1}{M^N} \tag{A2}$$

if $N(\mathbf{m}) = N$, where $P^{(\text{lab, hard})}$ denotes the probability distribution that results from applying the EM procedure for labeled configurations with "hard" constraint (A1). The probability distribution for labeled configurations **m** can be transformed into a probability distribution for unlabeled configurations **n** by using the multiplicity factor (eq. [3]):

$$P_{M,N}^{(\text{lab,hard})}(\mathbf{n}) = \mathcal{M}(\mathbf{n})\frac{1}{M^N} = \binom{N}{\mathbf{n}}\frac{1}{M^N} \tag{A3}$$

if $N(\mathbf{n}) = N$, which is a multinomial distribution.

Soft Constraint on N

The second possibility is to take all configurations **m** into account, thus including vectors **m** for which $N(\mathbf{m}) \neq N$. We require that the mean number of individuals equals N,

$$\sum_{\mathbf{m}} P(\mathbf{m})N(\mathbf{m}) = N. \tag{A4}$$

We use the technique of Lagrange multipliers to solve the EM problem. We denote the Lagrange multiplier for constraint (A4) α. The EM solution reads

$$P_{M,N}^{(\text{lab,soft})}(\mathbf{m}) = \frac{1}{Z}e^{-\alpha N(\mathbf{m})},$$

where $P^{(\text{lab, soft})}$ denotes the probability distribution that results from applying the EM procedure for labeled configurations with "soft" constraint (30). The normalization constant Z can be calculated as follows:

$$Z = \sum_{\mathbf{m}} e^{-\alpha N(\mathbf{m})} = \sum_{n=0}^{\infty} M^n e^{-\alpha n} = \frac{1}{1 - Me^{-\alpha}}.$$

Imposing constraint (A4) yields

$$N = -\frac{\partial}{\partial \alpha}\ln Z = \frac{Me^{-\alpha}}{1 - Me^{-\alpha}},$$

and we can solve for the Lagrange multiplier α,

$$\alpha = \ln\frac{M(N+1)}{N}.$$

As a result,

$$P_{M,N}^{(\text{lab,soft})}(\mathbf{m}) = \frac{1}{N+1}\left[\frac{N}{M(N+1)}\right]^{N(\mathbf{m})}, \tag{A5}$$

which gives the distribution for labeled configurations,

$$P_{M,N}^{(\text{lab,soft})}(\mathbf{n}) = \binom{N(\mathbf{n})}{\mathbf{n}}\frac{1}{N+1}\left[\frac{N}{M(N+1)}\right]^{N(\mathbf{n})}. \tag{A6}$$

APPENDIX B

EM for Unlabeled Configurations

We apply the EM algorithm under the assumption that unlabeled configurations **n** are a priori equally probable. We maximize entropy subject to the constraints on the total number of individuals. There are two ways to deal with this constraint.

Hard Constraint on N

The hard constraint restricts the set of configurations to vectors **n** with the exact number of individuals N,

$$N(\mathbf{n}) = N. \tag{B1}$$

There are no further constraints to impose, so the EM computation is trivial. Configurations **n** with $N(\mathbf{n}) \neq N$ have zero probability; configurations **n** with $N(\mathbf{n}) = N$ all have equal probability. It is a standard result in combinatorics that there are

$$\binom{N+M-1}{N}$$

configurations of the latter type. Hence,

$$P_{M,N}^{(\mathrm{unl,hard})}(\mathbf{n}) = 1\Big/\binom{N+M-1}{N} \tag{B2}$$

if $N(\mathbf{n}) = N$, where $P^{(\mathrm{unl,\ hard})}$ denotes the probability distribution that results from applying the EM procedure for unlabeled configurations with hard constraint (B1). From equation (16) we compute the one-cell abundance distribution,

$$P_{M,N}^{(\mathrm{unl,hard})}(n_1) = \binom{N-n_1+M-2}{N-n_1}\Big/\binom{N+M-1}{N} \tag{B3}$$

for $n_1 \leq N$, because there are

$$\binom{N-n_1+M-2}{N-n_1}$$

configurations that have n_1 individuals in one particular cell and $N-n_1$ in the remaining $M-1$ cells.

Soft Constraint on N

The soft constraint takes all configurations $\mathbf{n}$ into account and requires that the mean number of individuals equals N,

$$\sum_{\mathbf{n}} P(\mathbf{n})N(\mathbf{n}) = N. \tag{B4}$$

We use the technique of Lagrange multipliers to solve the EM problem. We denote the Lagrange multiplier for constraint (B4) α. The EM solution reads

$$P_{M,N}^{(\mathrm{unl,soft})}(\mathbf{n}) = \frac{1}{Z}e^{-\alpha N(\mathbf{n})},$$

where $P^{(\mathrm{unl,\ soft})}$ denotes the probability distribution that results from applying the EM procedure for unlabeled configurations with soft constraint (B4). The normalization constant Z can be calculated as follows:

$$Z = \sum_{\mathbf{n}} e^{-\alpha N(\mathbf{n})} = \sum_{n=0}^{\infty}\binom{n+M-1}{n}e^{-\alpha n} = \frac{1}{(1-e^{-\alpha})^M}.$$

Imposing constraint (B4) yields

$$N = -\frac{\partial}{\partial\alpha}\ln Z = M\frac{e^{-\alpha}}{1-e^{-\alpha}},$$

and we can solve for the Lagrange multiplier α,

$$\alpha = \ln\left(1+\frac{M}{N}\right).$$

As a result,

$$P_{M,N}^{(\mathrm{unl,soft})}(\mathbf{n}) = \left(\frac{M}{N+M}\right)^{M}\left(\frac{N}{N+M}\right)^{N(\mathbf{n})} = \prod_{m=1}^{M}\left(\frac{M}{N+M}\left(\frac{N}{N+M}\right)^{n_m}\right). \tag{B5}$$

APPENDIX C

Scale-Transformed EM Distributions

In this appendix, we investigate how the EM solutions for labeled and unlabeled configurations change under scale transformation. This transformation maps an abundance distribution on M_1 cells to a distribution on a coarser scale with M_2 cells, $M_2 < M_1$. The scales M_1 and M_2 are related by an integer scale factor $\ell = M_1/M_2$.

EM for Labeled Configurations

Consider the distribution $P_{M_1,N}^{(\mathrm{lab,hard})}$ on the fine scale M_1. It corresponds to randomly allocating N individuals to M_1 cells, each cell having probability $1/M_1$ of receiving an individual. To scale transform this distribution, we have to take ℓ cells together, so that individuals are now randomly allocated to $M_1/\ell = M_2$ regrouped cells, each regrouped cell having probability $\ell/M_1 = 1/M_2$. This is the distribution $P_{M_2,N}^{(\mathrm{lab,hard})}$ on the coarse scale M_2.

This result can be used to scale transform the distribution $P_{M_1,N}^{(\mathrm{lab,soft})}$. The latter distribution can be written as a combination of $P_{M_1,K}^{(\mathrm{lab,hard})}$ for different K. As we have shown above, each of these components is scale transformed to

$P^{(\mathrm{lab,hard})}_{M_2,K}$. Moreover, the coefficients of this combination, given by equation (14), do not depend on the scale M_1 or M_2. Hence, the scale transformation of $P^{(\mathrm{lab,soft})}_{M_1,N}$ is a combination of $P^{(\mathrm{lab,hard})}_{M_2,K}$ for different K, with coefficients also given by equation (14). This leads to the distribution $P^{(\mathrm{lab,soft})}_{M_2,N}$ on scale M_2.

EM for Unlabeled Configurations

Consider first the distribution $P^{(\mathrm{unl,hard})}_{M_1,N}$ on the fine scale M_1. It attributes the same probability to all unlabeled configurations. Hence, the scale-transformed (on the coarse scale M_2) probability for a configuration $\mathbf{n}$ is proportional to the number of configurations on scale M_2 that are compatible with $\mathbf{n}$. This number is given by

$$\mathcal{S}(\mathbf{n}) = \prod_{m=1}^{M_2} \binom{n_m + \ell - 1}{n_m}, \tag{C1}$$

because there are

$$\binom{n + \ell - 1}{n}$$

configurations that satisfy the condition $\sum_{i=1}^{\ell} n_i = n$. The probability distribution on scale M_2 follows directly from the multiplicity factor (C1),

$$\begin{aligned} P^{(\mathrm{avg,hard})}_{M_2,N,\ell}(\mathbf{n}) &= \left[1 \Big/ \binom{N + M_1 - 1}{N}\right] \mathcal{S}(\mathbf{n}) \\ &= \left[1 \Big/ \binom{N + \ell M_2 - 1}{N}\right] \prod_{m=1}^{M_2} \binom{n_m + \ell - 1}{n_m}, \end{aligned} \tag{C2}$$

where $P^{(\mathrm{avg,\ hard})}$ denotes the probability distribution that results from (1) applying the EM procedure for unlabeled configurations with hard constraint (B1) on a fine scale and (2) scale transforming the EM solution to a coarser scale.

Next, we compute the scale transformation of the distribution $P^{(\mathrm{unl,soft})}_{M_1,N}$. This multicell abundance distribution equals the product of M_1 independent one-cell (on scale M_1) abundance distributions (19). The scale transformation consists of regrouping cells on scale M_1 into one cell on scale M_2. Hence, the scale-transformed multicell abundance distribution equals the product of M_2 independent one-cell (on scale M_2) abundance distributions. Each factor in this product is given by the convolution of abundance distributions (eq. [19]), leading to a negative binomial distribution,

$$\begin{aligned} P^{(\mathrm{avg,soft})}_{M_2,N,\ell}(n_1) &= \binom{n_1 + \ell - 1}{n_1} \left(\frac{M_1}{N + M_1}\right)^{\ell} \left(\frac{N}{N + M_1}\right)^{n_1} \\ &= \binom{n_1 + \ell - 1}{n_1} \left(\frac{\ell M_2}{N + \ell M_2}\right)^{\ell} \left(\frac{N}{N + \ell M_2}\right)^{n_1}. \end{aligned} \tag{C3}$$

For the multicell abundance distribution, we obtain

$$P^{(\mathrm{avg,soft})}_{M_2,N,\ell}(\mathbf{n}) = \prod_{m=1}^{M_2} \binom{n_m + \ell - 1}{n_m} \left(\frac{\ell M_2}{N + \ell M_2}\right)^{\ell} \left(\frac{N}{N + \ell M_2}\right)^{n_m}. \tag{C4}$$

Our analysis here is concerned with scaling up from a fine scale M_1 to a coarse scale M_2, corresponding to integer scale factors $\ell = 2, 3, \ldots$. The opposite is equally possible: scaling down from a coarse scale M_1 to a fine scale M_2, corresponding to scale factors

$$\ell = \frac{1}{2}, \frac{1}{3}, \ldots.$$

This case requires a generalization of previous formulas for noninteger ℓ. Equation (C2) for the hard constraint becomes

$$P^{(\mathrm{avg,hard})}_{M_2,N,\ell}(\mathbf{n}) = \frac{N!\Gamma(\ell M_2)}{\Gamma(N+\ell M_2)}\prod_{m=1}^{M_2}\frac{\Gamma(n_m+\ell)}{n_m!\Gamma(\ell)}. \tag{C5}$$

Equation (C4) for the soft constraint becomes

$$P^{(\mathrm{avg,soft})}_{M_2,N,\ell}(\mathbf{n}) = \prod_{m=1}^{M_2}\frac{\Gamma(n_m+\ell)}{n_m!\Gamma(\ell)}\left(\frac{\ell M_2}{N+\ell M_2}\right)^{\ell}\left(\frac{N}{N+\ell M_2}\right)^{n_m}. \tag{C6}$$

Link between Averaged and Labeled Configurations Solution

We compute the limit $\ell \to \infty$ for the averaged distributions $P^{(\mathrm{avg,hard})}_{M,N,\ell}$ and $P^{(\mathrm{avg,soft})}_{M,N,\ell}$. With the hard constraint,

$$\begin{aligned}
\lim_{\ell\to\infty} P^{(\mathrm{avg,hard})}_{M,N,\ell}(\mathbf{n}) &= \lim_{\ell\to\infty}\frac{N!(\ell M-1)!}{(N+\ell M-1)!}\prod_{m=1}^{M}\frac{(n_m+\ell-1)!}{n_m!(\ell-1)!}\\
&= \binom{N}{\mathbf{n}}\lim_{\ell\to\infty}\frac{(\ell M-1)!}{(N+\ell M-1)!}\prod_{m=1}^{M}\frac{(n_m+\ell-1)!}{(\ell-1)!}\\
&= \binom{N}{\mathbf{n}}\lim_{\ell\to\infty}\frac{1}{(\ell M)^N}\prod_{m=1}^{M}\ell^{n_m}\\
&= \binom{N}{\mathbf{n}}\frac{1}{M^N}\\
&= P^{(\mathrm{lab,hard})}_{M,N}(\mathbf{n}).
\end{aligned} \tag{C7}$$

With the soft constraint,

$$\begin{aligned}
\lim_{\ell\to\infty} P^{(\mathrm{avg,soft})}_{M,N,\ell}(\mathbf{n}) &= \lim_{\ell\to\infty}\prod_{m=1}^{M}\frac{(n_m+\ell-1)!}{n_m!(\ell-1)!}\left(\frac{\ell M}{N+\ell M}\right)^{\ell}\left(\frac{N}{N+\ell M}\right)^{n_m}\\
&= \binom{N(\mathbf{n})}{\mathbf{n}}\frac{1}{N(\mathbf{n})!}\lim_{\ell\to\infty}\prod_{m=1}^{M}\ell^{n_m}\left(\frac{\ell M}{N+\ell M}\right)^{\ell}\left(\frac{N}{\ell M}\right)^{n_m}\\
&= \binom{N(\mathbf{n})}{\mathbf{n}}\frac{1}{N(\mathbf{n})!}\left(\frac{N}{M}\right)^{N(\mathbf{n})}\lim_{\ell\to\infty}\left(\frac{\ell M}{N+\ell M}\right)^{\ell M}\\
&= \binom{N(\mathbf{n})}{\mathbf{n}}\frac{1}{N(\mathbf{n})!}\left(\frac{N}{M}\right)^{N(\mathbf{n})}\lim_{\ell\to\infty}\left(1-\frac{N}{\ell M}\right)^{\ell M}\\
&= \binom{N(\mathbf{n})}{\mathbf{n}}\frac{1}{N(\mathbf{n})!}\left(\frac{N}{M}\right)^{N(\mathbf{n})}e^{-N}.
\end{aligned} \tag{C8}$$

This can be interpreted as the combination of a Poisson distribution for the number of individuals,

$$\lim_{\ell\to\infty} P^{(\text{avg},\text{soft})}_{M,N,\ell}(N(\mathbf{n})) = e^{-N}\frac{N^{N(\mathbf{n})}}{N(\mathbf{n})!}, \tag{C9}$$

and the distribution for the vector $\mathbf{n}$ conditional on the number of individuals,

$$\begin{aligned}\lim_{\ell\to\infty} P^{(\text{avg},\text{soft})}_{M,N,\ell}(\mathbf{n}|N(\mathbf{n}) = K) &= \lim_{\ell\to\infty} P^{(\text{avg},\text{hard})}_{M,K,\ell}(\mathbf{n}) \\ &= P^{(\text{lab},\text{hard})}_{M,K}(\mathbf{n}).\end{aligned} \tag{C10}$$

APPENDIX D

Alternative EM for Labeled Configurations

We reconsider the EM problem for labeled configurations. Instead of assuming that all vectors $\mathbf{m}$ are a priori equally probable, we assume that the prior probability of a vector $\mathbf{m}$ is proportional to $1/N(\mathbf{m})!$. This implies that vectors with the same number of individuals are a priori equally probable but that vectors $\mathbf{m}$ with a large number of individuals are a priori less probable than vectors $\mathbf{m}$ with a smaller number of individuals.

The EM procedure with the latter prior distribution and hard constraint (A1) is identical to the computation leading to $P^{(\text{lab},\text{hard})}_{M,N}$ (eqq. [A2], [A3]). We consider here the EM problem with soft constraint (A4). Using the Lagrange multiplier α,

$$P^{(\text{lab},\text{alt},\text{soft})}_{M,N}(\mathbf{m}) = \frac{1}{Z}\frac{1}{N(\mathbf{m})!}e^{-\alpha N(\mathbf{m})},$$

where $P^{(\text{lab, alt, soft})}$ denotes the probability distribution that results from applying the EM procedure for labeled configurations with alternative prior distribution $1/N(\mathbf{m})!$ and soft constraint (A4). The normalization constant Z can be calculated as follows:

$$Z = \sum_{\mathbf{m}}\frac{1}{N(\mathbf{m})!}e^{-\alpha N(\mathbf{m})} = \sum_{n=0}^{\infty}\frac{1}{n!}M^n e^{-\alpha n} = \exp\left(Me^{-\alpha}\right).$$

Imposing constraint (A4) yields

$$N = -\frac{\partial}{\partial\alpha}\ln Z = Me^{-\alpha},$$

and we can solve for the Lagrange multiplier α,

$$\alpha = \ln\frac{M}{N}.$$

As a result,

$$P^{(\text{lab},\text{alt},\text{soft})}_{M,N}(\mathbf{m}) = e^{-N}\frac{1}{N(\mathbf{m})!}\left(\frac{N}{M}\right)^{N(\mathbf{m})}, \tag{D1}$$

which gives the multicell abundance distribution

$$P^{(\text{lab},\text{alt},\text{soft})}_{M,N}(\mathbf{n}) = \binom{N(\mathbf{n})}{\mathbf{n}}e^{-N}\frac{1}{N(\mathbf{m})!}\left(\frac{N}{M}\right)^{N(\mathbf{m})}. \tag{D2}$$

This is exactly distribution (C8): a Poisson distribution for the number of individuals and distribution $P_{M,N}^{(\mathrm{lab,hard})}$ conditional on the number of individuals. As the Poisson distribution has coefficient of variation $\mathrm{CV} = 1/N^{1/2}$, hard and soft constraints are equivalent if N is large.

APPENDIX E

Harte et al. (2008)'s EM Problem

We construct an EM problem for the multicell abundance distribution that generalizes Harte et al. (2008)'s EM problem for the one-cell abundance distribution. The EM problem is formulated in terms of unlabeled configurations $\mathbf{n}$ and imposes both a hard and a soft constraint on the number of individuals N. The hard constraint restricts the set of configurations: a configuration $\mathbf{n}$ with one or more of its components $n_m > N$ has zero probability. The soft constraint states that averaged over the remaining configurations, the mean number of individuals equals N; see equation (B4).

This EM problem can be solved with the technique of Lagrange multipliers. With the Lagrange multiplier for the soft constraint (B4) denoted α, the EM solution reads

$$P_{M,N}^{(\mathrm{unl,hard/soft})}(\mathbf{n}) = \frac{1}{Z} e^{-\alpha N(\mathbf{n})},$$

where $P^{(\mathrm{unl,\ hard/soft})}$ denotes the probability distribution that results from applying the EM procedure for unlabeled configurations with both hard and soft constraints. The normalization constant Z can be calculated as follows:

$$Z = \sum_{\mathbf{n}} e^{-\alpha N(\mathbf{n})} = \left(\sum_{n=0}^{N} e^{-\alpha n} \right)^{M} = \left(\frac{1 - e^{-\alpha(N+1)}}{1 - e^{-\alpha}} \right)^{M}.$$

Imposing constraint (B4) yields

$$N = -\frac{\partial}{\partial \alpha} \ln Z = M \frac{e^{-\alpha}}{1 - e^{-\alpha}} \frac{1 - (N+1)e^{-N\alpha} + Ne^{-(N+1)\alpha}}{1 - e^{-(N+1)\alpha}}.$$

This equation can be solved numerically for the Lagrange multiplier. This is the same equation that Harte et al. (2008) solve to obtain their Lagrange multiplier (see their eq. [B-5]). As a result,

$$\begin{aligned} P_{M,N}^{(\mathrm{unl,hard/soft})}(\mathbf{n}) &= \left(\frac{1 - e^{-\alpha}}{1 - e^{-\alpha(N+1)}} \right)^{M} e^{-\alpha N(\mathbf{n})} \\ &= \prod_{m=1}^{M} \left(\frac{1 - e^{-\alpha}}{1 - e^{-\alpha(N+1)}} e^{-\alpha n_m} \right). \end{aligned} \tag{E1}$$

This is a product of one-cell abundance distributions, each of which is given by

$$P_{M,N}^{(\mathrm{unl,hard/soft})}(n_1) = \frac{1 - e^{-\alpha}}{1 - e^{-\alpha(N+1)}} e^{-\alpha n_1} \tag{E2}$$

for $n_1 \leq N$.

This is exactly the one-cell abundance distribution obtained by Harte et al. (2008; see their eq. [9]). Therefore, our EM problem embeds the one-cell abundance distribution of Harte et al. (2008) in a multicell abundance distribution.

The EM distribution (E1) is not scale consistent. To see this, consider first the distribution (E1) on the fine scale M_1. Configurations $\mathbf{n}$ with all components $n_m \leq N$ have a nonzero probability. We scale transform this distribution to the coarse scale M_2 (scale factor ℓ). The resulting distribution assigns a nonzero probability to configurations $\mathbf{n}$ with all components $n_m \leq \ell N$. Next, consider the distribution (E1) obtained by applying EM directly on the coarse scale M_2. This distribution has only nonzero probability configurations $\mathbf{n}$ with all components $n_m \leq N$. Hence, the scale-transformed EM distribution and the direct EM distribution are different, and so scale consistency is not satisfied.

Literature Cited

Banavar, J. R., and A. Maritan. 2007. The maximum relative entropy principle. Preprint, http://arxiv.org/abs/cond-mat/0703622.

Chave, J., D. Alonso, and R. S. Etienne. 2006. Comparing models of species abundance. Nature 414:E1–E2.

Coleman, B. R. 1981. On random placement and species-area relations. Mathematical Biosciences 54:191–215.

Conlisk, E., M. Bloxham, J. Conlisk, B. Enquist, and J. Harte. 2007. A new class of models of spatial distribution. Ecological Monographs 77:269–284.

Dewar, R. C., and A. Porté. 2008. Statistical mechanics unifies different ecological patterns. Journal of Theoretical Biology 251:389–403.

Etienne, R. S. 2005. A new sampling formula for neutral biodiversity. Ecology Letters 8:253–260.

———. 2007. A neutral sampling formula for multiple samples and an "exact" test of neutrality. Ecology Letters 10:608–618.

Etienne, R. S., and H. Olff. 2005. Confronting different models of community structure to species-abundance data: a Bayesian model comparison. Ecology Letters 8:493–504.

Haegeman, B., and M. Loreau. 2008. Limitations of entropy maximization in ecology. Oikos 117:1700–1710.

———. 2009. Trivial and non-trivial applications of entropy maximization in ecology: a reply to Shipley. Oikos 118:1270–1278.

Harte, J., E. Conlisk, A. Ostling, J. Green, and A. Smith. 2005. A theory of spatial structure in ecological communities at multiple spatial scales. Ecological Monographs 75:179–197.

Harte, J., T. Zillio, E. Conlisk, and A. B. Smith. 2008. Maximum entropy and the state variable approach to macroecology. Ecology 89:2700–2711.

Jaynes, E. T. 1957. Information theory and statistical mechanics. Physical Review 106:620–630.

———. 2003. Probability theory: the logic of science. Cambridge University Press, Cambridge.

Johnson, N. L., S. Kotz, and N. Balakrishnan. 1997. Discrete multivariate distributions. Wiley, New York.

MacArthur, R. 1960. On the relative abundance of species. American Naturalist 94:25–36.

Maddux, R. D. 2004. Self-similarity and the species-area relationship. American Naturalist 163:616–626.

Ostling, A., J. Harte, J. L. Green, and A. P. Kinzig. 2004. Self-similarity, the power law form of the species-area relationship, and a probability rule: a reply to Maddux. American Naturalist 163:627–633.

Pueyo, S., F. He, and T. Zillio. 2007. The maximum entropy formalism and the idiosyncratic theory of biodiversity. Ecology Letters 11:1017–1028.

Shipley, B., D. Vile, and E. Garnier. 2006. From plant traits to plant communities: a statistical mechanistic approach to biodiversity. Science 314:812–814.

Volkov, I., J. R. Banavar, S. P. Hubbell, and A. Maritan. 2003. Neutral theory and relative species abundance in ecology. Nature 424:1035–1037.

Associate Editor: Axel G. Rossberg
Editor: Mark A. McPeek