Applications of the principle of maximum entropy: from physics to ecology

2010 J. Phys.: Condens. Matter 22 063101

(http://iopscience.iop.org/0953-8984/22/6/063101)

IOP PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **22** (2010) 063101 (13pp) doi:10.1088/0953-8984/22/6/063101

TOPICAL REVIEW

Applications of the principle of maximum entropy: from physics to ecology

Jayanth R Banavar[1], **Amos Maritan**[2] **and Igor Volkov**[1,3]

[1] Department of Physics, 104 Davey Laboratory, The Pennsylvania State University, University Park, PA 16802, USA
[2] Dipartimento di Fisica 'G Galilei', Università di Padova CNISM and INFN, via Marzolo 8, 35131 Padova, Italy
[3] Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA

Received 10 November 2009, in final form 17 December 2009
Published 22 January 2010
Online at stacks.iop.org/JPhysCM/22/063101

Abstract

There are numerous situations in physics and other disciplines which can be described at different levels of detail in terms of probability distributions. Such descriptions arise either intrinsically as in quantum mechanics, or because of the vast amount of details necessary for a complete description as, for example, in Brownian motion and in many-body systems. We show that an application of the principle of maximum entropy for estimating the underlying probability distribution can depend on the variables used for describing the system. The choice of characterization of the system carries with it implicit assumptions about fundamental attributes such as whether the system is classical or quantum mechanical or equivalently whether the individuals are distinguishable or indistinguishable. We show that the correct procedure entails the maximization of the relative entropy subject to known constraints and, additionally, requires knowledge of the behavior of the system in the absence of these constraints. We present an application of the principle of maximum entropy to understanding species diversity in ecology and introduce a new statistical ensemble corresponding to the distribution of a variable population of individuals into a set of species not defined *a priori*.

Contents

1. Introduction

The principle of maximum entropy [1–12] is a widely used variational method for the analysis of both complex equilibrium and non-equilibrium systems and is being increasingly employed in a variety of contexts such as ecology [13–15], nuclear magnetic resonance spectroscopy [16], x-ray diffraction [17], electron microscopy [18], and neuroscience [19] for inference from incomplete data. In many instances, it is convenient and/or useful to describe a system and known constraints either in terms of a full description or in a coarse-grained manner. Of course, one would expect and require that the results of any analysis be robust under coarse-graining.

A key issue in combinatorics is that of distinguishability or lack thereof. Imagine rolling a pair of dice—an outcome of two specific distinct numbers on the dice (say, a three and a five) is twice as likely as getting the same specific

number on both dice (say, a three and a three). This is the premise underlying classical statistical mechanics. The two dice are distinguishable. The same experiment carried out with quantum dice, which are indistinguishable, yields exactly the same probability for the two outcomes in which the two numbers are the same or different. The maximum entropy principle is traditionally applied in statistical mechanics to study the distribution of balls (electrons, atoms, molecules etc)—each colored differently and distinguishable or colored the same and indistinguishable—into a set of boxes (energy levels). Boltzmann statistics results when one considers distinguishable balls, and Bose–Einstein statistics for the case of indistinguishable balls. Fermi–Dirac statistics arises when there is a ceiling on the maximum occupancy of indistinguishable balls in a box.

Consider a snapshot of a tropical forest comprised of trees of many different species. A cornerstone of studies of biodiversity is the relative species abundance, which measures the fraction of species having a given abundance. This measure is particularly important for rare species, i.e. species having a low abundance, because these species could become extinct (at least in the local region) more readily than the more abundant species. The species–area relationship is a very useful benchmark as well—it measures the number of distinct species as a function of the sampled area. An important use of this measure arises when one wishes to estimate the effect on biodiversity of diminishing the area available to an ecosystem due to habitat destruction or climate change. One can imagine that the trees in a forest are akin to distinguishable balls that have been categorized into species or boxes. The measures in ecology are therefore somewhat analogous to standard physics distributions and one might ask whether they can be elucidated using the principle of maximum entropy to determine the most probable outcome given certain constraints.

We will show that the details of the dynamics play a pivotal role in the combinatorics to be used in the maximum entropy principle. More importantly, the very choice of how one characterizes a system carries with it implicit assumptions about fundamental attributes such as whether the system is classical or quantum mechanical or equivalently whether the individuals are distinguishable or indistinguishable. Thus, unless one is careful, the results that one obtains may be an artifact of an improper choice.

Our presentation is pedagogical as befits a topical review. There are a number of specialized reviews of the principle of maximum entropy with application to physics [12], information theory [20] and natural language [21] just to cite a few examples. The paper is organized as follows. In section 2, the concept of entropy is introduced in a rather intuitive way and the maximum entropy ('maxent') principle is explained. Section 3 contains two paradigmatic examples of the use of the maxent principle: the case of distinguishable and indistinguishable individuals. The relative entropy is derived within a simple example and its properties are discussed in section 4 together with the maximum relative entropy ('maxrent') principle [15]. The *a priori* probability entering in the definition of relative entropy is related to the system dynamics in section 5. Section 6 contains a critique of the use of the maxent principle in ecology. A new statistical ensemble is introduced which is suitable for describing the relative species abundance in ecology. We conclude with a brief summary in section 7.

2. General considerations

Consider the familiar example of rolling a cubic dice N times. The total possible number of distinct outcomes is represented by E and for a dice $E = 6$. In ecology, one may similarly carry out a thought experiment in which N represents the number of independent realizations or snapshots of an ecological community under equivalent conditions. Each realization can be characterized by various attributes, e.g., the species abundances. E, in this case, would represent the total number of possible distinct measures of the species abundances. In the absence of any additional information, one might assign equal probability to all E outcomes stated as the *principle of insufficient reason* by Laplace. However, in the presence of new information, which can be expressed as constraints, the challenge is to assign probabilities to the E outcomes which ensures that the constraints are satisfied *without making any unwarranted additional assumptions*.

Let $\vec{n} \equiv (n_1, n_2, \ldots, n_E)$ denote a situation in which the N realizations yield n_1 instances of outcome 1, n_2 instances of outcome 2, ..., and n_E instances of outcome E. Each realization is postulated to yield one out of E possible outcomes. Thus the total number of conceivable distinct results of the outcome of all N realizations is E^N. Of these, the number corresponding to $\vec{n} \equiv (n_1, n_2, \ldots, n_E)$ is given, from simple combinatorics, by

$$W(\vec{n}) = \frac{N!}{\prod_{i=1}^{E} n_i!} \tag{1}$$

with the constraint

$$\sum_{i=1}^{E} n_i = N. \tag{2}$$

The numerator in equation (1) represents all possible choices of the N realizations whereas the denominator takes into account the fact that unlike interchanges between different outcomes, interchanges within the same outcome are unobservable (figure 1).

When $n_1, n_2, \ldots$ are all large, using Stirling's approximation, $n_i! \approx \mathrm{e}^{-n_i} n_i^{n_i}$, one obtains

$$\frac{\ln W(\vec{n})}{N} \underset{N\to\infty}{=} -\sum_i P_i \ln P_i \equiv \mathcal{H}(\vec{P}), \tag{3}$$

where P_i, the frequency of occurrence of the ith event, is given by

$$P_i = \frac{n_i}{N}. \tag{4}$$

$\mathcal{H}$ is the entropy of the 'distribution' probability $\vec{P} \equiv (P_1, P_2, \ldots, P_E)$. Equation (3) tells us that if we repeat the 'experiment' N times, then the number of times we get the distribution $\vec{n} \equiv (n_1, n_2, \ldots) = N\vec{P}$ is proportional to

$$W(\vec{n}) \propto \exp\left(N\mathcal{H}(\vec{P})\right). \tag{5}$$

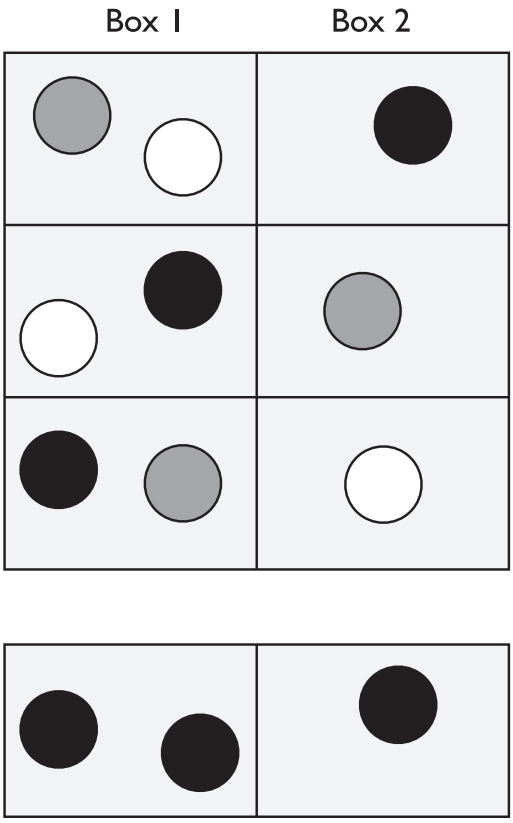

Figure 1. Different counting schemes for distinguishable and indistinguishable individuals. As an example, consider $\mathcal{N} = 3$, $\mathcal{S} = 2$. For the case of distinguishable individuals (shaded circles in the top three panels), there are three distinct configurations with $\vec{n} = (n_1 = 2, n_2 = 1)$, $W(\vec{n}) \overset{(1)}{=} 3$ (interchanging the two individuals within a species does not lead to a new configuration). In terms of the $\vec{i}$ representation (see section 3), the configurations above are $\vec{i} = (i_1, i_2, i_3) = (1, 1, 2)$, (1 ,2, 1) and (2, 1, 1) where the first individual is white, the second is grey, and the third is black. For indistinguishable individuals (black circles) the configurations above are identical and there is no way to detect a difference among the three cases. Thus for any $\vec{n}$, the corresponding number of ways of obtaining it is $W_I(\vec{n}) = 1$, i.e. there is just one way to get the distribution $\vec{n} = (n_1, n_2, \ldots, n_S)$.

Thus the most probable $\vec{n} \equiv (n_1, n_2, \ldots, n_E)$ among the total number of conceivable distinct results of the outcome of all N realizations is obtained by maximizing the entropy (3) under the constraint

$$\sum_{i=1}^{E} P_i = 1, \tag{6}$$

which is termed the normalization condition. If this is the only constraint in the problem, then the maximum of $\mathcal{H}(\vec{P})$ is attained with $P_i = 1/E$, as one would intuitively expect (see also section 3). A few important properties of the entropy are recalled and derived in appendix A. It is slightly technical and is not necessary for understanding the rest of the paper.

The existence of other knowledge about the system results in additional constraints which serve to limit the range of possible $\vec{n}$s or equivalently the possible $\vec{P}$s. This additional knowledge can often be encapsulated in the form of constraints on averages of certain quantities. For example, a quantity Q, whose average value is $\bar{Q}$ and which has value Q_i when one has an outcome i (for example, for the number rolled in a dice throw, $Q_i = i$), obeys the constraint

$$\langle Q \rangle \equiv \sum_{i=1}^{E} P_i Q_i = \bar{Q}. \tag{7}$$

In order to find the maximum of $H(P) = -\sum_i P_i \ln P_i$ subject to the constraints given by equations (6) and (7) one can introduce the Lagrange function with Lagrange multipliers α and β [22] (see appendix B). Taking derivatives with respect to P_j and setting them to zero:

$$\begin{aligned} 0 &= \frac{\partial}{\partial P_j}\left[-\sum_i P_i \ln P_i - \alpha \sum_i P_i - \beta \sum_i P_i Q_i\right] \\ &= -1 - \ln P_j - \alpha - \beta Q_j \end{aligned} \tag{8}$$

one gets the following expression:

$$P_i = \mathrm{e}^{-1-\alpha-\beta Q_i}. \tag{9}$$

Applying the normalization constraint equation (6), one gets

$$P_i = \mathrm{e}^{-\beta Q_i}/Z, \tag{10}$$

where $Z = \sum_i \mathrm{e}^{-\beta Q_i}$. Usually constraint equation (7) cannot be solved explicitly and can be expressed in the compact form

$$\bar{Q} = -\frac{\partial Z}{\partial \beta}. \tag{11}$$

It can be shown that $W(\vec{n})$ in equation (1) is unique and has a sharp maximum corresponding to the most probable distribution as $N \to \infty$.

3. Distinguishable and indistinguishable individuals

3.1. Boltzmann statistics

Consider $\mathcal{N}$ distinguishable individuals distributed among $\mathcal{S}$ species (boxes). Let $P(k)$, the normalized relative species abundance, denote the probability that a given species is comprised of k individuals. Let the αth individual (α runs from 1 to $\mathcal{N}$) belong to the i_αth species (the range of i_α is from 1 to $\mathcal{S}$) and thus $E = \mathcal{S}^{\mathcal{N}}$. The outcome (event) can be denoted by the $i \equiv (i_1, i_2, \ldots, i_{\mathcal{N}})$ (individual 1 in species i_1, individual 2 in species i_2 etc). Let us impose a constraint that the average number of individuals in a specific species, e.g. species 1, is equal to $\bar{k}$.

This corresponds to defining $Q_i = k_i \equiv \sum_{\alpha=1}^{\mathcal{N}} \delta_{i_\alpha,1}$, the number of individuals in species 1 in the $i \equiv (i_1, i_2, \ldots, i_{\mathcal{N}})$ event, and $\delta_{i,j}$ is the Kronecker delta function, equal to 1 if $i = j$ and zero otherwise. Thus the constraint becomes $\langle k \rangle = \sum_i P_i k_i = \bar{k}$. Using the general results of section 2 one obtains

$$P_i = \frac{\mathrm{e}^{-\beta k_i}}{Z} \tag{12}$$

with

$$Z = \sum_i \mathrm{e}^{-\beta k_i} = \sum_{i_1,\ldots,i_N} \mathrm{e}^{-\beta \sum_{\alpha=1}^{\mathcal{N}} \delta_{i_\alpha,1}}$$
$$= \left(\sum_{i=1}^{\mathcal{S}} \mathrm{e}^{-\beta\delta_{i,1}}\right)^{\mathcal{N}} = (\mathcal{S} - 1 + \mathrm{e}^{-\beta})^{\mathcal{N}}. \quad (13)$$

The constraint $\bar{k} = \langle k \rangle$ leads to

$$\bar{k} = -\frac{\partial}{\partial\beta} \ln Z = \frac{\mathcal{N}}{1 + (\mathcal{S} - 1)\mathrm{e}^{\beta}} \overset{\mathcal{S}\gg 1}{\to} \frac{\mathcal{N}}{\mathcal{S}} \mathrm{e}^{-\beta}, \quad (14)$$

i.e. $\beta = \ln(\frac{\mathcal{N}}{\mathcal{S}}/\bar{k})$. Note that in the absence of the constraint on $\langle k \rangle$ there is no Lagrange multiplier, that is $\beta = 0$, implying the obvious result $\langle k \rangle = \mathcal{N}/\mathcal{S}$.

In order to obtain the relative species abundance, $P(k)$, we carry out a coarse-graining procedure in which we sum P_i over all configurations in which there are k individuals in the species of interest. (The coarse-graining, as defined here, is merely a change of description to a coarser level. For example, one may consider tossing 10 coins and asking what the outcome for each coin toss was—i.e. was the first coin a head, was the second a head etc. This is a fine-level description when the coins are distinguishable. Alternatively one may simply record the total number of heads (and tails) and this would correspond to a coarse-grained description of the same coin toss experiment.) When $\mathcal{N}$ and $\mathcal{S}$ are large, one obtains

$$P(k) = \sum_{i=1}^{E} P_i \delta_{k,k_i} = \bar{k}^k \mathrm{e}^{-\bar{k}}/k!. \quad (15)$$

This is the familiar Poisson distribution or the grand-canonical Boltzmann distribution, when all energy levels are the same.

3.2. Bose statistics

Let us now consider the case of indistinguishable individuals. The best we can do is to study how many individuals there are in a given species, called the occupation number representation. We are unable to discern the identity of any individual and thus we work directly with $P(k)$. We seek to maximize the entropy

$$\mathcal{H}(P) \equiv -\sum_k P(k) \ln P(k) \quad (16)$$

subject to the same constraints as before. Following the same procedure as in the previous subsection, one obtains the familiar grand-canonical Bose–Einstein distribution (with all energy levels being the same)

$$P(k) = \mathrm{e}^{-\beta k}(1 - \mathrm{e}^{-\beta}), \quad (17)$$

which is a pure exponential function. One finds that

$$\bar{k} = \langle k \rangle = \frac{1}{\mathrm{e}^{\beta} - 1}, \quad (18)$$

yielding

$$P(k) = \frac{\bar{k}^k}{(1 + \bar{k})^{k+1}}. \quad (19)$$

Note that equations (19) and (15) are different. Indeed they correspond to different underlying (implicit) hypotheses regarding the nature of the distinguishability of individuals. This is reflected in the maximum entropy principle by implementing a specific representation, i.e. the label representation for distinguishable individuals used in the previous subsection and occupation number representation for the indistinguishable case used in equation (16). In other words, how one chooses to characterize the system and the level of description one uses carry with them implicit assumptions pertaining to the distinguishability or lack thereof of the individuals. The issue of indistinguishability may be viewed as another type of constraint: all *microscopic* configurations corresponding to the interchange of any two individuals ought to be considered as the same configuration. The choice of the occupation number representation takes care of this constraint automatically and it is in fact the appropriate one for deriving quantum statistics. When the individuals are indistinguishable, all the information that one has is encapsulated by $P(k)$. The conundrum is that the result obtained on applying the maximum entropy principle to P_i and then coarse-graining the result to obtain $P(k)$ in equation (15) is different from that obtained on applying the maximum entropy principle directly to $P(k)$, equation (19). In other words, the operations of the entropy maximization and of coarse-graining do *not* commute.

4. Relative entropy and the maxrent principle

We turn now to a resolution of this puzzle by invoking the concept of relative entropy [23, 24, 15]. We suggest that the correct application of the principle of maximum entropy entails the maximization of the relative entropy

$$\mathcal{H}_{\mathrm{C-G}}(\vec{P}) \equiv -\sum_i P_i \ln \frac{P_i}{P_{0i}} \quad (20)$$

subject to the constraints imposed by our partial knowledge of the system. The subscript C-G stands for coarse-grained. P_{0i} is the reference probability and has the physical meaning that, on maximizing the entropy, P_i is equal to P_{0i} in the absence of any constraints. The crucial observation is that one must have knowledge of P_{0i} in order to apply the method successfully. The lesson learned from the success of the method in physics is that when one uses as complete a description of a system as possible, the reference term is uniform. Thus for distinguishable individuals one can use the label representation (which is the most detailed representation) and set $P_{0i} = \text{const}$ and obtain Boltzmann statistics. Should one choose to use an occupation number representation for distinguishable individuals, one would need to transform the uniform P_{0i} in the label representation to the occupation number representation and again obtain Boltzmann statistics. However, were one to use the occupation number representation and employ a uniform P_{0i}, one would obtain Bose–Einstein statistics instead of Boltzmann statistics. This underscores the fact that the most complete description of indistinguishable individuals necessarily involves use of the occupation number representation—the label representation is not suitable for indistinguishable individuals.

In order to understand the form of the relative entropy, equation (20), let us consider a simple example pertaining to our original case of distinguishable individuals which allowed us to introduce the concept of entropy and the maximum entropy principle. Suppose that each species has a fine structure, i.e. the ith species contains g_i subspecies. In this case we have to deal with a probability for the αth individual ($\alpha = 1, \ldots, \mathcal{N}$) to be found in the i_αth species and in the κ_{i_α} th subspecies ($\kappa_i = 1, \ldots, g_i$). The outcome is now given by $(i_1, \kappa_{i_1}, i_2, \kappa_{i_2}, \ldots, i_\mathcal{N}, \kappa_{i_\mathcal{N}}) \equiv i\kappa$. The entropy is simply obtained by generalizing equation (3) to

$$\mathcal{H}(\vec{P}) = -\sum_{i\kappa} P_{i\kappa} \ln P_{i\kappa}. \tag{21}$$

Let us assume that the constraints do not depend on $\kappa \equiv (\kappa_{i_1}, \kappa_{i_2}, \ldots, \kappa_{i_\mathcal{N}})$:

$$\sum_{i\kappa} P_{i\kappa} = 1, \tag{22}$$

$$\sum_{i\kappa} P_{i\kappa} Q_i = \bar{Q}. \tag{23}$$

The former is the normalization condition and the latter was introduced earlier as equation (7). Maximizing the entropy with the two constraints yields

$$P_{i\kappa} = \mathrm{e}^{-\beta Q_i}/Z, \tag{24}$$

as in equation (12), which is independent of κ. Thus the probability of observing the outcome $i \equiv (i_1, i_2, \ldots, i_\mathcal{N})$, i.e. the first individual in species 1, the second individual in species 2 etc independent of the subspecies they belong to, is given by the (marginalized) probability

$$P_i \equiv \sum_\kappa P_{i\kappa} \propto \mathrm{e}^{-\beta Q_i} P_{0i} \tag{25}$$

with $P_{0i} \propto g_{i_1} g_{i_2}, \ldots, g_{i_\mathcal{N}}$. P_{0i} is the reference probability.

Because the constraints do not depend on κ, one may substitute

$$P_{i\kappa} = \frac{P_i}{P_{0i}} \tag{26}$$

in equations (21)–(23). The constraints become

$$\sum_i P_i = 1 \tag{27}$$

$$\sum_i P_i Q_i = \bar{Q}. \tag{28}$$

The key finding is that the correct answer (25) is obtained if one maximizes the relative entropy

$$\mathcal{H}_{\mathrm{C-G}}(\vec{P}) \equiv -\sum_i P_i \ln \frac{P_i}{P_{0i}} \tag{29}$$

subject to the constraint equations (27) and (28). This demonstrates that the coarse-graining procedure requires the inclusion of the reference term P_{0i} in order to get the correct answer independent of whether the maximum entropy principle is applied before or after the coarse-graining. Thus in the example above, the reference probability is $P_{0i\kappa} = 1$ before the coarse-graining, whereas it is $P_{0i} = \sum_\kappa P_{0i\kappa}$ after the coarse-graining (cf equation (25)).

We therefore suggest that the correct and consistent application of the maximum entropy principle entails the maximization of the relative entropy [23] instead of the Shannon entropy in equation (3) subject again to the constraints obtained from partial knowledge that one has about the system. The reference term has been discussed in the literature in the different context of going from a discrete to a continuous system and is 'proportional to the limiting density of discrete points' [7], where it is needed for dimensional reasons. The reference term is, however, not commonly invoked as an essential ingredient in the discrete case. It has been shown by Shore and Johnson [5] that 'given a continuous prior density and new constraints, there is only one posterior density satisfying these constraints that can be chosen by a procedure that satisfies the axioms'. The unique posterior can be obtained by maximizing the relative entropy and the axioms pertain to uniqueness, invariance, system independence and subset independence. If P_{0i} can be chosen to be a constant or simply equal to 1, equation (29) becomes equivalent to equation (3).

We return to the puzzle stated earlier pertaining to the non-commutability of the application of the maximum entropy principle and coarse-graining. The puzzle is resolved by the use of a reference term $\frac{1}{k!}$ in equation (16), which emerges as the large-N limit of $\mathcal{N}!/(k!(\mathcal{N}-k)!)$, yielding the Poisson distribution equation (15). Indeed, in the derivation of equation (15), it was implicitly assumed that $P_{0,i}$ is a constant. On coarse-graining to a description involving the variable k, one obtains $P_0(k) \propto \frac{\mathcal{N}!}{k!(\mathcal{N}-k)!}$, yielding equation (15). (This result is obtained by summing P_{0i} over all configurations with k individuals in a given species.) If, instead, one assumes that $P_0(k)$ is a constant, which is appropriate when individuals are indistinguishable, then one derives the Bose–Einstein distribution, equation (17). Recently, Dunkel *et al* have used the notion of the relative entropy to explore the relativistic version of Maxwell's velocity distribution of an ideal gas. The importance of the relative entropy has been underscored by Dewar and Porte, who have coined the name maxrent for the maximization of the relative entropy.

The success of the principle of maximum entropy hinges on the choice of the reference probability, P_{0i}, and the identification of the correct constraints not encapsulated in P_{0i}. In the statistical mechanics examples studied above, the constraint is imposed by fixing, e.g., the average energy while the choice of P_{0i} is guided by the postulate that all states are *a priori* equally probable when one works at the finest level of description for the system being studied. Of course, this follows from the dynamics of the system.

5. System dynamics

Consider the dynamics, in terms of a Markov process, in the occupation number representation. We will use the subscripts BE and B to denote the Bose–Einstein and Boltzmann cases respectively. If the transition rate, $W^{\mathrm{quantum}}(n_j \to n_j + 1)$ ($W^{\mathrm{quantum}}(n_j \to n_j - 1)$) is proportional to $n_j + 1$ (n_j) then, in the stationary state, $P_{0,\mathrm{BE}}(\vec{n}) = \mathrm{const}$ in agreement

with the implicit choice made for the Bose–Einstein case, equation (17). These transition rates follow from the symmetry of the quantum wavefunction describing indistinguishable individuals [25]. For classical (distinguishable) individuals, the transition rate $W^{\text{classical}}(n_j \to n_j + 1)$ is simply constant whereas the transition rate $W^{\text{classical}}(n_j \to n_j - 1)$ is proportional to n_j. The stationary state in this case is given by $P_{0,B}(\vec{n}) = 1/\prod_i n_i!$ and, substituting in equation (29), one obtains Boltzmann statistics in the occupation number representation.

We now return to the problem of determination of the relative species abundance, $P(k)$. Consider the simple case in which all species are demographically equivalent [26] and are governed by similar death and birth rates. A naive application of the maximum entropy principle without the appropriate non-trivial reference term and with the constraint that the average population is fixed yields a simple exponential form for the species abundance

$$P(k) \propto \mathrm{e}^{-\beta k}. \tag{30}$$

as in equation (17) for the case of indistinguishable individuals. In order to choose the reference entropy, we turn again to the dynamics as a guide. Consider a Markov process with transition rates $W^{\text{eco}}(k \to k \pm 1) = k + c$ where c is a constant term that, for simplicity, is species independent. When $c = 0$, one has a simple birth–death process, whose rate is proportional to the number of individuals of a given species. A non-zero value of c introduces density dependence in the birth and death rates with a positive value of c corresponding to a rare-species advantage [27]. The stationary state corresponding to these dynamics provides information pertaining to the reference probability $P_{0,k} \propto 1/(k + c)$. On applying the principle of maximum relative entropy with this reference probability, one finds

$$P(k) \propto \mathrm{e}^{-\beta k}/(k + c). \tag{31}$$

When $c = 0$, we obtain the celebrated Fisher log-series [28]. (Note that this result can also be obtained from the standard application of the principle of maximum entropy by imposing a constraint on the average value of $\ln n$, a constraint with no ecological basis.) When c is positive, one obtains the result derived using a density-dependent neutral approach [27] which fits the relative species abundance data of several tropical forests fairly well. The key point is that if one chooses to work in a coarse-grained description, as we did here, it is crucial to obtain a reference probability arising from the dynamics in the absence of any constraint. Thus, ignoring the reference probability corresponds to making precise assumptions on the dynamics that has led the system to the observed state.

6. Using maxent in ecology

6.1. A new statistical ensemble for ecological systems

As noted above, one can use the principle of maximum relative entropy to readily derive an expression for the relative species abundance (RSA) of an ecosystem using the dynamics as a guide. There have been recent attempts to apply the principle of the maximum entropy method to ecology. There are pitfalls that one encounters when one applies the principle in a naive manner to non-equilibrium phenomena. One also has to recognize the difference between distinguishable and non-distinguishable entities as discussed in section 3. And, as shown in section 4, one ought to work with the set of variables which provides as complete a description of the system as possible and hope that, in this description, the reference term, P_{0i}, is constant.

In order to assess the consequences of the application of the principle of maximum entropy, let us begin with the apparently plausible assumption that the abundance of each species within a single trophic level is observable, i.e. the species of trees in a tropical forest are labeled and distinguishable. We will assume that the trees in the forest belong to functional groups and that there are g species within each functional group. g plays the role of degeneracy of the energy levels in statistical mechanics. Let m_i^α denote the population of the αth functional group ($\alpha = 1, \ldots, g$) of the ith species ($i = 1, \ldots, S$). Let $\mathcal{P}(\vec{m})$ denote the probability distribution function of the m satisfying the constraints

$$\sum_{\vec{m}} \mathcal{P}(\vec{m}) = 1, \tag{32}$$

$$\sum_{\vec{m}} \mathcal{P}(\vec{m}) \sum_{i,\alpha} m_i^\alpha = N, \tag{33}$$

where the first constraint is simply the normalization and the second ensures that one has a fixed average population. As noted earlier, the $\vec{m}$ representation is the appropriate one for deriving quantum statistics—the individuals are indistinguishable and all the information that one has is encapsulated by $\mathcal{P}(\vec{m})$. The key point is that the $\vec{m}$ description along with the choice of the reference term being equal to 1 is tantamount to the inconsistent assumption of distinguishable species and indistinguishable trees. Proceeding, nevertheless, with a naive application of the principle of maximum entropy yields

$$\mathcal{P}(\vec{m}) = Z^{-1}\mathrm{e}^{-\beta \sum_{i,\alpha} m_i^\alpha} \qquad \text{where } Z = [1 - \mathrm{e}^{-\beta}]^{-gS}. \tag{34}$$

If one is interested only in the probability distribution of the total population of each species, $n_i = \sum_\alpha m_i^\alpha$, then the marginalized distribution

$$P(\vec{n}) = \left\langle \prod_i \delta_{\sum_\alpha m_i^\alpha, n_i} \right\rangle = \sum_{\vec{m}:\, \sum_\alpha m_i^\alpha = n_i} \mathcal{P}(\vec{m}) \tag{35}$$

is readily obtainable from equation (34):

$$P(\vec{n}) = Z^{-1}\mathrm{e}^{-\beta \sum_i n_i} \prod_i \binom{n_i + g - 1}{n_i}, \tag{36}$$

and each $n_i = 0, 1, \ldots$. Note that the n-representation corresponds to a coarser description than the m-representation. The above result, derived by Harte *et al* for the $g = 1$ case, corresponds to Bose–Einstein statistics for the trees with the species playing the role of the energy levels, all with the same energy. (Harte *et al* [14] considered an additional variable,

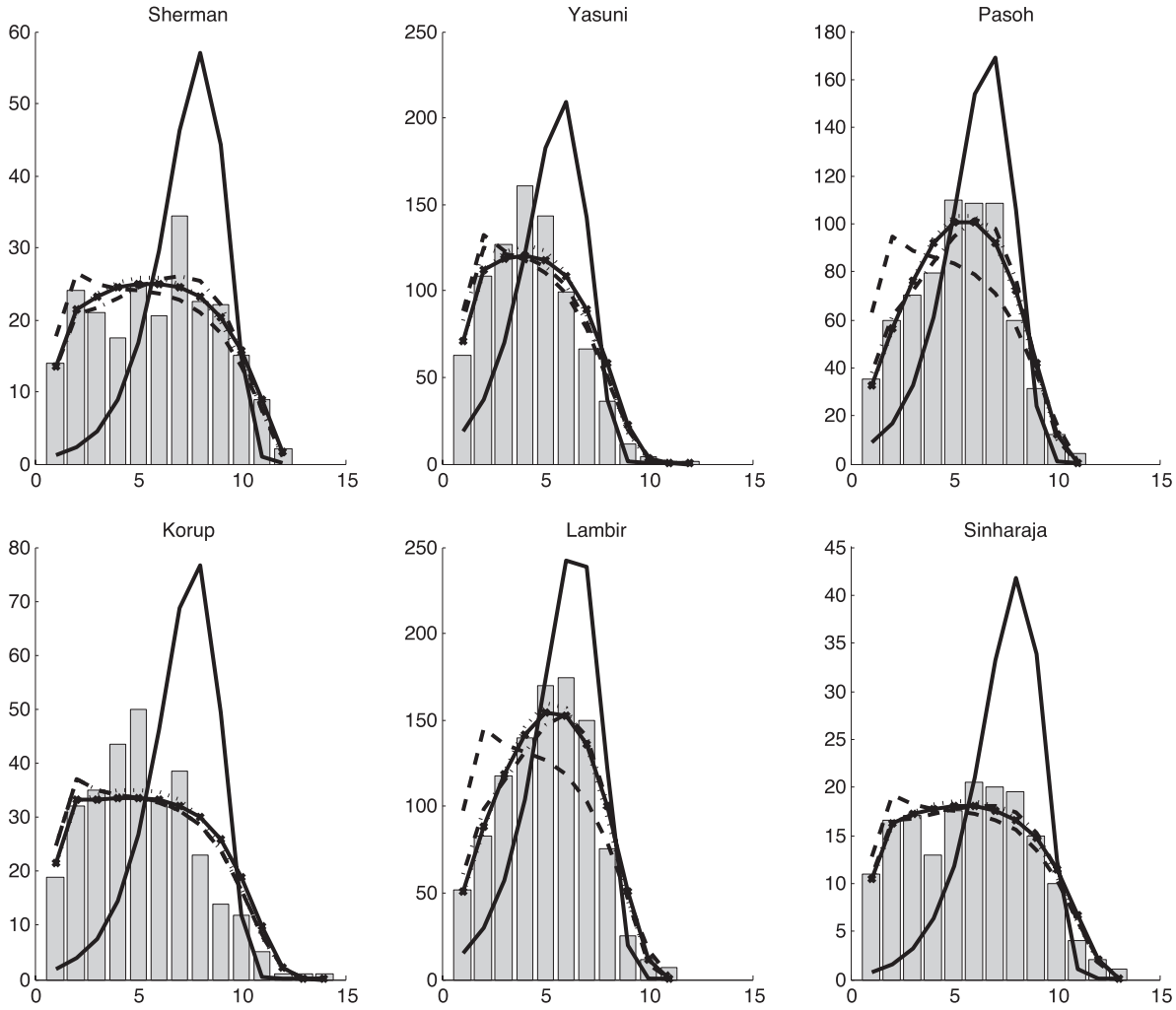

Figure 2. Fits of five models to the tree species abundance data from the Sherman, Yasuni, Pasoh, Lambir, Korup and Sinharaja plots, for trees >10 cm in stem diameter at breast height (see table 1). Dotted, dashed, solid, cross and dash–dotted lines correspond to: Model 1: the density-dependent equation (31); Model 2: Fisher log-series equation (31) with $c = 0$; Model 3: exponential distribution equation (38); Model 4: equation (50); and Model 5: equation (37). The frequency distributions are plotted using Preston's binning method. The numbers on the x axis represent Preston's octave classes. The second and third models (Fisher log-series and exponential distribution) perform relatively poorly while the three other models provide better fits.

the metabolic energy, and applied the maximum entropy principle for a joint distribution of $\vec{n}$ and the metabolic energy. Integrating over the metabolic energy, they find an extra $1/n_i$ for each of the S species yielding the Fisher log-series for the RSA.) The RSA can be obtained from equation (36) by summing over all ns but one:

$$P_{\mathrm{RSA}}(n) = (1 - \mathrm{e}^{-\beta})^g \mathrm{e}^{-\beta n} \binom{n+g-1}{n} \tag{37}$$

$$\xrightarrow{g=1} (1 - \mathrm{e}^{-\beta})\mathrm{e}^{-\beta n}. \tag{38}$$

If we further coarsen the description in terms of the variables, $\phi_k(\vec{n}) = \sum_i \delta_{n_i,k}$, the number of species with abundance k (note that $k = 0$ is also included here because the species have labels and therefore one can observe the absence of species), one finds (for $g = 1$)

$$\hat{P}_n(\phi) = \left\langle \prod_{k\geqslant 0} \delta_{\phi_k(\vec{n}),\phi_k} \right\rangle = Z^{-1} \frac{S!}{\prod_k \phi_k!} \mathrm{e}^{-\beta \sum_k k\phi_k}. \tag{39}$$

The RSA is obtained by averaging ϕ_n:

$$\frac{\langle \phi_n \rangle}{S} = P_{\mathrm{RSA}}(n). \tag{40}$$

The result obtained above is based on inconsistent assumptions pertaining to the notion of distinguishable species and indistinguishable individuals.

We now turn to an alternative approach for deducing the relative species abundance based on a new representation. It is based on doing away with the idea of labeling the species—after all, one does not necessarily observe exactly the same species in all forests around the globe and thus specifying the abundance of a given species is not appropriate. The configurations of our system consist of partitioning a variable population of individuals into a set of species not defined *a priori*. Thus the observable quantity is $\vec{\phi} = (\phi_1, \phi_2, \ldots)$, where ϕ_k is the number of species with population k and we are interested in $P(\vec{\phi})$, the probability of observing the

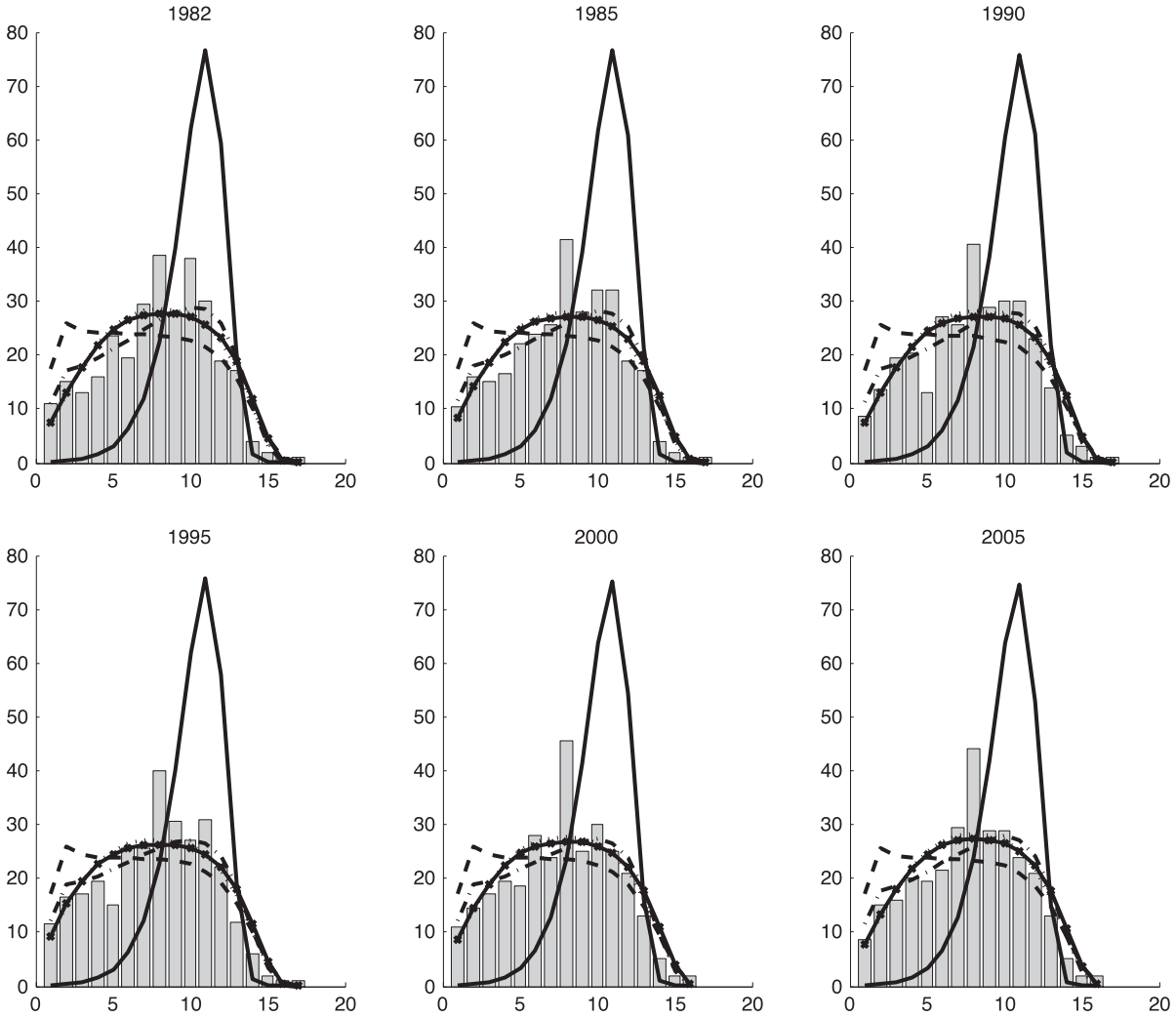

Figure 3. Fits of five models to the tree species abundance data of the BCI plots (1982–2005 censuses), for trees >1 cm in stem diameter at breast height (see table 2) The line style is the same as in figure 2.

Table 1. Maximum likelihood estimates of the five models for the six data sets for tropical forests. In the six plots coordinated by Center for Tropical Forest Science of the Smithsonian (http://www.ctfs.si.edu), we considered trees with diameter at breast height >10 cm. S is the number of species, and J is the total abundance. The subscripts in the parameters correspond to the particular model. Log-likelihood estimators L were calculated using binning-independent methods [27] (the smaller values of L correspond to better fits).

Model	1	2	3	4	5					
$\langle\varphi_n\rangle$	$\theta\frac{x^n}{n+c}$	$\theta\frac{x^n}{n}$	θx^n	$\frac{\theta}{e^{\beta(n+\mu)}-1}$	$\theta x^n\frac{\Gamma(n+g)}{\Gamma(n+1)}$					
Plot	S	J	c_1	x_1	θ_1	L_1	x_2	θ_2	L_2	
Sherman, Panama	227	21 905	0.49	0.998	39.64	314.24	0.998	35.29	315.89	
Yasuni, Ecuador	821	17 546	0.51	0.988	213.18	303.64	0.99	178.57	311.41	
Pasoh, Malaysia	678	26 554	1.95	0.993	189.5	365.31	0.995	126.74	397.02	
Korup, Cameroon	308	24 591	0.24	0.998	53.04	323.13	0.998	49.61	323.96	
Lambir, Malaysia	1004	33 175	2.02	0.991	301	391.24	0.994	195.3	442.21	
Sinharaja, Sri Lanka	167	16 936	0.38	0.998	28.26	258.52	0.998	25.73	259.34	
Plot	x_3	θ_3	L_3	β_4	μ_4	θ_4	L_4	g_5	x_5	θ_5	L_5
Sherman	0.99	2.38	433.59	0.0026	0.39	0.097	312.81	0.1	0.998	28.16	312.8
Yasuni	0.953	40.3	599.87	0.016	0.34	3.17	319.39	0.04	0.989	170.74	310.46
Pasoh	0.974	17.76	519.53	0.01	1.66	1.81	365.71	0.24	0.991	84.53	366.83
Korup	0.987	3.91	574.37	0.003	0.15	0.15	327.51	0	0.998	49.52	323.97
Lambir	0.97	31.33	620.14	0.012	1.67	3.43	400.44	0.24	0.989	135.75	402.57
Sinharaja	0.99	1.66	364.2	0.0025	0.3	0.067	258.91	0.063	0.998	22.49	258.47

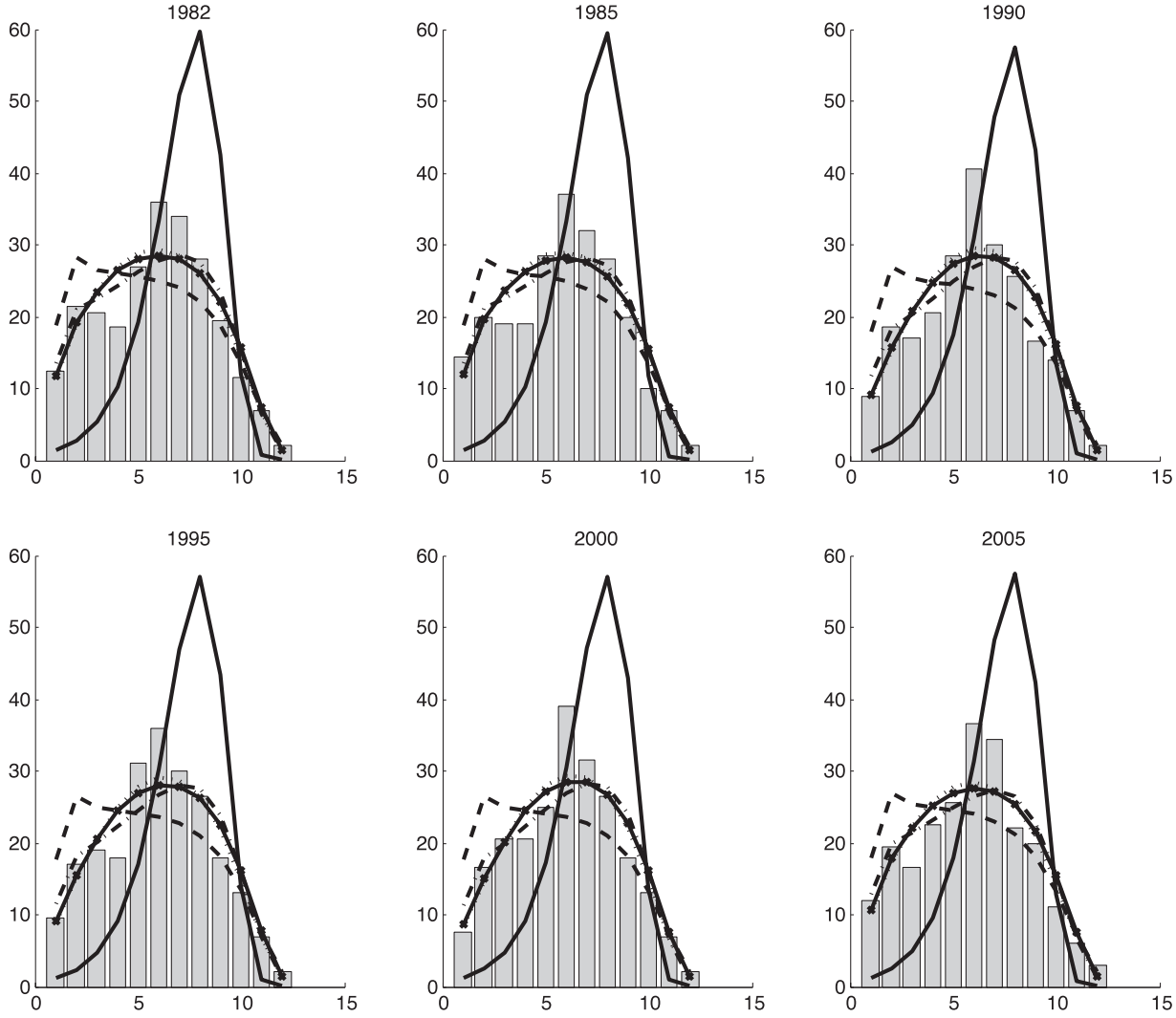

Figure 4. Fits of five models to the tree species abundance data of the BCI plots (1982–2005 censuses), for trees >10 cm in stem diameter at breast height (see table 3) The line style is the same as in figure 2.

configuration $\vec{\phi}$. Note that $k = 0$ is excluded because the species are not labeled.

We impose *three* constraints:

$$\sum_{\vec{\phi}} P(\vec{\phi}) = 1 \qquad \text{normalization} \tag{41}$$

$$\langle S \rangle = \sum_{\vec{\phi}} P(\vec{\phi}) \sum_{k \geqslant 1} \phi_k \qquad \text{average number of species} \tag{42}$$

$$\langle N \rangle = \sum_{\vec{\phi}} P(\vec{\phi}) \sum_{k \geqslant 1} \phi_k k \qquad \text{average number of individuals.} \tag{43}$$

Maximizing the entropy

$$\mathcal{H}(P) = -\sum_{\vec{\phi}} P(\vec{\phi}) \ln P(\vec{\phi}) \tag{44}$$

with the three constraints we get

$$P(\vec{\phi}) = \hat{Z}^{-1} \mathrm{e}^{-\beta \sum_{k>0} \phi_k \epsilon_k} \tag{45}$$

$$\epsilon_k = k + \mu \tag{46}$$

$$\hat{Z}(\beta, \mu) \equiv \prod_{k \geqslant 1} (1 - \mathrm{e}^{-\beta \epsilon_k}), \tag{47}$$

where β and μ are the Lagrange multipliers corresponding to the constraint equations (42) and (43), respectively. μ arises from the constraint on the average number of species, whereas β originates from the average population constraint. The above distribution is the same as that of a gas of indistinguishable particles occupying a discrete equally spaced ladder-like spectrum with the occupation number of the kth level being ϕ_k. The relative species abundance is easily calculated to be

$$\hat{P}_{\mathrm{RSA}}(n) \propto \langle \phi_n \rangle = \frac{1}{\mathrm{e}^{\beta \epsilon_n} - 1}, \tag{48}$$

and corresponds to the familiar Bose–Einstein distribution. Introducing the degeneracy g one obtains

$$\hat{P}(\phi) = \prod_{k \geqslant 1} \left\{ (1 - \mathrm{e}^{-\beta \epsilon_k})^g \mathrm{e}^{-\beta \phi_k \epsilon_k} \binom{\phi_k + g - 1}{\phi_k} \right\}, \tag{49}$$

Table 2. Maximum likelihood estimates of the five models for the six censuses of the Barro Colorado island plot. Included are the grown trees and saplings with diameter at breast height >1 cm.

Year	S	J	c_1	x_1	θ_1	L_1	x_2	θ_2	L_2		
1982	306	235 313	1.93	0.9998	42.82	735.65	0.9999	34.68	746.15		
1985	307	242 045	1.61	0.9998	41.86	726.56	0.9999	34.69	735.61		
1990	304	244 011	1.88	0.9998	42.12	725.96	0.9999	34.27	737.15		
1995	303	229 007	1.26	0.9998	40.45	709.41	0.9998	34.42	716.14		
2000	301	213 765	1.50	0.9998	41.37	686.23	0.9998	34.47	694.55		
2005	299	208 387	1.84	0.9998	42.28	686.52	0.9998	34.32	697.36		
Year	x_3	θ_3	L_3	β_4	μ_4	θ_4	L_4	g_5	x_5	θ_5	L_5
1982	0.999	0.398	1018.42	0.000 26	1.72	0.01	738.50	0.12	0.9998	23.12	736.55
1985	0.999	0.390	1031.74	0.000 25	1.44	0.01	729.41	0.10	0.9998	24.51	728.28
1990	0.999	0.379	1026.36	0.000 25	1.69	0.01	728.87	0.10	0.9998	23.78	729.19
1995	0.999	0.401	1021.88	0.000 26	1.12	0.01	712.11	0.09	0.9998	25.53	710.60
2000	0.999	0.424	991.82	0.000 28	1.33	0.01	689.05	0.09	0.9998	25.44	688.85
2005	0.999	0.430	975.43	0.000 29	1.65	0.01	689.34	0.109	0.9998	24.19	690.05

Table 3. Maximum likelihood estimates of the five models for the six censuses of the Barro Colorado Island plot. Included are the grown trees only with diameter at breast height >10 cm.

Year	S	J	c_1	x_1	θ_1	L_1	x_2	θ_2	L_2		
1982	238	20 878	1.04	0.998	46.61	315.68	0.998	37.66	320.80		
1985	237	20 712	0.96	0.998	45.87	311.19	0.998	37.53	315.54		
1990	229	21 226	1.71	0.998	48.19	309.81	0.998	35.87	319.02		
1995	227	21 442	1.68	0.998	47.38	317.43	0.998	35.43	326.26		
2000	227	21 193	1.86	0.998	48.47	307.55	0.998	35.51	318.05		
2005	229	20 852	1.16	0.998	45.21	307.12	0.998	35.98	312.72		
Year	x_3	θ_3	L_3	β_4	μ_4	θ_4	L_4	g_5	x_5	θ_5	L_5
1982	0.989	2.74	429.63	0.0034	0.90	0.15	316.44	0.13	0.997	28.63	316.11
1985	0.989	2.74	426.57	0.0033	0.82	0.15	311.90	0.12	0.997	28.97	311.32
1990	0.989	2.50	411.93	0.0033	1.49	0.15	310.70	0.16	0.997	25.20	312.05
1995	0.989	2.43	418.77	0.0032	1.46	0.14	318.16	0.16	0.997	24.83	319.30
2000	0.989	2.46	408.38	0.0033	1.63	0.15	308.51	0.16	0.997	24.61	310.66
2005	0.989	2.54	418.89	0.0032	1.00	0.14	307.87	0.13	0.997	27.24	308.11

leading to

$$\hat{P}_{\mathrm{RSA}}(n) = \langle \phi_n \rangle = \frac{g}{\mathrm{e}^{\beta\epsilon_n} - 1}, \tag{50}$$

to be compared with equation (37).

Figures 2–4 show the fits of five distinct models to empirical data. The figures show that the data set is adequately fit by more than a single model. Note that the optimal fit for Model 5 occurs for g close to 0 (less than 0.25). The derivation of equation (37) was carried out assuming a non-zero integer value of g and thus the best-fit values are worrisome. An ecological community is governed by niche effects and is characterized by interactions between species and by interactions between the species and the temporally and spatially heterogeneous environment. Quantities such as the relative species abundance can often be fit admirably using expressions derived using simple assumptions. The existence of a good fit does not of course necessarily imply that the underlying assumptions are correct. Rather, analytically tractable frameworks can be used to fit the gross patterns observed and deviations from the predictions can be used to assess what new ingredients must be added in order to understand and predict the behavior of ecological communities. We caution the reader that data-fitting exercises such as the one that we are carrying out do not, in and of themselves, determine the validity of a particular approach. At best, they provide a guide to whether a given approximation is capable of explaining the data or not.

6.2. Plant spatial distribution

The application of the maximum entropy principle to the spatial distribution of trees in a forest yields the Poisson distribution: the probability of observing n trees in a subarea a is given by $P_{\mathrm{Poisson}}(n, a) = \mathrm{e}^{-\rho a}(\rho a)^n/n!$, where ρ is the density of trees in the plot. This is a standard textbook result [29] and can be derived along the same lines as equation (15) by replacing species 1 (in the previous discussion) by a specific subarea and the remaining species by other subareas. The species–area relationship, i.e. the average number of species in the subarea a, is given by [14] $\sum[1 - P(0, a)]$, where the summation is performed over all species. The clumping can be imposed as a constraint by generalizing the maximum entropy principle to include spatial effects which would lead to a field theory approach that is analogous to the one used in both equilibrium and non-equilibrium statistical physics [30].

The Poisson distribution has the necessary property that a merger of two subareas preserves the form of the distribution with effective scale-dependent parameters, i.e.

$$P(n, a+b) = \sum_{m=0}^{n} P(m, a) P(n-m, b). \qquad (51)$$

The analysis of Harte *et al* [14] yields instead the geometric distribution: $P_{\text{geom}}(n, a) \propto [\rho a/(\rho a + 1)]^n$. The geometric distribution does not satisfy the above convolution equation and can at best strictly hold only at one length scale. Indeed, these results are based on the implicit assumption of indistinguishable (quantum) trees yielding an effective interaction between trees analogous to that responsible for Bose–Einstein condensation. In order to make the Harte *et al* model operational at all scales, one needs to supply additional information regarding the spatial correlations between the quadrats, thus introducing new constraints and further decreasing the entropy of the probability distribution.

One may compare data from the Barro Colorado island tropical forest [31] to the results obtained by Harte *et al* [14] and the Poisson distribution (figure 5). Consider a species with abundance N and partition the area into $k \times k$ quadrats. The Harte *et al* model [14] predicts that the fraction of quadrats in which at least one individual of the species is present is given by $g(a) = aN/(aN + 1)$, where $a = 1/k^2$. A random homogeneous distribution of trees (Poisson distribution) predicts a higher fraction $f_P(a) = 1 - \exp(-aN)$ even though the two distributions coincide in the limit of infinitesimal aN. Let us define $d(f, g)$, the distance between the two functions $f(a)$ and $g(a)$, to be equal to the largest difference between them, i.e. $d(f, g) = \max |f(a) - g(a)|$, with $a \in [0, 1]$. Of the 305 species in the BCI forest, there are 114 species that deviate from the Harte *et al* model more than that model deviates from the Poisson distribution. Also, one finds that there are 123 species that are closer to f_P than to f_g.

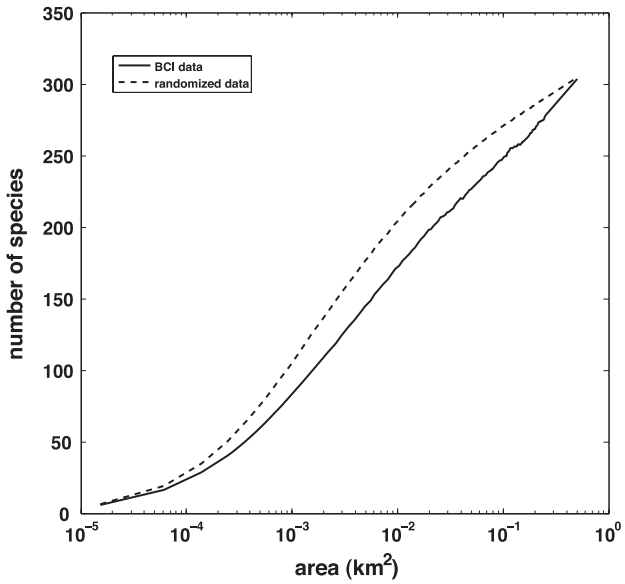

Figure 5. Species–area relationship for the BCI plot (solid line) and averages over randomized data sets (dashed line). The data randomization was carried out using an iterative procedure where at each step a random pair of trees was picked and their positions interchanged. The dashed curve is the species–area relationship averaged over 100 random plots. The key reason for the difference between the maximum entropy result (the dashed curve) and the actual data (the solid curve) is the absence of clumping in the random data. Interestingly, Harte *et al* [14] have demonstrated that the application of the principle of maximum entropy under the assumption of indistinguishable individuals yields a species–area relationship in excellent accord with data. For the random sampling of distinguishable individuals, this corresponds to weighting the random plots in a physically unjustifiable manner. Consider the species–area relationship for a plot of area A_0 with N_0 trees. One may think of an alternative randomization procedure which may be carried out as follows. Consider an initially empty subarea $A < A_0$. We envisage carrying out N_0 steps; at each step one populates the subarea with a tree with a probability $p = A/A_0$. The resulting abundance N of the subarea is distributed according to the binomial distribution: $P(N) = \frac{N_0!}{N!(N_0-N)!} p^N (1-p)^{N_0-N}$. As is well-known [29], the binomial distribution becomes a Poisson distribution in the limit of $N_0 \to \infty$ with fixed pN_0. On aggregating several subareas together, the abundance probability remains binomial. The model proposed by Harte *et al*, which yields clustering akin to that of the real data, is equivalent to another different, artificial randomization procedure in which the iteration stops as soon as a step does not lead to a new tree in the subplot. The probability distribution is then represented by the geometric distribution $P(N) \propto p^N$. Note that unlike the binomial, the geometric distribution is not preserved when several areas are aggregated together. This difference in randomization is an observable consequence of the application of the maximum entropy principle for a system with distinguishable (binomial distribution) and indistinguishable (geometric distribution) individuals.

7. Summary

The maximum entropy principle is an inference technique for constructing an estimate of a probability distribution using available information. In order to have a chance of obtaining the correct results, one must carefully choose the appropriate variables characterizing the system, one must consider issues of distinguishability or lack thereof, one must know or guess the reference entropy, and one must choose the constraints wisely. We suggest that, in order to guarantee that the results do not depend on the description level, one ought to maximize *the relative entropy* subject to the known constraints. This provides a natural interpretation of the relative entropy [23] in the context of statistical physics. In order to be successful, the method requires knowledge of the reference probability, which, in turn, depends on the system dynamics. Alternatively [6, 19], one could maximize the ordinary entropy $\mathcal{H}(\vec{P})$, equation (3), and continue to add additional constraints until one obtains the correct $\vec{P}$. In order to obtain the correct answer, in the absence of the reference entropy, one requires the knowledge of which optimal constraints to use (e.g. the constraint on the average value of $\ln n$ in the ecology illustration) or the use of a large enough number of constraints [6] to ensure convergence. Unfortunately, in general, there is no *a priori* guarantee that either of these approaches will be successful. Inspired by the ecology application, we have introduced a novel statistical ensemble with indistinguishable 'particles' and indistinguishable 'levels' yielding a distribution corresponding to that of a quantum oscillator.

Acknowledgments

We are indebted to Sandro Azaele and Sikai Zhu for collaborating with us on related research. We are grateful to Roderick Dewar and John Harte for insightful discussions. This work was supported by Fondazione Cariparo—Padova. The BCI forest dynamics research project was made possible by National Science Foundation grants to Stephen P Hubbell, support from the Center for Tropical Forest Science, the Smithsonian Tropical Research Institute, the John D and Catherine T MacArthur Foundation, the Mellon Foundation, and the Celera Foundation.

Appendix A. Properties of the entropy, $\mathcal{H}$

We summarize here some of the key properties of $\mathcal{H}$, the entropy of the 'distribution' probability $\vec{P} \equiv (P_1, P_2, \ldots, P_E)$. This section is not strictly essential for the understanding of the rest of the paper.

(1)

$$\mathcal{H}(\vec{P}) \geqslant 0. \tag{52}$$

This follows from the observation that, in equation (3), $0 \leqslant P_i \leqslant 1$ and $\lim_{x\to 0^+} x \ln x = 0$.

(2) $\mathcal{H}(\vec{P})$ is a concave function.

A function $f(x)$ of the real variable x is said to be convex (i.e. $-f(x)$ is concave) if

$$f(\mu x_1 + (1-\mu)x_2) \leqslant \mu f(x_1) + (1-\mu) f(x_2) \tag{53}$$

is valid for all pairs x_1, x_2 and $0 \leqslant \mu \leqslant 1$. If equation (53) holds as an equality only when $\mu = 0, 1$, f is said to be strictly convex. It is easy to show by induction that equation (53) implies

$$f\left(\sum_i \mu_i x_i\right) \leqslant \sum_i \mu_i f(x_i), \qquad \mu_i \geqslant 0, \quad \sum_i \mu_i = 1. \tag{54}$$

Before proving the concavity of $\mathcal{H}(\vec{P})$, we also need the following result: if $f(x)$ has a second derivative $f''(x) \geqslant 0 (>0)$, then f is convex (strictly convex). Indeed, on using Taylor's theorem,

$$f(y) = f(x) + f'(x)(y-x) + f''(\xi)\frac{(y-x)^2}{2},$$

where ξ is a suitable value in between x and y. Thus $f(y) \geqslant f(x) + f'(x)(y-x)$ because $f''(\xi) \geqslant 0$. Taking $x = \mu x_1 + (1-\mu)x_2$ and $y = x_1$ and x_2,

$$f(x_1) \geqslant f(x) + (1-\mu)(x_1 - x_2) f'(x),$$

$$f(x_2) \geqslant f(x) + \mu(x_2 - x_1) f'(x)$$

and equation (53) follows. Strict inequality implies strict convexity. Because $-\mathcal{H}(\vec{P})$ is a sum of convex functions $f(x) = x \ln x$ ($f''(x) = 1/x > 0$ if $x > 0$), it is itself a convex function, i.e. $\mathcal{H}(\vec{P})$ is concave:

$$\mathcal{H}(\mu \vec{P}_1 + (1-\mu)\vec{P}_2) \geqslant \mu \mathcal{H}(\vec{P}_1) + (1-\mu)\mathcal{H}(\vec{P}_2). \tag{55}$$

(3) $\mathcal{H}(\vec{P})$ has only one maximum if the constraints are linear in P. This result follows from property (2) above and does not depend on its specific expression, equation (3). However using equation (3), one sees that the matrix of second derivatives, $\partial^2 \mathcal{H}(\vec{P})/\partial P_i \partial P_j = -\delta_{i,j}/P_i$, is negative definite and so at most one maximum exists.

In the case of just one constraint (6), one can easily see that

$$\mathcal{H}(\vec{P}_0) > \mathcal{H}(\vec{P}) \qquad \forall \vec{P} \neq \vec{P}_0, \tag{56}$$

where $P_{0i} = 1/E$ is the uniform distribution. In order to obtain this result, let $f(x) = x \ln x$. Then

$$-\frac{\mathcal{H}(\vec{P}_0)}{E} = -\frac{\ln E}{E} = f\left(\frac{1}{E}\right) = f\left(\frac{\sum_i P_i}{E}\right) < \sum_i \frac{1}{E} f(P_i) = -\frac{\mathcal{H}(\vec{P})}{E} \Rightarrow (56).$$

Thus the uniform distribution has the maximum entropy in the absence of constraints.

(4) Why is the most probable distribution interesting and what is the utility of the entropy? The answer is, in part, contained in the 'concentration theorem' of Jaynes [7].

First, let us observe that if $\vec{P}$ is the distribution which maximizes the entropy $\mathcal{H}_{\max} = \mathcal{H}(\vec{P}) > \mathcal{H}(\vec{P}')$, $\vec{P} \neq \vec{P}'$. The number of times that we obtain the distribution $\vec{n}' \equiv (n'_1, n'_2, \ldots) = N\vec{P}'$ compared to the corresponding number in which $\vec{n} \equiv (n_1, n_2, \ldots) = N\vec{P}$ is observed is given by

$$\frac{W(\vec{n})}{W(\vec{n}')} \propto \mathrm{e}^{N(\mathcal{H}_{\max} - \mathcal{H}(P'))}[1 + \mathrm{O}(1/N)], \tag{57}$$

where the exponential follows from (3). The concentration theorem says that the fraction of the distributions $\vec{P}'$ such that

$$N(\mathcal{H}_{\max} - \mathcal{H}(P')) \equiv N\Delta\mathcal{H} = x \tag{58}$$

is given by

$$P(x) = \frac{x^b \mathrm{e}^{-x}}{\Gamma(b+1)}, \tag{59}$$

where $b = \frac{E-m-2}{2}$ for m constraints, including the normalization ($m = 2$ if we have only the constraint equations (6) and (eq:1.8.1)). Since $x = N\Delta\mathcal{H} = \sum_i \frac{\Delta P_i^2}{2P_i} N$ ($\Delta P_i \equiv P'_i - P_i$), due to the exponential decay in (59), the most relevant distributions, P_i, are such that

$$|\Delta P_i| \lesssim \sqrt{\frac{P_i}{N}} \tag{60}$$

which is equivalent to the well-known result that

$$\frac{\Delta n_i}{N} \lesssim \frac{1}{\sqrt{N}}. \tag{61}$$

Appendix B. A primer on Lagrange multipliers

A standard trick which is used to maximize/minimize a function subject to constraints is to introduce Lagrange multipliers. Suppose we want to determine the maximum of $f(x, y) = -x^2 - y^2$ with the constraint $\varphi(x, y) = y + x = 2$.

Of course this can be done immediately by finding $y = 2 - x$ from the constraints and maximizing $f(x, 2 - x) = -2x^2 + 4x - 4$ with respect to x. This gives $x = 1$, $y = 1$, $f(1, 1) = -2$. However, in practice it is not easy/convenient to eliminate some of the variables from the constraints. Rather one introduces a new function

$$F(x, y) = f(x, y) + \lambda\varphi(x, y) \tag{62}$$

and maximizes/minimizes it with respect to both x and y as they were independent variables not subject to constraints. The parameter λ is the so-called Lagrange multiplier. The equations to be solved for the maximum/minimum are

$$0 = \frac{\partial F}{\partial x} = -2x + \lambda, \tag{63}$$

$$0 = \frac{\partial F}{\partial y} = -2y + \lambda \tag{64}$$

which give x and y as a function of the parameter λ. This free parameter is then used in order to satisfy the constraints $2 = x + y = \lambda$ which immediately leads to the exact answer $x = y = 1$. Even if rather trivial, this example illustrates the general method.

References

[1] Boltzmann L 1964 *Lectures on Gas Theory* (London: Cambridge University Press)
[2] Shannon C E 1948 *Bell Syst. Tech. J.* **27** 379–423
[3] Jaynes E T 1957 *Phys. Rev.* **106** 620–30
[4] Jaynes E T 1957 *Phys. Rev.* **108** 171–90
[5] Shore J E and Johnson R W 1980 *IEEE Trans. Inf. Theory* **26** 26–37
[6] Mead L R and Papanicolaou N 1984 *J. Math. Phys.* **25** 2408
[7] Jaynes E T 2003 *Probability Theory* (London: Cambridge University Press) p 375
[8] Dewar R C 2003 *J. Phys. A: Math. Gen.* **36** 631–41
[9] Dewar R C 2005 *J. Phys. A: Math. Gen.* **38** L371–81
[10] Whitfield J 2005 *Nature* **436** 905–7
[11] Dunkel J, Talkner P and Hanggi P 2007 *New J. Phys.* **9** 144
[12] Caticha A 2008 Lectures on probability, entropy and statistical physics arXiv:0808.0012
[13] Shipley B, Vile D and Garnier E 2006 *Science* **314** 812–4
[14] Harte J, Zillio T, Conlisk E and Smith A 2008 Maximum entropy and the state variable approach to macroecology *Ecology* **89** 2700–11
[15] Dewar R C and Porte A 2008 *J. Theor. Biol.* **251** 389–403
[16] Sibisi S, Skilling J, Brereton R G, Laue E D and Staunton J 1984 *Nature* **311** 446–7
[17] Kitaura R *et al* 2002 *Science* **298** 2358–61
[18] Dong W *et al* 1992 *Nature* **355** 605–9
[19] Schneidman E, Berry M J, Segev R and Bialek W 2006 *Nature* **440** 1007–12
[20] Cover T M and Thomas J A 2006 *Elements of Information Theory* 2nd edn (New York: Wiley)
[21] Berger A L, Della-Pietra S A and Della-Pietra V J 1996 A maximum entropy approach to natural language processing *Comput. Linguist.* **22** 39–71
[22] Vapnyarskii I B 2001 Lagrange multipliers *Encyclopaedia of Mathematics* ed M Hazewinkel, (Dordrecht: Kluwer Academic) http://eom.springer.de/L/l057190.htm
[23] Kullback S 1959 *Information Theory and Statistics* (New York: Wiley)
[24] Banavar J R and Maritan A 2007 The maximum relative entropy principle arXiv:cond-matt/0703622v1
[25] Feynman R P, Leighton R B and Sands M 1970 *The Feynman Lectures on Physics* vol 3 (Reading, MA: Addison-Wesley)
[26] Hubbell S P 2001 *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton, NJ: Princeton University Press)
[27] Volkov I, Banavar J R, He F, Hubbell S P and Maritan A 2005 *Nature* **438** 658–61
[28] Fisher R A, Corbet A S and Williams C B 1943 *J. Anim. Ecol.* **12** 42–58
[29] Sivia D S 1996 *Data Analysis: A Bayesian Tutorial* (Oxford: Oxford University Press)
[30] Kardar M 2007 *Statistical Physics of Fields* (Cambridge: Cambridge University Press)
[31] Hubbell S P, Condit R and Foster R B 2005 *Barro Colorado Forest Census Plot Data* available at http://ctfs.si/edu/datasets/bci