Similarity-Reduced Diversities: the Effective Entropy and the Reduced Entropy

François Bavaud¹

Accepted: 26 July 2021 / Published online: 8 September 2021
© The Author(s) 2021

Abstract

The paper presents and analyzes the properties of a new diversity index, the effective entropy, which lowers Shannon entropy by taking into account the presence of similarities between items. Similarities decrease exponentially with the item dissimilarities, with a freely adjustable discriminability parameter controlling various diversity regimes separated by phase transitions. Effective entropies are determined iteratively, and turn out to be concave and subadditive, in contrast to the reduced entropy, proposed in Ecology for similar purposes. Two data sets are used to illustrate the formalism, and underline the role played by the dissimilarity types.

Keywords Similarity-reduced diversity · Confusion matrix · Rate distortion function · Phase transitions · Rao quadratic entropy

1 Introduction

Evaluating the variety or diversity of a collection of items occurring with unequal frequencies is a much explored domain of data analysis, with well-known proposals such as the logarithmic Shannon entropy, or Rényi or Tsallis power generalizations (see, e.g., Masi, 2005; Jost, 2006). In those *plain* or *naive* approaches, items are implicitly meant as completely dissimilar from each other. Yet, examples abound in which some pairs of items are *similar* to some extent, thus arguably reducing the corresponding naive diversities. The quest for formally sound and fruitfully interpretable *similarity-dependent* measures of diversity has been explicitly formulated and investigated in the ecological literature (see, e.g., Champely & Chessel, 2002; Pavoine et al., 2005; Leinster & Cobbold, 2012; Chiu et al., 2014; Marcon, 2016 and references therein).

This paper proposes an original similarity-dependent logarithmic measure of relative diversity (6), we shall refer to as the *effective entropy*. It depends upon the relative

✉ François Bavaud
fbavaud@unil.ch

¹ Department of Language and Information Sciences, Institute of Geography and Sustainability
University of Lausanne, Lausanne, Switzerland

frequencies and dissimilarities between items, as well as on a freely adjustable parameter controlling for the *discriminability* between similar items. Its construction can be motivated by a simple perceptual rationale: if two items are similar enough, and if the discriminability parameter is low enough, then one item can wrongly be perceived as another, hence lowering the diversity of the item collection. Further elaborating on this idea leads to a statistical mechanical framework (minimization of the free energy) whose associated formalism turns out to overlap in large parts with the *rate distortion setup* in information theory (e.g., Cover & Thomas, 2006) or the *regularized optimal transportation* (e.g., Cuturi, 2013) — with a completely different motivation and purpose.

The resulting effective entropy can be computed iteratively in general (Section 2, Theorems 1 and 2), sometimes in a single step for particular cases. One of the salient features of the framework is the apparition of *phase transitions* when lowering the discriminability parameter, from the *low-temperature regime* where all the items are perceived, at least partially, to the *high-temperature regime*, where only the *dominating items* persist (Section 3, Theorems 3, 4 and 5).

Two examples (copper concentration data, Section 3.1; world cities, Section 4.1) illustrate the formalism and its versatility, as well as the influence of the nature (city-block, Euclidean, ultrametric) of the item dissimilarities (Theorems 6 and 7). *Similarities* can also be considered directly, as far they are non-negative, symmetric, and endowed with a unit dominating diagonal.

The effective entropy (6) turns out to provide an (apparently close) upper bound to another logarithmic measure of diversity (12) considered in Ecology and computable in a single step, referred in this paper to as the *reduced entropy*. The presence of entirely similar items with distinct labels should not increase the diversity — a property automatically satisfied by the effective and the reduced entropies, in contrast to Shannon entropy. Also, the former satisfy the so-called monotonicity and modularity properties (Theorem 8). Both take on their maximum values for non-uniform item distributions in general (Theorem 10), and converge to Shannon entropy in the *naïve limit* of identity similarities, as expected.

Yet, the *concavity property*, asserting that the diversity of the whole should not be lower than the average diversity of its parts, is in general satisfied by the effective entropy only (Theorem 9). Also, like Shannon entropy, the joint effective entropy is maximum for independent distributions (Theorem 11), a property violated again in general for the reduced entropy. Remedying the formal defects of the latter provided the initial motivation of the present study.

2 The Formalism

Basic ingredients consist of n items, objects or species, with relative frequencies $f_i > 0$ normalized to $\sum_{i=1}^n f_i = 1$. In addition, the differences between items are specified by a finite $n \times n$ matrix of *dissimilarities* $\mathbf{D} = (d_{ij})$ obeying

$$d_{ij} \geq 0 \quad (1a)$$

$$d_{ij} = d_{ji} \quad (1b)$$

$$d_{ii} = 0 \quad (1c)$$

$$d_{ij} = 0 \Rightarrow d_{ik} = d_{jk}, \quad (1d)$$

where (1d) defines an *even* or *semi-proper* dissimilarity (see, e.g., Critchley & Fichet, 1994).

2.1 Transition Matrices and Percept Weights

Let z_{ij} denote the probability that item i is perceived or received as item j . According to the context, the pair i - j can be referred to as the stimulus-percept, stimulus-response, origin-destination, input-output, symbol-reproduction, or source-estimate pair. By construction, $z_{ij} \geq 0$ and $z_{i\bullet} = \sum_j z_{ij} = 1$ (where “ $\bullet$ ” denotes summation over the replaced index). The set of all such memberships or *transition matrices* $\mathbf{Z} = (z_{ij}) \in \mathbb{R}^{n \times n}$ will be denoted as $\mathcal{Z}$.

The *weight of percept* j is obtained as

$$\rho_j = \sum_i f_i z_{ij} \geq 0 \quad (2)$$

and obeys $\rho_\bullet = 1$ by construction.

2.2 Energy, Mutual Information, and Free Energy

The average stimulus-percept dissimilarity, cost or *energy* is

$$U[\mathbf{Z}] = \sum_{i,j} f_i z_{ij} d_{ij} \geq 0, \quad (3)$$

whose minimum is $\mathbf{Z} = \mathbf{I}_n = (\delta_{ij})$, the identity matrix (no stimulus-percept confusion). This minimum is unique if $\mathbf{D}$ is *proper*, that is if $d_{ij} > 0$ for $i \neq j$.

By contrast, the confusion is maximum when the percept is independent of the stimulus, that is when the joint stimulus-percept distribution $f_i z_{ij}$ is equal to the corresponding independent distribution $f_i \rho_j$, or equivalently when $z_{ij} = \rho_j$ for the distribution ρ in (2). Such an independent transition minimizes the stimulus-percept *mutual information*, which reads here as

$$K[\mathbf{Z}] = \sum_{i,j} f_i z_{ij} \ln \frac{z_{ij}}{\rho_j} \geq 0. \quad (4)$$

The above antagonist tendencies can be combined in the (dimensionless) *free energy*

$$F[\mathbf{Z}] = \beta U[\mathbf{Z}] + K[\mathbf{Z}], \quad (5)$$

where $\beta > 0$ is a free *discriminability* parameter, known in statistical mechanics as the *inverse temperature* $\beta = 1/T$. In the *low-temperature limit* $\beta \rightarrow \infty$, only the energy term contributes and the optimal transition matrix $\mathbf{Z}$ tends to $\mathbf{I}_n$ (perfect discrimination). In the *high-temperature limit* $\beta \rightarrow 0$, only the entropy term contributes and $\mathbf{Z} \rightarrow \mathbf{1}_n \rho'$ (complete confusion).

Note that in physics, the free energy is generally defined as $U + \beta^{-1}K$ instead of F in (5). The latter choice, without incidence on the form of the optimal transition matrix $\mathbf{Z}$, has the advantage of making the effective entropy (defined below) dimensionless, and directly comparable to Shannon entropy. Also, $\mathbf{D}$, U and β^{-1} have the dimensions of an energy; they can be made dimensionless by a proper rescaling of $\mathbf{D}$ (see Section 3.1).

2.3 Effective Entropy E and Reduced Entropy R

The *effective entropy*, whose study constitutes the main scope of the paper, is defined as

$$E \equiv E(f, \mathbf{D}, \beta) = \min_{\mathbf{Z} \in \mathcal{Z}} F[\mathbf{Z}]. \quad (6)$$

The optimal transition $\mathbf{Z}$ minimizing $F[\mathbf{Z}]$ is determined by the non-linear equation (see Theorem 1 below)

$$z_{ij} = \frac{s_{ij}\rho_j}{\tau_i}, \quad \text{where} \quad s_{ij} = e^{-\beta d_{ij}}, \quad (7)$$

$$\rho_j[\mathbf{Z}] = \sum_i f_i z_{ij} \quad \text{and} \quad \tau_i[\mathbf{Z}] = \sum_j s_{ij} \rho_j. \quad (8)$$

The components of the $n \times n$ similarity matrix $\mathbf{S} = (s_{ij})$ in (7) obey

$$s_{ij} \geq 0 \quad (9a)$$

$$s_{ij} = s_{ji} \quad (9b)$$

$$s_{ii} = 1 \quad (9c)$$

$$s_{ij} = 1 \Rightarrow s_{ik} = s_{jk}. \quad (9d)$$

Similarities (9) are related to dissimilarities as $\mathbf{S} = \exp(-\beta \mathbf{D})$, to be understood as the *componentwise* or *Hadamard exponential*. In terms of similarities, the free energy (5) reads, by direct substitution, as

$$F[\mathbf{Z}] = \sum_{i,j} f_i z_{ij} \ln \frac{z_{ij}}{s_{ij}\rho_j} \geq 0. \quad (10)$$

Let $\kappa(\boldsymbol{\rho}||\mathbf{f}) = \sum_j \rho_j \ln(\rho_j/f_j) \geq 0$ denote the Kullback-Leibler divergence between percept and stimulus weights, and consider the functional

$$G[\mathbf{Z}] = F[\mathbf{Z}] + \kappa(\boldsymbol{\rho}||\mathbf{f}) = \sum_{i,j} f_i z_{ij} \ln \frac{z_{ij}}{s_{ij}f_j} \geq 0. \quad (11)$$

Its minimizer $\mathbf{Z}^0$, and minimum value $R = G[\mathbf{Z}^0]$, we shall refer to as the *reduced entropy*, are readily found (in a single step) to be

$$z_{ij}^0 = \frac{s_{ij}f_j}{b_i}, \quad \text{where} \quad b_i = \sum_j s_{ij}f_j, \text{ and} \quad (12a)$$

$$R = \min_{\mathbf{Z} \in \mathcal{Z}} G[\mathbf{Z}] = - \sum_i f_i \ln b_i \quad (\text{reduced entropy}). \quad (12b)$$

Quantity b_i in (12) is the *banality* of item i , measuring its average similarity to other items (Marcon, 2016), proposed by Leinster and Cobbold (2012), as well as by Ricotta and Szeidl (2006) in the variant $s_{ij} = 1 - d_{ij}/d_{\max}$. By construction,

$$0 \leq E \leq R \leq H, \quad \text{where} \quad H = - \sum_i f_i \ln f_i \quad (\text{Shannon entropy}). \quad (13)$$

The lower bound for R follows from (11), and the upper bound for R from $f_i \leq b_i \leq 1$. Also, substituting (7) in (10) yields the expression

$$E = - \sum_i f_i \ln \tau_i. \quad (14)$$

The normalization factor τ_i in (7) and (8) has a form similar to b_i in (12): they both measure the average similarity, where the average is taken on $\mathbf{f}$ for b_i , but on $\boldsymbol{\rho}$ for τ_i . One can refer to b_i as the source or *stimulus banality*, and to τ_i as the outcome or *percept banality*.

2.4 High- and Low-Temperature Similarities: Aggregation of Equivalent Items

In the high-temperature limit $\beta \rightarrow 0$, $\mathbf{S} \rightarrow \mathbf{J}_n$ (the $n \times n$ unit matrix filled with ones), and thus $\mathbf{b} \rightarrow \mathbf{1}_n$ (the unit vector) and finally $E, R \rightarrow 0$.

In the low-temperature limit $\beta \rightarrow \infty$, $s_{ij} \rightarrow 0$ for $i \neq j$, unless $d_{ij} = 0$: recall from (1d) that $\mathbf{D}$ is supposed to be semi-proper only, which implies that the relation $i \sim j \Leftrightarrow d_{ij} = 0$ is an equivalence relation. As a result, $\lim_{\beta \rightarrow \infty} s_{ij} = 1$ if i and j belong to the same equivalence class C , and $\lim_{\beta \rightarrow \infty} s_{ij} = 0$ otherwise. Hence, $\mathbf{S} = \bigoplus_C \mathbf{J}_C$ (direct sum of matrices) is block diagonal, and

$$\lim_{\beta \rightarrow \infty} R = H_{\text{agg}} = - \sum_C F_C \ln F_C, \quad \text{where} \quad F_C = \sum_{i \in C} f_i. \quad (15)$$

See the Appendix for a proof of (15). In other terms, in the low-temperature limit, the reduced entropy *automatically aggregates* equivalent items (i.e., whose dissimilarity is zero), yielding a corresponding entropy $H_{\text{agg}} \leq H$. The same limit holds for the effective entropy.

Similar considerations hold for the effective and reduced entropies at any temperature: items equivalent in the sense $d_{ij} = 0$ can be first aggregated into equivalence classes, yielding *proper dissimilarities* between classes, on which the above formalism can then be applied. From now on, we consider proper dissimilarities only, i.e., such that $d_{ij} > 0$ for $i \neq j$.

2.5 Behavior and Computation of the Optimal Solution

Depending on β , the percept weights ρ_j may be positive or zero, a circumstance causing the phase transitions observed in the case studies below. The *effective variety*

$$v_{\text{eff}} = \sum_{j=1}^n I(\rho_j > 0) \quad (16)$$

counts the number of detected percepts, and ranges from 1 to n .

Let $\langle n \rangle = \{1, \dots, n\}$ denote the collection of items, and for $A \subseteq \langle n \rangle$, let $\mathcal{Z}_A$ be the set of transitions whose non-zero percept weights are exactly those of the *set of occurring percepts* A , that is $\rho_j > 0$ for $j \in A$, and $\rho_j = 0$ for $j \notin A$. Equivalently, $z_{\bullet j} > 0$ for $j \in A$, and $z_{\bullet j} = 0$ for $j \notin A$.

Theorem 1 (first-order condition) *Define the sub-indicator*

$$c_j = \sum_i \frac{f_i s_{ij}}{\tau_i}. \quad (17)$$

A necessary and sufficient condition for $\mathbf{Z}$ to be the minimizer of $F[\mathbf{Z}]$ is the first-order condition

$$z_{ij} = \frac{s_{ij} \rho_j}{\tau_i} \quad \text{where} \quad \tau_i = \sum_j \rho_j s_{ij} \quad \text{and} \quad c_j \leq 1 \quad \text{when} \quad \rho_j = 0. \quad (18)$$

Also, $c_j = 1$ when $\rho_j > 0$.

In particular, the first identity in (18) is necessary and sufficient when all $\rho_j > 0$. Also, $c_j \rho_j = \rho_j$. Identity $c_j = 1$ is a necessary condition for the apparition of percept j , whence the name “sub-indicator”.

Theorem 2 (iterative solution) *The series of iterations $\mathbf{Z}^{(0)} \rightarrow \boldsymbol{\rho}^{(0)} \rightarrow \mathbf{Z}^{(1)} \rightarrow \boldsymbol{\rho}^{(1)} \rightarrow \dots$ with*

$$\rho_j^{(t)} = \sum_i f_i z_{ij}^{(t)} \quad z_{ij}^{(t+1)} = \frac{s_{ij} \rho_j^{(t)}}{\tau_i^{(t)}} \quad \text{where} \quad \tau_i^{(t)} = \sum_j s_{ij} \rho_j^{(t)} \quad (19)$$

converge towards the unique minimizer (7) of $F[\mathbf{Z}]$, provided $\mathbf{Z}^{(0)} \in \mathcal{Z}_{(n)}$, the set of transitions with strictly positive row margins.

3 Temperature Regimes and Illustrations

3.1 Example A: Copper Concentration Data

The R dataset `chem{MASS}` consists of 24 univariate copper concentrations x_i , on which dissimilarities are defined as $d_{ij} = |x_i - x_j|$. The values 2.20 2.20 2.40 2.40 2.50 2.70 2.80 2.90 3.03 3.03 3.10 3.37 3.40 3.40 3.40 3.50 3.60 3.70 3.70 3.70 3.70 3.77 5.28 28.95 contain ties occurring twice ($x = 2.20, 2.40, 3.03$), three times ($x = 3.40$) and four times ($x = 3.70$). They must be preliminarily aggregated to define a proper dissimilarity (Section 2.4), resulting in a sample of $n = 16$ observations with non-uniform weights f , namely with values $1/24$ for the unique values, $2/24$ for values with two ties in the original dataset, $3/24$ for values with three ties, and $4/24$ for values with four ties.

Temperatures and dissimilarities always appear in combination βd_{ij} . For comparison sake, the temperature scale will, here and in the sequel, be fixed by further dividing the dissimilarities by the quantity

$$\Delta = \frac{1}{2} \sum_{ij} f_i f_j d_{ij} \quad (20)$$

which corresponds to the *inertia* of the configuration when $\mathbf{D}$ is squared Euclidean, and is referred to as *Rao quadratic entropy* in Ecology (see, e.g., Champely & Chessel, 2002; Pavoine et al., 2005 and references therein). Equivalently, the dissimilarity scale is normalized to $\Delta = 1$.

Figure 1 depicts the effective variety (16), effective entropy (6) and reduced entropy (12b) as a function of the inverse temperature. As expected, the number of detected percepts *tends* to decrease with the temperature, yet with possible monotonicity violations (disappearances followed by re-emergence of percepts $\rho_j > 0$, as detailed in Figs. 2 and 3). Also, the effective entropy appears closely approximated by the more directly computable reduced entropy.

The left panel of Fig. 2 depicts the *rate distortion function* of information theory. In this framework (Cover & Thomas, 2006, chapter 10), source i is sent through a communication channel, and decoded as j . The mutual information K or *channel rate* between sources and their reconstruction is a measure of their dependence, reaching its maximum H (for fixed f) for the perfect transmission $\mathbf{Z} = \mathbf{I}$, in which case the *distortion* U is zero (Section 3.3). Conversely, a zero-rate or random channel ($K = 0$) generates a distortion of at least $U = \partial^*$ (Section 3.4).

The right panels of Figs. 2 and 3 depict the sub-indicators $c_j(\beta)$ (17) for $j = 1, \dots, 16$. By construction, the effective variety is bounded above by the number of sub-indicators equal to one. Actually, the two quantities happen to numerically *coincide* in this case study,

Fig. 1 Left: effective variety v_{eff} ; right: effective entropy $E(\beta)$ and reduced entropy $R(\beta)$, as a function of the inverse temperature β

Fig. 2 Left: rate distortion function, plotting the mutual information $K(\beta)$ as a function of the energy $U(\beta)$. Right: plot of the sub-indicators $c_j(\beta)$ (17) for $j = 1, 2, 3, 4$

Fig. 3 Plot of the sub-indicators $c_j(\beta)$ (17) for $j = 5, 6, 7, 8$ (left), $j = 9, 10, 11, 12$ (middle) and $j = 13, 14, 15, 16$ (right)

in the sense $\sum_{j=1}^n I(\rho_j > 10^{-5}) = \sum_{j=1}^n I(c_j > 1 - 10^{-5})$. Said otherwise, $\rho_j > 0$ iff $c_j = 1$, and the sub-indicator c_j turns out to behave here as an indicator.

3.2 High- and Low-Temperature Regimes

The previous sections suggest the existence of two critical temperatures β_H and β_L , with $0 < \beta_H < \beta_L < \infty$, such that

1. for $\beta > \beta_L$, all $\rho_j > 0$ (that is $\mathbf{Z} \in \mathcal{Z}_{(n)}$): *low-temperature phase or regime*
2. for $\beta < \beta_H$, $\rho_j > 0$ iff j belongs to some subset $A^* \subset \langle n \rangle$, the *dominating set: high-temperature regime*
3. for $\beta_H < \beta < \beta_L$, the set A of occurring percepts with $\rho_j > 0$ varies, generally non monotonically with β , from A^* to $\langle n \rangle$: *medium-temperature regime*.

3.3 Low-Temperature Regime

For $\beta \rightarrow \infty$, $\mathbf{Z} = \mathbf{I}$ and $\rho_j \rightarrow f_j > 0$. By continuity, for β large enough, all sub-indicators (17) are equal to one, and the inverse similarity matrix, noted $\mathbf{S}^{-1} = (s^{ij})$, exists (recall $\lim_{\beta \rightarrow \infty} \mathbf{S} = \mathbf{I}$ for $\mathbf{D}$ proper).

Theorem 3 (low-temperature regime) *Assume that $\mathbf{S}^{-1} = (s^{ij})$ exists, and define the (possibly non positive) signed weights by*

$$r_j := \sum_{i=1}^n f_i \frac{s^{ij}}{s^{i\bullet}}. \quad (21)$$

Then all percepts emerge (that is $\rho_j > 0$ for all j) iff $r_j > 0$. In this case, the optimal solution (7), (8) is given by

$$\rho_j = r_j, \quad \tau_i = \frac{f_i}{s^{i\bullet}}, \quad z_{ij} = \frac{s^{i\bullet} s_{ij} r_j}{f_i} \quad \text{and} \quad E = H + \sum_i f_i \ln s^{i\bullet}(\beta). \quad (22)$$

Note that $\sum_j r_j = 1$ always holds, whence the name signed weights. Thus the condition $\min_j r_j \geq 0$, characterizing the low-temperature regime, also implies $\max_j r_j \leq 1$; see the left panels of Figs. 4, 7, 8, 9, 10, 11, 12, 13 and 14.

The invertibility of $\mathbf{S}$ for all $\beta > 0$ holds for a large class of dissimilarities, namely squared Euclidean proper dissimilarities (Theorem 6b). Yet, the above theorem does not rule out the possible coexistence of all percepts together with a non-invertible $\mathbf{S}$, although such a case has not been met in our investigations.

3.4 High-Temperature Regime

For $\beta = 0$, the free energy consists of $K[\mathbf{Z}]$ only, which is minimized (and equal to zero) by the independent transition $z_{ij} = \rho_j$, where ρ is an arbitrary distribution. Let us introduce the definitions

$$\partial_j = \sum_i f_i d_{ij}, \quad \bar{\partial} = \sum_j f_j \partial_j = 2\Delta, \quad (23a)$$

$$\partial^* = \min_j \partial_j \quad \text{and} \quad A^* = \{j \mid \partial_j = \partial^*\}, \quad (23b)$$

Fig. 4 Left: the low-temperature regime $\beta > \beta_L$ is characterized by $\min_j r_j \geq 0$, and hence $r_j = \rho_j$ (Theorem 3). Right: histogram of ρ_j in the high-temperature limit: the mass is evenly concentrated on the dominating states $A^* = \{9, 10\}$, that is $\rho_9 = \rho_{10} = 0.5$

appearing in the study of the high-temperature regime. The *eccentricity* ∂_j measures the (weighted) *average dissimilarity* between item j and the other items in the configuration (f, D) , and obeys $\partial_j = d_{jf} + \Delta$ when D is squared Euclidean, where d_{jf} is the dissimilarity between item j and the barycenter of the configuration (Huygens principle). The average eccentricity $\bar{\partial}$ is twice the Rao entropy (20), and is equal to 2 under the normalization $\Delta = 1$. *Medoids* are items with minimal eccentricity ∂^* . The collection of medoids constitutes the *dominating set* A^* .

In the copper concentration data, the dominating set A^* contains the minimizers of the median average deviation $\partial_j = \sum_i f_i |x_i - x_j|$, and consists of the two observations $j = 9$ and $j = 10$ (see Fig. 4). In the world cities dataset with geodesic distances (Section 4.1), A^* consists of Berlin only.

For $\beta \rightarrow 0$, the energy becomes $U[Z] = \sum_j \rho_j \partial_j$, which is minimum iff the non-zero components or *support* $\text{supp}(\rho)$ of ρ belong to the dominating set A^* . This dominating set persists in the high-temperature regime $0 < \beta \ll 1$:

Theorem 4 (high-temperature regime) *For $0 < \beta \ll 1$, the free energy is minimum for*

$$z_{ij} = \frac{s_{ij} \rho_j}{\sum_{k \in A^*} s_{ik} \rho_k} \quad \text{where} \quad \text{supp}(\rho) \subset A^*. \quad (24)$$

Also, the effective and relative entropies obey

$$E(\beta) = \beta \partial^* + O(\beta^2) \quad \text{and} \quad R(\beta) = \beta \bar{\partial} + O(\beta^2). \quad (25)$$

The dominating set is stable in the sense $c_j(\beta=0) = 1$ for all j , with derivatives $c'_j(\beta=0) = 0$ for $j \in A^*$ and $c'_j(\beta=0) < 0$ for $j \notin A^*$, which implies (Theorem 1) $\rho_j(\beta) = 0$ for $0 < \beta \ll 1$ and $j \notin A^*$.

Inequality $\partial^* \leq \bar{\partial}$ reflects the general relation $E(\beta) \leq R(\beta)$. Also, when A^* contains a single element, the relation $E(\beta) = \beta \partial^*$ is exact in the high-temperature regime.

3.5 Medium-Temperature Regime

The medium-temperature regime, neither characterized by the presence of all percepts nor the sole presence of dominating percepts, is arguably the most convoluted: by increasing the temperature, percepts can disappear, then reappear later (Figs. 1, 3 and 5). However, fairly simple and revealing relations still exist for the percept banality $\tau_i(\mathbf{f})$ in (8), considered as a function of $\mathbf{f}$ (that is for $\mathbf{S}$ fixed and evaluated on the optimal transition $\mathbf{Z}(\mathbf{f})$):

Theorem 5 (derivatives of the percept banalities) *For all $\beta > 0$, the derivatives of the percept banalities with respect to the item weights satisfy*

$$\frac{1}{\tau_i} \frac{\partial \tau_i}{\partial f_j} = \frac{1}{\tau_j} \frac{\partial \tau_j}{\partial f_i} = \sum_k \frac{f_k}{\tau_k^2} \frac{\partial \tau_k}{\partial f_i} \frac{\partial \tau_k}{\partial f_j} \quad \text{and} \quad \sum_k \frac{f_k}{\tau_k} \frac{\partial^2 \tau_k}{\partial f_i \partial f_j} = 0. \quad (26)$$

Identities (26) permit to prove the concavity of the effective entropy for all $\beta > 0$ (Theorem 9). Arguably relevant in further perturbation studies, they are not further interpreted nor investigated here.

4 Types of Dissimilarities and Illustrations

4.1 Example B: World Cities

Sparsely populated places, spatially close to more populated places, tend to be overlooked, and designated by the latter — a mechanism captured by the present formalism.

The population N_i , latitudes ϕ_i , longitudes θ_i of $n = 30$ world cities have been extracted from the (now outdated) R dataset `world.cities{maps}`. The selected sample contains the five most populated cities for each of the six continents. Relative weights are $f_i = N_i/N_\bullet$ and geodesic distances are given by

$$d_{ij}^{\text{geo}} = \arccos(\kappa_{ij}) \quad \text{where} \quad \kappa_{ij} = \sin \phi_i \sin \phi_j + \cos \phi_i \cos \phi_j \cos(\theta_i - \theta_j). \quad (27)$$

Fig. 5 For decreasing temperature (that is for increasing β , from the left to the right panels), more percepts tend to emerge. Right: in the low-temperature limit, the histogram of the percepts coincides with the original copper concentration data, that is $\rho_j = f_j$

Fig. 6 Left: cylindrical projection of the 313 world cities with more than 10^6 inhabitants, where the size of symbols reflects the population. The five most populated cities of each continent form the dataset of $n = 30$ objects, weighted by their population size. Right: dendrogram for the $n = 30$ retained cities, resulting from single linkage on D^{geo} , where D^{geo} are the geodesic distances given in (27). The ultrametric dissimilarities d^{ultra}_{ij} were obtained as the height at which i and j are merged. Cutting the dendrogram as above (at 0.683) yields three groups (Americas, Oceania, Africa-Asia-Europe)

The left panel of Fig. 6 depicts the corresponding cylindrical projection, and the right panel depicts the dendrogram resulting from the *single linkage hierarchical ascendent classification* (HAC) applied on dissimilarities D^{geo} .

Figures 7, 8, 9, 10, 11 and 12 give, keeping the city weights f unchanged, the behavior of the signed weights r and percept weights ρ for the various regimes, as well as the effective and reduced entropies, and the effective variety. They do so, respectively, for the dissimilarities D^{geo} , their square, their cube, their squared root, D^{ultra} (the ultrametric dissimilarity obtained from the dendrogram of Fig. 6) and for random dissimilarities D^{random} (whose univariate coordinates are independently drawn as the square of a Student variate with three degrees of freedom).

Fig. 7 Similarities $S = \exp(-\beta D^{\text{geo}})$, where D^{geo} are the geodesic distances (27): signed weights r and percept weights ρ for the various regimes (left panel), effective and reduced entropies (middle panel), and the effective variety (right panel)

Fig. 8 Same legend as Fig. 7, for similarities $S = \exp(-\beta D^{\text{geo}2})$, where $d_{ij}^{\text{geo}2} = (d_{ij}^{\text{geo}})^2$

Fig. 9 Same legend as Fig. 7, for similarities $S = \exp(-\beta D^{\text{geo}3})$, where $d_{ij}^{\text{geo}3} = (d_{ij}^{\text{geo}})^3$

Fig. 10 Same legend as Fig. 7, for similarities $S = \exp(-\beta \sqrt{D^{\text{geo}}})$

Fig. 11 Same legend as Fig. 7, for similarities $S = \exp(-\beta D^{\text{ultra}})$, where D^{ultra} is the ultrametric dissimilarity obtained from single linkage applied on D^{geo} (Fig. 6 right panel)

Fig. 12 Same legend as Fig. 7, for similarities $S = \exp(-\beta D^{\text{random}})$, where the symmetric off-diagonal components of the random dissimilarity D^{random} (see the text)

The understanding of the class of *exponential similarity matrices* of the form $S = \exp(-\beta D)$ (componentwise or Hadamard exponential), where $\beta > 0$ and D is a dissimilarity matrix, benefits from a wealth of acute studies in matrix analysis (see, e.g., Horn & Johnson, 1991; Critchley & Fichtel, 1994; Martínez et al., 1994; Bapat & Raghavan, 1997; Deza & Laurent, 1997; Reams, 1999; Bavaud, 2011; Dellacherie et al., 2014 and references therein). Theorem 6 summarizes some salient features from these sources:

Theorem 6 (exponential similarity matrices) *Let $S = \exp(-\beta D)$. Then*

- 6a) S is positive semi-definite (p.s.d.) for all $\beta > 0$ iff D is squared Euclidean.
- 6b) S is positive definite (p.d.) for all $\beta > 0$, and hence invertible, iff D is squared Euclidean and proper.
- 6c) if D is ultrametric (and hence squared Euclidean) and proper, then $S^{-1} = (s^{ij})$ is, for all $\beta > 0$, a strictly diagonally dominant Stieltjes matrix, that is a p.d. matrix with $s^{ij} \leq 0$ for $i \neq j$ and $s^{i\bullet} > 0$.
- 6d) proper ultrametric dissimilarities possess an inverse $(D^{\text{ultra}})^{-1} = (d^{ij})$ obeying $d^{i\bullet} > 0$.

Recall that $D = (d_{ij})$ is squared Euclidean iff D is of the form $d_{ij} = \|x_i - x_j\|^2$ for some vectors $x_i \in \mathbb{R}^p$. The dissimilarities D^{geo} and $\sqrt{D^{\text{geo}}}$ (Hadamard square root) are squared Euclidean (see, e.g., Critchley & Fichtel, 1994), in contrast to $D^{\text{geo}2}$, $D^{\text{geo}3}$ or D^{random} which are not. The difference between the reduced and effective entropies appears smaller for squared Euclidean dissimilarities, and still smaller for ultrametric dissimilarities which are investigated next.

4.2 Ultrametric (Dis)similarities

The single-linkage ultrametric dissimilarity D^{ultra} extracted from the dendrogram in Fig. 6 is squared Euclidean too (see, e.g., Critchley & Fichtel, 1994), and yields a *strictly ultrametric similarity* satisfying $s_{ij} \geq \max(s_{ik}, s_{jk})$, as well as $s_{ii} > s_{ij}$ for all $i \neq j$.

From Theorem (6d), the normalized *circumweights* defined by

$$g_i = \frac{d^{i\bullet}}{d^{\bullet\bullet}} > 0$$

satisfy

$$\partial_j^g = \sum_i g_i d_{ij} = \frac{1}{d^{\bullet\bullet}} \sum_i d^{i\bullet} d_{ij} = \frac{1}{d^{\bullet\bullet}} = 2\Delta^g. \quad (28)$$

Fig. 13 Same legend as Fig. 7, for similarities $S = \exp(-\beta D^{\text{ultra}})$ and item weights given by the circumweights $\mathbf{g}$ in (28)

In other terms, there exists, for each proper ultrametric dissimilarity, a weight distribution $\mathbf{g}$ such that the barycenter of the configuration is at *equal distance* $\Delta^g = 1/(2d^{\bullet\bullet})$ from all items, which lie on the circumcircle centered on the barycenter. This vividly illustrates the poor low-dimensional compressibility of ultrametric configurations when performing multidimensional scaling. This distance turns out to be 0.683 in the world cities example (see the right panel of Fig. 6), corresponding to a geodesic distance of about $0.683 \times 6371 \cong 4351$ km.

When endowed with circumweights, all items dominate by construction in the high-temperature regime (i.e., $A^* = \langle n \rangle$), but they do not necessarily appear in the medium-temperature regime (Fig. 13). As pointed out by Pavoine et al. (2005), ultrametric dissimilarities grant that the maximum value of Rao quadratic entropy $\Delta(\mathbf{f})$ is precisely attained for $\mathbf{f} = \mathbf{g}$, where all items or species are present (i.e., $\text{supp}(\mathbf{g}) = \langle n \rangle$), as expected from a decent measure of biodiversity.

Figure 14 depicts the case of ultrametric *equidistant similarities* $d_{ij} = d > 0$ for $i \neq j$, for which the critical temperatures can be explicitly computed:

Theorem 7 (equidistant dissimilarities) *Let $d_{ij} = d > 0$ for $i \neq j$ (and $d_{ii} = 0$), with distinct weights $\mathbf{f}$ decreasingly ordered as $f_{[1]} > f_{[2]} > \dots > f_{[n]}$. Let $s(\beta) = \exp(-\beta d / \Delta)$.*

Fig. 14 Same legend as Fig. 7, for equidistant dissimilarities $d_{ij} = d > 0$ for $i \neq j$, and city weights $\mathbf{f}$ unchanged

The low-temperature regime holds for $s(\beta) \in [0, s_-)$, and the high-temperature regime holds for $s(\beta) \in (s_+, 1)$, where

$$s_- = \frac{f_{[n]}}{1 - (n-1)f_{[n]}} < s_+ = \frac{f_{[2]}}{f_{[1]}}.$$

Equivalently,

$$\beta_L = \frac{1}{2} \left(1 - \sum_i f_i^2 \right) \ln \frac{1 - (n-1)f_{[n]}}{f_{[n]}} > \beta_H = \frac{1}{2} \left(1 - \sum_i f_i^2 \right) \ln \frac{f_{[1]}}{f_{[2]}}.$$

5 Further Properties as Measures of Diversity

The exponential of the entropy $\exp(H)$ can be interpreted as a measure of the effective number of alternatives, taking on its maximum value n for uniform distributions $f_i = 1/n$. Similarly, the reduced entropy R has been introduced as a similarity-dependent measure of ecological diversity, where $\exp(R)$ (denoted ${}^1D^S(f)$ by Leinster and Cobbold, 2012) can be interpreted as an effective, similarity-reduced number of alternatives. However, neither $R(f)$ nor $E(f)$ take on their maximum on the uniform distribution in general (see Theorem 10 below).

The larger the similarity between items, the smaller should be the corresponding diversity. This *monotonicity* property is satisfied for the reduced and effective entropies: if $\tilde{s}_{ij} \geq s_{ij}$ for all i, j , then $\tilde{R} \leq R$ (from (12)) and $\tilde{E} \leq E$ (from (10)).

Another property shared by the reduced and effective entropies is the so-called *modularity* (Leinster & Cobbold 2012), to be compared with the decomposition of Shannon entropy under *aggregation*: let X be a categorical variable, and let G be a coarser variable resulting from the aggregation of some categories of X , that is $H(G|X) = 0$. Then $H(X) = H(G) + H(X|G) = H(G) + \sum_g \pi_g H(X|g)$. Here we have:

Theorem 8 (modularity) *Let the n items be partitioned into m communities or groups $g = 1, \dots, m$, in such a way that the similarity between two items belonging to different groups is zero. Then*

$$R = H(G) + \sum_g \pi_g R^g \quad \text{and} \quad E = H(G) + \sum_g \pi_g E^g, \quad (29)$$

where $\pi_g = \sum_{i \in g} f_i$ is the relative weight of group g , $H(G) = -\sum_g \pi_g \ln \pi_g$, and the quantities R^g and E^g denote the reduced and effective entropies within group g only.

What if the rather restrictive conditions of the previous theorem, namely strict partitioning of items into groups (that is $H(G|X) = 0$) and zero inter-group similarity (that is $S = \bigoplus_g S^g$) are lifted? The question directly points to the issue of *concavity*, a desirable property of diversity measures, which says that the diversity of the whole is not less than the average diversity of its parts: such is, notably, the case for the variance (the total variance is bounded below by the within-groups variance, the difference being the between-groups variance), and Shannon entropy ($H(X) \geq H(X|G)$).

The effective entropy turns out to be concave as well (Theorem 9 below). Let f_i^g (obeying $f_i^g \geq 0$ and $f_{\bullet}^g = 1$) denote the relative weight of item i in group $g = 1, \dots, m$, let $\pi_g \geq 0$ with $\pi_{\bullet} = 1$ denote the relative importance of group g , and let $f_i = \sum_g \pi_g f_i^g$

denote the item weights of the whole. Also, let $F[f, \mathbf{Z}]$ denote the free energy (10) considered as a function of the two independent variables f and $\mathbf{Z}$, and finally let $E(f) = \min_{\mathbf{Z} \in \mathcal{Z}} F[f, \mathbf{Z}]$.

Theorem 9 (concavity of the effective entropy) *The effective entropy is concave in f , that is*

$$E(f) \geq \sum_{g=1}^m \pi_g E(f^g) \quad \text{for } f = \sum_g \pi_g f^g. \quad (30)$$

As the proof in the [Appendix](#) makes clear, virtually no condition on $\mathbf{S}$ in (10) is required to ensure the validity of Theorem 9. By contrast, the concavity of the reduced entropy $R(f)$, although supposedly simpler to manipulate than $E(f)$, seems harder to establish: without additional conditions on $\mathbf{S}$, $R(f)$ is *not concave in general*, as shown by the following counter-example: let $n = 3$ and

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & 0.7 \\ 0 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{pmatrix} f = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.5 \end{pmatrix} f^1 = \begin{pmatrix} 0.2 \\ 0.2 \\ 0.6 \end{pmatrix} f^2 = \begin{pmatrix} 0.3 \\ 0.3 \\ 0.4 \end{pmatrix}.$$

Then $f = \frac{1}{2}(f^1 + f^2)$ and $R(f) - \frac{1}{2}R(f^1) - \frac{1}{2}R(f^2) = 0.33667 - \frac{1}{2} \cdot 0.26791 - \frac{1}{2} \cdot 0.40622 = -0.00039$, whose negative sign violates the concavity of $R(f)$. By contrast, one finds $E(f) - \frac{1}{2}E(f^1) - \frac{1}{2}E(f^2) = 0.214 - \frac{1}{2} \cdot 0.143 - \frac{1}{2} \cdot 0.214 = 0.036$, non-negative as it must.

In the naive approach, the uniform distribution $f_i = 1/n$ is well-known to maximize Shannon entropy $H(f)$. This is not the case anymore for the effective entropy, whose non-uniform maximizing distribution f° can be explicitly computed for low temperatures:

Theorem 10 (maximum effective entropy) *Let β be large enough so that $\mathbf{S}$ is invertible with $s^{i\bullet} > 0$ for all i . Then the item distribution $f_i^\circ = s^{i\bullet}/s^{\bullet\bullet}$ maximizes the effective entropy $E(f) \leq E(f^\circ) = \ln s^{\bullet\bullet}$. Also, f° constitutes a stationary point of the reduced entropy $R(f)$. Furthermore, $E(f^\circ) = R(f^\circ)$, $\tau_i^\circ = 1/s^{\bullet\bullet}$, $\rho_j^\circ = f_j^\circ$ and $z_{ij}^\circ = s_{ij}s^{j\bullet}$.*

Thus $\exp(E(f^\circ)) = s^{\bullet\bullet}$ measures, for β large, the maximum number of effective items, and tends to n , as it must, in the naive limit $\beta \rightarrow \infty$. By contrast, f° is a stationary point of the reduced entropy $R(f)$, whose lack of concavity cannot however exclude the presence of multiple extrema in general.

Finally, we address the question of the *subadditivity* of the effective and reduced entropies, which expresses for Shannon entropy as the independence bound $H(X, Y) \leq H(X) + H(Y)$, where equality holds iff X and Y are independent. Here X can be thought of as a categorical variable whose modalities $i = 1, \dots, n$ are similar in part, as expressed by the $n \times n$ similarity matrix s_{ij}^X . Likewise, Y is a categorical variable with modalities $k = 1, \dots, m$ and associated $m \times m$ similarity matrix s_{kl}^Y . The weights of the cross-modalities are $f_{(ik)}^{XY}$ with $f_i^X = f_{(i\bullet)}^{XY}$ and $f_k^Y = f_{(\bullet k)}^{XY}$, and we assume the similarities between cross-modalities to be defined as

$$s_{(ik)(jl)}^{XY} = s_{ij}^X \cdot s_{kl}^Y \quad \text{that is} \quad \mathbf{S}^{XY} = \mathbf{S}^X \otimes \mathbf{S}^Y \quad (\text{Kronecker product}) \quad (31)$$

which is equivalent to the definition $d_{(ik)(jl)}^{XY} = d_{ij}^X + d_{kl}^Y$.

Theorem 11 (subadditivity) *Consider a bivariate distribution f^{XY} with margins f^X and f^Y , and let $\hat{f}_{(ik)}^{XY} = f_i^X f_k^Y$ be the corresponding independent distribution. Then*

$$E(\hat{f}^{XY}) = E(f^X) + E(f^Y) \quad \text{and} \quad R(\hat{f}^{XY}) = R(f^X) + R(f^Y).$$

Furthermore, $E(f^{XY}) \leq E(\hat{f}^{XY})$, which makes E subadditive.

As the proof in the [Appendix](#) makes clear, subadditivity does not hold for the reduced entropy, which, after its lack of concavity (Theorem 9) suffers here a second setback. For a minimal example, consider $n = m = 2$ and

$$S^X = \begin{pmatrix} 1 & .8 \\ .8 & 1 \end{pmatrix} S^Y = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix} (f_{(ik)}^{XY}) = \begin{pmatrix} .41 & .29 \\ .19 & .11 \end{pmatrix} (\hat{f}_{(ik)}^{XY}) = \begin{pmatrix} .42 & .28 \\ .18 & .12 \end{pmatrix}.$$

with $f^X = (.7, .3)$ and $f^Y = (.6, .4)$. Then $E(f^{XY}) - E(f^X) - E(f^Y) = -1.7 \cdot 10^{-16} < 0$, but $R(f^{XY}) - R(f^X) - R(f^Y) = 1.2 \cdot 10^{-5} > 0$.

6 Conclusion

This paper has introduced and analyzed the properties of a new diversity index, the effective entropy. It is based upon a formalism which is both tractable and non-trivial, and it is no coincidence that parts of this formalism also serve in statistical mechanics, statistics, operation research and information theory for the exposition and solution of their own relevant issues. The effective entropy appears to satisfy many properties expected from a diversity index, and handles the presence of item (dis)similarities in a systematic, controlled way. It remedies the formal deficiencies of the reduced entropy, for which it provides a lower bound, and provides an explicit mechanism of diversity reduction due to the confusion between close items.

For a configuration of weighted items with given pair dissimilarities, the effective and the reduced entropies constitute a one-parameter diversity family indexed by a discriminability or inverse temperature parameter. Both entropies converge towards Shannon entropy in the low-temperature limit, provided items are pairwise distinct. In the high-temperature limit, the behavior of the effective and reduced entropy is governed by the minimum eccentricity, respectively the average eccentricity, that is (up to a factor two) Rao quadratic entropy.

Identifying the exact conditions on the (dis)similarities making the reduced entropy concave or subadditive constitutes an obvious future challenge. Determining a well-behaved “power” generalization of the presently proposed “logarithmic” effective dissimilarity, in the spirit of Rényi or Tsallis entropies, constitutes another.

Appendix

Proof of $\lim_{\beta \rightarrow \infty} E = H_{agg}$ for semi-proper dissimilarities (15): for $\beta \rightarrow \infty$, the energy term (3) must vanish, making $z_{ij} = 0$ whenever i and j belong to different equivalence classes $C[i] \neq C[j]$. For i, j in the same class C , the mutual information (4) is

minimum under independence (within C), that is for $z_{ij} = a_j$; identity $\rho_j = \sum_i f_i z_{ij} = \sum_{i \in C} f_i a_j = F_C a_j$ entails

$$K = \sum_C \sum_{i,j \in C} f_i \frac{\rho_j}{F_C} \ln \frac{\rho_j}{F_C \rho_j} = - \sum_C \sum_{i \in C} f_i \underbrace{\sum_{j \in C} \frac{\rho_j}{F_C}}_{=1} \ln F_C = - \sum_C F_C \ln F_C = H_{\text{agg}}.$$

□

Proof of Theorem 1 Differentiating $F[\mathbf{Z}] + \sum_i \lambda_i (1 - z_{i\bullet})$ with respect to z_{ij} (where λ_i is the Lagrange multiplier associated with constraint $z_{i\bullet} = 1$) yields the first-order condition $\beta d_{ij} + \ln \frac{z_{ij}}{\rho_j} = \lambda_i / f_i$, equivalent to (8); note that non-negativity of $\mathbf{Z}$ is automatically ensured. The second derivative $\frac{\partial^2 F[\mathbf{Z}]}{\partial z_{ij} \partial z_{i'j'}} = \delta_{jj'} (\frac{f_i \delta_{ii'}}{z_{ij}} - \frac{f_i f_{i'}}{\rho_j})$ (see, e.g., Bavaud, 2009) can be shown to be positive semi-definite. Eigenvalues are zero for eigenvectors parallel to z_{ij} (that is $\sum_{i'} \frac{\partial^2 F[\mathbf{Z}]}{\partial z_{ij} \partial z_{i'j'}} z_{i'j} = 0$), but positive otherwise, implying the strict convexity of $F[\mathbf{Z}]$ under admissible variations $z_{ij} \rightarrow z_{ij} + \varepsilon_{ij}$ with $\varepsilon_{i\bullet} = 1$, and thus the unicity of the optimal solution in the interior of the convex set $\mathcal{Z}$. Summing the first identity in (7) over $\sum_i f_i \dots$ yields $\rho_j = c_j \rho_j$, that is $c_j = 1$ for $\rho_j > 0$.

The above derivation is incomplete at the boundary of $\mathcal{Z}$ when $\rho_j = 0$, for some j (implying $z_{ij} = 0$ for all i). Here the condition $c_j \leq 1$ can be derived (with some additional work) by starting from the Kuhn-Tucker conditions of convex optimization, or directly as follows: adding mass on percept j can be performed by the transformation $\tilde{z}_{ij} = \alpha_i \in (0, 1)$ and $\tilde{z}_{ik} = (1 - \alpha_i) z_{ik}$ for $k \neq j$. The new free energy reads $F[\tilde{\mathbf{Z}}] = F[\mathbf{Z}] + \bar{\alpha} \zeta(\gamma) + 0(\alpha^2)$ where $\bar{\alpha} = \sum_i f_i \alpha_i$, $\gamma_i = f_i \alpha_i / \bar{\alpha} \geq 0$ and $\zeta(\gamma) = \sum_i \gamma_i [\beta d_{ij} + \ln \tau_i + \ln(\gamma_i / f_i)]$ is convex in γ . Hence $\rho_j = 0$ minimizes the free energy if $\min_\gamma \zeta(\gamma) = \zeta(\gamma^0) \geq 0$. One finds $\gamma_i^0 = f_i s_{ij} / (\tau_i c_j)$ and $\zeta(\gamma^0) = -\ln c_j$, and the latter condition becomes $c_j \leq 1$. □

Proof of Theorem 2 Consider the functional $M[\mathbf{Z}, \boldsymbol{\rho}] = \sum_{ij} f_i z_{ij} (\beta d_{ij} + \ln \frac{z_{ij}}{\rho_j})$, where $\mathbf{Z}$ is a transition matrix and $\boldsymbol{\rho}$ some distribution. Then (a) $\boldsymbol{\rho}^{(t)} = \arg \min_{\boldsymbol{\rho}} M[\mathbf{Z}^{(t)}, \boldsymbol{\rho}]$ and (b) $\mathbf{Z}^{(t+1)} = \arg \min_{\mathbf{Z}} M[\mathbf{Z}, \boldsymbol{\rho}^{(t)}]$. Hence

$$F[\mathbf{Z}^{(t+1)}] = M[\mathbf{Z}^{(t+1)}, \boldsymbol{\rho}^{(t+1)}] \stackrel{(a)}{\leq} M[\mathbf{Z}^{(t+1)}, \boldsymbol{\rho}^{(t)}] \stackrel{(b)}{\leq} M[\mathbf{Z}^{(t)}, \boldsymbol{\rho}^{(t)}] = F[\mathbf{Z}^{(t)}]$$

Applying $\sum_i f_i \dots$ to the second identity in (19) yields $\lim_{t \rightarrow \infty} \rho_j^{(t+1)} / \rho_j^{(t)} = c_j$. Thus $c_j < 1$ implies $\rho_j = \lim_{t \rightarrow \infty} \rho_j^{(t)} = 0$, in accordance with Theorem (1).

The condition $\mathbf{Z}^{(0)} \in \mathcal{Z}_{(n)}$ is crucial, since restricting $\mathbf{Z}^{(0)}$ to $\mathcal{Z}_A$, say, where A is a strict subset of $\langle n \rangle$, that is starting with a $\mathbf{Z}^{(0)}$ with column margins $z_{\bullet j}^{(0)} = 0$ for $j \notin A$, will automatically generate subsequent transitions $\mathbf{Z}^{(t)} \in \mathcal{Z}_A$, converging to the optimal solution of the *restricted problem*

$$E_A = \min_{\mathbf{Z} \in \mathcal{Z}_A} F[\mathbf{Z}] \geq E = \min_{\mathbf{Z} \in \mathcal{Z}} F[\mathbf{Z}],$$

corresponding to a so-called *metastable solution* in statistical mechanics, here betrayed by the possible presence of some $c_j > 1$ for $j \notin A$, in violation of Theorem 1.

The iterative process (19) has been rediscovered several times in various disciplines. Known as the Blahut-Arimoto algorithm in information theory, it constitutes a variant of the EM algorithm of Statistics, which (among other purposes) serves at identifying the optimal

soft clustering minimizing the functional $F^{\text{cluster}} = \beta \sum_{i=1}^n \sum_{g=1}^m f_i z_{ig} d_{ig} + K[\mathbf{Z}]$. Here $z_{ig} \geq 0$ (with $z_{i\bullet} = 1$) is the membership probability of item i in cluster $g = 1, \dots, m$, and d_{ig} is the dissimilarity between i and the optimal cluster representative, which turns out to be its weighted barycenter for squared Euclidean dissimilarities (see, e.g., Bavaud, 2009) and references therein). In contrast to problem (6), the cluster representatives g are not constrained to $g \in \langle n \rangle$: for $m \geq n$, the minimum free energy F^{cluster} provides a *lower bound* for the effective entropy (6). $\square$

Proof of Theorem 3 By Theorem 1, $c_j = 1$ whenever $\rho_j > 0$, thus $\mathbf{S}\boldsymbol{\gamma} = \mathbf{1}$ where $\gamma_i = f_i/\tau_i$ by (17). Multiplying by $\mathbf{S}^{-1}$ yields $f_i/\tau_i = s^{i\bullet}$, and hence $\boldsymbol{\rho} = \mathbf{S}^{-1}\boldsymbol{\tau}$, that is $\rho_j = \sum_i f_i s^{ij}/s^{i\bullet}$ which is (21). The third and fourth identities of (22) are obtained by simple substitution from (7), resp. (14). Identities $\sum_j z_{ij} = 1$ and $\sum_i f_i z_{ij} = \rho_j$ are straightforward to demonstrate, showing (21) and (22) to yield the optimal solution, provided $\min_j r_j \geq 0$. $\square$

Proof of Theorem 4 (24) and (25) follow from the properties discussed in the beginning of Section 3.4. For $\beta = 0$, $\mathbf{S} = \mathbf{J}$ and all sub-indicators (17) are identically one. First-order expansion $s_{ij} = 1 - \beta d_{ij} + 0(\beta^2)$ yields

$$c_j(\beta) = \sum_i f_i (1 - \beta d_{ij} + \beta \sum_{k \in A^*} \rho_k d_{ik}) + 0(\beta^2) = 1 - \beta(\partial_j - \partial^*) + 0(\beta^2)$$

which demonstrates the stability of the solution. The high-temperature expansions (25), as well as the exactness of the relation $E = \beta \partial^*$ when A^* consists of a single element, follows analogously from (10). $\square$

Proof of Theorem 5 Following Theorem 1, $c_j = 1$ for $\rho_j > 0$. Derivating (17) with respect to f_k , and a second time with respect to f_l , yields for $\rho_j > 0$

$$\frac{s_{kj}}{\tau_k} = \sum_i \frac{f_i s_{ij}}{\tau_i^2} \frac{\partial \tau_i}{\partial f_k} \quad (\text{A.1})$$

$$\frac{s_{kj}}{\tau_k^2} \frac{\partial \tau_k}{\partial f_l} + \frac{s_{lj}}{\tau_l^2} \frac{\partial \tau_l}{\partial f_k} = 2 \sum_i \frac{f_i s_{ij}}{\tau_i^3} \frac{\partial \tau_i}{\partial f_k} \frac{\partial \tau_i}{\partial f_l} - \sum_i \frac{f_i s_{ij}}{\tau_i^2} \frac{\partial^2 \tau_i}{\partial f_k \partial f_l} \quad (\text{A.2})$$

Multiplying both identities with ρ_j (either positive or zero) and summing over j yields using (8) the following identities valid for all k, l

$$1 = \sum_i \frac{f_i}{\tau_i} \frac{\partial \tau_i}{\partial f_k} \quad (\text{A.3})$$

$$\frac{1}{\tau_k} \frac{\partial \tau_k}{\partial f_l} + \frac{1}{\tau_l} \frac{\partial \tau_l}{\partial f_k} = 2 \sum_i \frac{f_i}{\tau_i^2} \frac{\partial \tau_i}{\partial f_k} \frac{\partial \tau_i}{\partial f_l} - \sum_i \frac{f_i}{\tau_i} \frac{\partial^2 \tau_i}{\partial f_k \partial f_l}. \quad (\text{A.4})$$

On the other hand, derivating (A.3) with respect to f_l yields

$$\frac{1}{\tau_l} \frac{\partial \tau_l}{\partial f_k} = \sum_i \frac{f_i}{\tau_i^2} \frac{\partial \tau_i}{\partial f_k} \frac{\partial \tau_i}{\partial f_l} - \sum_i \frac{f_i}{\tau_i} \frac{\partial^2 \tau_i}{\partial f_k \partial f_l}. \quad (\text{A.5})$$

The symmetry in $k \leftrightarrow l$ in the r.h.s. of (A.5) implies $\frac{1}{\tau_l} \frac{\partial \tau_l}{\partial f_k} = \frac{1}{\tau_k} \frac{\partial \tau_k}{\partial f_l}$. Substituting the latter in (A.4) and comparing with (A.5) finally yields $\sum_i \frac{f_i}{\tau_i} \frac{\partial^2 \tau_i}{\partial f_k \partial f_l} = 0$. $\square$

Proof of Theorem 6 A proof of (6a) can be found in Horn & Johnson (1991 p. 476), Bapat & Raghavan (1997 p. 163) or in Reams (1999), and applies to conditionally semi-negative definite (c.s.n.d.) matrices, that is to symmetric $n \times n$ matrices $\mathbf{D} = (d_{ij})$ such that $\sum_{ij} d_{ij} h_i h_j \leq 0$ for all h with $h_\bullet = 0$. On the other hand, $\mathbf{D}$ is squared Euclidean iff it is c.s.n.d. with a null diagonal (see, e.g., Critchley and Fichet, 1994; Deza & Laurent, 1997). Lemma 2.5 in Reams (1999) demonstrates (6b). Property (6c) and its proof appear in (Martínez et al., 1994); see also (Dellacherie et al., 2014) and references therein. Property (6d) follows from adapting Lemma 1 in Martínez et al. (1994) to $\mathbf{D}^{\text{ultra}}$ instead of $\mathbf{S}$ (see also Pavoine et al., 2005 and references therein). $\square$

Proof of Theorem 7 After decreasing labelling, $\partial_j = d(1 - f_j)$, hence $j = 1$ dominates. Also, $\Delta/d = \frac{1}{2}(1 - \sum_i f_i^2)$ and $\tau_j = s + (1 - s)\rho_j$ where $s = s(\beta)$. Sub-indicators (17) are of the form $c_j = \varphi(s) + (1 - s)f_j/[s + (1 - s)\rho_j]$, where the function $\varphi(s)$ does not depend on j . Hence, for $j \neq 1$, $c_j < c_1 = 1$ (and thus $\rho_1 = 1$ and $\rho_j = 0$) iff $f_j/s < f_1$, that is iff $s > \max_{j \neq 1} f_j/f_1 = s_+$.

On the other hand, the inverse of $\mathbf{S} = (1 - s)\mathbf{I}_n + s\mathbf{J}_n$ is $\mathbf{S}^{-1} = \alpha\mathbf{I}_n + \gamma\mathbf{J}_n$ where $\alpha = 1/(1 - s)$ and $\gamma = -s/[(1 - s)(1 - s + ns)]$. The signed weights (21) read $r_j = (\alpha f_j + \gamma)/(\alpha + n\gamma)$ and $\min_j r_j > 0$ iff $\alpha f_n + \gamma > 0$, that is iff $s < f_n/[1 - (n - 1)f_n] = s_-$. $\square$

Proof of Theorem 8 The normalized item weights in group g are $f_i^g = f_i/\pi_g$ if $i \in g$, and $f_i^g = 0$ otherwise. By hypothesis, $\mathbf{S} = \bigoplus_g \mathbf{S}^g$, and thus, for $i \in g$, $b_i = \sum_{j \in g} f_j s_{ij} = \pi_g \sum_{j \in g} f_j^g s_{ij}^g = \pi_g b_i^g$. Then $R = -\sum_i f_i \ln b_i = -\sum_g \pi_g \sum_{i \in g} f_i^g \ln(\pi_g b_i^g) = -\sum_g \pi_g \ln \pi_g - \sum_g \pi_g \sum_{i \in g} f_i^g \ln b_i^g = H(G) + \sum_g \pi_g R^g$.

Similarly, the normalized percept weights in group g are $\rho_j^g = \rho_j/\pi_g$ if $j \in g$, and $\rho_j^g = 0$ otherwise. For $i \in g$, $\tau_i = \sum_j s_{ij} \rho_j = \pi_g \sum_{j \in g} s_{ij}^g \rho_j^g = \pi_g \tau_i^g$. By hypothesis, the transitions $z_{ij} = s_{ij} \rho_j / \tau_i$ are zero if i and j belong to different groups; if they belong to the same group g , then $z_{ij} = z_{ij}^g = s_{ij}^g \rho_j^g / \tau_i^g$, where $z_{i\bullet}^g = 1$ and $\sum_i f_i^g z_{ij}^g = \frac{1}{\pi_g} \sum_i f_i z_{ij} = \frac{1}{\pi_g} \rho_j = \rho_j^g$. Substituting the expression for $z_{ij} = z_{ij}^g$ in (10) yields

$$\begin{aligned} E &= \sum_g \sum_{ij \in g} \pi_g f_i^g z_{ij}^g \ln \frac{z_{ij}^g}{s_{ij}^g \pi_g \rho_j^g} = -\sum_g \pi_g \ln \pi_g + \sum_g \pi_g \sum_{ij \in g} f_i^g z_{ij}^g \ln \frac{z_{ij}^g}{s_{ij}^g \rho_j^g} = \\ &= H(G) + \sum_g \pi_g E^g. \end{aligned}$$

$\square$

Two Proofs of Theorem 9 Taking into account $\partial \rho_j / \partial f_k = z_{kj}$, the Hessian reads $\partial^2 F[\mathbf{f}, \mathbf{Z}] / \partial f_k \partial f_l = -\sum_j z_{kj} z_{lj} / \rho_j$, which is negative semi-definite, thus establishing the concavity of $F[\mathbf{f}, \mathbf{Z}]$ in $\mathbf{f}$. Thus $F[\mathbf{f}, \mathbf{Z}] \geq \sum_g \pi_g F[\mathbf{f}^g, \mathbf{Z}]$ and

$$E(\mathbf{f}) = \min_{\mathbf{Z} \in \mathcal{Z}} F[\mathbf{f}, \mathbf{Z}] \geq \min_{\mathbf{Z} \in \mathcal{Z}} \left(\sum_g \pi_g F[\mathbf{f}^g, \mathbf{Z}] \right) \geq \sum_g \pi_g \min_{\mathbf{Z} \in \mathcal{Z}} F[\mathbf{f}^g, \mathbf{Z}] = \sum_g \pi_g E(\mathbf{f}^g)$$

A second proof of Theorem 9 follows from Theorem 5: to establish the concavity of a quantity such as $E(\mathbf{f})$, it is enough to consider two groups of equal importance $\pi_1 = \pi_2 = \frac{1}{2}$, such that $f_i^1 = f_i + \eta_i$ and $f_i^2 = f_i - \eta_i$, where $\boldsymbol{\eta}$ must satisfy $\boldsymbol{\eta}_\bullet = 0$ but can otherwise

be chosen arbitrarily (provided it is small enough to insure $f_i^1, f_i^2 \geq 0$; recall $f_i > 0$). The excess $\zeta(\boldsymbol{\eta}) := E(\boldsymbol{f}) - \frac{1}{2}E(\boldsymbol{f} + \boldsymbol{\eta}) - \frac{1}{2}E(\boldsymbol{f} - \boldsymbol{\eta})$ satisfies $\zeta(0) = 0$ and

$$\frac{\partial \zeta(\eta=0)}{\partial \eta_i} = -\frac{1}{2} \frac{\partial E(\boldsymbol{f})}{\partial f_i} + \frac{1}{2} \frac{\partial E(\boldsymbol{f})}{\partial f_i} = 0 \quad \text{and} \quad \frac{\partial^2 \zeta(\eta=0)}{\partial \eta_i \partial \eta_j} = -\frac{\partial^2 E(\boldsymbol{f})}{\partial f_i \partial f_j}.$$

Thus, $\zeta(\boldsymbol{\eta}) = -\frac{1}{2} \sum_{ij} E_{ij}(\boldsymbol{f}) \eta_i \eta_j + 0(\eta^3)$, and $\zeta(\boldsymbol{\eta}) \geq 0$ in the neighborhood of $\boldsymbol{\eta} = 0$ iff the Hessian $E_{ij}(\boldsymbol{f})$ of $E(\boldsymbol{f})$ is *conditionally negative semi-definite*, that is obeys $\sum_{ij} E_{ij}(\boldsymbol{f}) \eta_i \eta_j \leq 0$ for all $\boldsymbol{\eta}$ such that $\boldsymbol{\eta}_\bullet = 0$.

Now (14), (A.3) and (26) yield

$$E_i(\boldsymbol{f}) := \frac{\partial E(\boldsymbol{f})}{\partial f_i} = -\ln \tau_i - \sum_k \frac{f_k}{\tau_k} \frac{\partial \tau_k}{\partial f_i} = -\ln \tau_i - 1 \quad (\text{A.6})$$

$$E_{ij}(\boldsymbol{f}) := \frac{\partial^2 E(\boldsymbol{f})}{\partial f_i \partial f_j} = -\frac{1}{\tau_i} \frac{\partial \tau_i}{\partial f_j} = -\sum_k \frac{f_k}{\tau_k^2} \frac{\partial \tau_k}{\partial f_i} \frac{\partial \tau_k}{\partial f_j} \quad (\text{A.7})$$

the later expression being obviously negative semi-definite, and hence conditionally negative semi-definite. $\square$

Proof of Theorem 10 For $\boldsymbol{f} = \boldsymbol{f}^\circ$, one verifies in (21) that $r_j = f_j^\circ > 0$, hence $\boldsymbol{\rho}^\circ = \boldsymbol{f}^\circ$ by (22). As a result, (8) or (22) show $\tau_i^\circ = 1/s^{\bullet\bullet}$ and finally $z_{ij}^\circ = s_{ij}s^{j\bullet}$. Since $\boldsymbol{\tau}^\circ$ is constant, (A.6) together with $f_i^\circ > 0$ then demonstrates $\boldsymbol{f}^\circ$ to be a stationary point for $E(\boldsymbol{f})$, and hence, given the concavity of the latter, its unique maximizer, with value $E(\boldsymbol{f}^\circ) = \ln s^{\bullet\bullet}$. Finally, identity $b_i^\circ = 1/s^{\bullet\bullet}$ entails $R(\boldsymbol{f}^\circ) = E(\boldsymbol{f}^\circ)$, and distribution $\boldsymbol{f}^\circ$ is easily shown to constitute a stationary point of $R(\boldsymbol{f})$. $\square$

Proof of Theorem 11 Let $\hat{b}_{(ik)}^{XY}$ denote the bivariate banalities evaluated at $\boldsymbol{f}^{XY} = \hat{\boldsymbol{f}}^{XY}$, with similar notations for the other quantities. Definition (31) immediately entails $\hat{b}_{(ik)}^{XY} = b_i^X b_k^Y$ and thus $R(\hat{\boldsymbol{f}}^{XY}) = R(\boldsymbol{f}^X) + R(\boldsymbol{f}^Y)$. Turning to the effective entropy, one readily verifies that, whenever $\boldsymbol{f}^{XY} = \hat{\boldsymbol{f}}^{XY}$, the optimal joint transition $\hat{\boldsymbol{Z}}^{XY}$ obeying (7) is simply $\hat{\boldsymbol{Z}}^{XY} = \boldsymbol{Z}^X \otimes \boldsymbol{Z}^Y$, where $\boldsymbol{Z}^X$ and $\boldsymbol{Z}^Y$ are the corresponding univariate optimal transitions. In particular, $\hat{\rho}_{(jl)}^{XY} = \rho_j^X \rho_l^Y$, $\hat{\tau}_{(ik)}^{XY} = \tau_i^X \tau_k^Y$ and thus $E(\hat{\boldsymbol{f}}^{XY}) = E(\boldsymbol{f}^X) + E(\boldsymbol{f}^Y)$.

Consider now a perturbation of the form $\hat{f}_{(ik)}^{XY} = \hat{f}_{(ik)}^{XY} + \epsilon h_{(ik)}$, where ϵ is small and $\boldsymbol{h}$ obeys $h_{(i\bullet)} = 0$ and $h_{(\bullet k)} = 0$ to preserve the margins. Taylor expansion around $\epsilon = 0$ yields

$$E(\boldsymbol{f}^{XY}) = E(\boldsymbol{f}^X) + E(\boldsymbol{f}^Y) + \epsilon \sum_{ik} \hat{E}_{(ik)} h_{(ik)} + \frac{\epsilon^2}{2} \sum_{ikjl} \hat{E}_{(ik)(jl)} h_{(ik)} h_{(jl)} + 0(\epsilon^3)$$

where $\hat{E}_{(ik)}$ and $\hat{E}_{(ik)(jl)}$ are the corresponding derivatives (A.6) and (A.7) evaluated at $\hat{\boldsymbol{f}}^{XY}$. One has $\hat{E}_{(ik)} = -\ln \hat{\tau}_{(ik)}^{XY} - 1 = -\ln \tau_i^X - \ln \tau_k^Y - 1$, and the contribution of this first order term is zero in view of the conditions imposed on $\boldsymbol{h}$. The contribution of the second order term is negative in view of the negative semi-definiteness of $\hat{E}_{(ik)(jl)}$ established in the proof of Theorem 9, and finally $E(\boldsymbol{f}^{XY}) \leq E(\boldsymbol{f}^X) + E(\boldsymbol{f}^Y) + 0(\epsilon^3) = E(\hat{\boldsymbol{f}}^{XY}) + 0(\epsilon^3)$, showing $\hat{\boldsymbol{f}}^{XY}$ to be a local maximum of the effective entropy within the set of bivariate distributions with fixed margins. The latter set being convex, this local maximum is the global maximum in view of the concavity of $E(\boldsymbol{f})$.

The above argument fails for the reduced entropy: its first-order derivative turns out to be $\hat{R}_{(ik)} = -\ln b_i^X - \ln b_k^Y - \alpha_i^X \alpha_k^Y$ where $\alpha_i^X = \sum_j s_{ij}^X f_j^X / b_j^X$ and $\alpha_k^Y = \sum_l s_{kl}^Y f_l^Y / b_l^Y$, and h can always be chosen so that $-\varepsilon \sum_{ik} \alpha_i^X \alpha_k^Y h_{ik} > 0$, thus ruining subadditivity. $\square$

Acknowledgements I'm much indebted to Eric Marcon, who introduced me generously in the subject.

Funding Open Access funding provided by Université de Lausanne. This work was supported by the Swiss National Science Foundation under Grant No. 190377.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bapat, R., & Raghavan, T. (1997). *Nonnegative matrices and applications* Vol. 64. Cambridge: Cambridge University Press.
- Bavaud, F. (2009). Aggregation invariance in general clustering approaches. *Advances in Data Analysis and Classification*, 3(3), 205–225.
- Bavaud, F. (2011). On the Schoenberg transformations in data analysis: theory and illustrations. *Journal of Classification*, 28(3), 297–314.
- Champely, S., & Chessel, D. (2002). Measuring biological diversity using Euclidean metrics. *Environmental and Ecological Statistics*, 9(2), 167–177.
- Chiu, C.-H., Jost, L., & Chao, A. (2014). Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. *Ecological Monographs*, 84(1), 21–44.
- Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*, 2nd edn. Hoboken: Wiley.
- Critchley, F., & Fichtel, B. (1994). The partial order by inclusion of the principal classes of dissimilarity on a finite set, and some of their basic properties. In B. Van Cutsem (Ed.) *Classification and dissimilarity analysis, Lecture Notes in Statistics*, (Vol. 93 pp. 5–65). Springer.
- Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.) *Advances in Neural Information Processing Systems*, (Vol. 26 pp. 2292–2300). Curran Associates, Inc.
- Dellacherie, C., Martinez, S., & San Martin, J. (2014). *Inverse M-matrices and ultrametric matrices, vol. 2118 of Lecture Notes in Mathematics*. Berlin: Springer.
- Deza, M. M., & Laurent, M. (1997). *Geometry of cuts and metrics, volume 15 of algorithms and combinatorics*. Berlin: Springer.
- Horn, R. A., & Johnson, C. R. (1991). *Topics in matrix analysis*. Cambridge: Cambridge University Press.
- Jost, L. (2006). Entropy and diversity. *Oikos*, 113(2), 363–375.
- Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: the importance of species similarity. *Ecology*, 93(3), 477–489.
- Marcon, E. (2016). Mesurer la biodiversité et la Structuration Spatiale. Habilitation à diriger des recherches. Université de Guyane. <https://hal-agroparistech.archives-ouvertes.fr/tel-01502970>.
- Martínez, S., Michon, G., & Martín, J.S. (1994). Inverse of strictly ultrametric matrices are of Stieltjes type. *SIAM Journal on Matrix Analysis and Applications*, 15(1), 98–106.
- Masi, M. (2005). A step beyond Tsallis and rényi entropies. *Physics Letters A*, 338(3–5), 217–224.
- Pavoine, S., Ollier, S., & Pontier, D. (2005). Measuring diversity from dissimilarities with Rao's quadratic entropy: Are any dissimilarities suitable? *Theoretical Population Biology*, 67(4), 231–239.
- Reams, R. (1999). Hadamard inverses, square roots and products of almost semidefinite matrices. *Linear Algebra and its Applications*, 288, 35–43.
- Ricotta, C., & Szeidl, L. (2006). Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical Population Biology*, 70(3), 237–243.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.