entropy

Article

Maximum Entropy Estimation of Probability Distribution of Variables in Higher Dimensions from Lower Dimensional Data

Jayajit Das [1,2,3,4,*]**, Sayak Mukherjee** [1,2] **and Susan E. Hodge** [1,2]

[1] Battelle Center for Mathematical Medicine, Research Institute at the Nationwide Children's Hospital, 700 Children's Drive, OH 43205, USA; E-Mails: mukherjee.39@osu.edu (S.M.); Susan.Hodge@nationwidechildrens.org (S.E.H.)

[2] Department of Pediatrics, The Ohio State University, Columbus, OH 43205, USA

[3] Department of Physics, The Ohio State University, Columbus, OH 43210, USA

[4] Department of Biophysics Program, The Ohio State University, Columbus, OH 43210, USA

* Author to whom correspondence should be addressed; E-Mail: das.70@osu.edu.

Academic Editor: Kevin H. Knuth

Abstract: A common statistical situation concerns inferring an unknown distribution $Q(x)$ from a known distribution $P(y)$, where X (dimension n), and Y (dimension m) have a known functional relationship. Most commonly, $n \leq m$, and the task is relatively straightforward for well-defined functional relationships. For example, if Y_1 and Y_2 are independent random variables, each uniform on $[0, 1]$, one can determine the distribution of $X = Y_1 + Y_2$; here $m = 2$ and $n = 1$. However, biological and physical situations can arise where $n > m$ and the functional relation $Y \rightarrow X$ is non-unique. In general, in the absence of additional information, there is no unique solution to Q in those cases. Nevertheless, one may still want to draw some inferences about Q. To this end, we propose a novel maximum entropy (MaxEnt) approach that estimates $Q(x)$ based only on the available data, namely, $P(y)$. The method has the additional advantage that one does not need to explicitly calculate the Lagrange multipliers. In this paper we develop the approach, for both discrete and continuous probability distributions, and demonstrate its validity. We give an intuitive justification as well, and we illustrate with examples.

Keywords: maximum entropy; joint probability distribution; microbial ecology

1. Introduction

We are often interested in quantitative details about quantities that are difficult or even impossible to measure directly. In many cases we may be fortunate enough to find measureable quantities that are related to our variables of interest. Such examples are abundant in nature. Consider a community of microbes coexisting in humans or other metazoan species [1,2]. It is possible to measure the relative abundances of different species in the microbial community in individual hosts, but it could be difficult to directly measure parameters that regulate interspecies interactions in these diverse communities. Knowing the quantitative values of the parameters representing microbial interactions is of great interest, both because of their role in development of therapeutic strategies against diseases such as colitis, and for basic understanding, as we have discussed in [3].

Inference of these unknown variables from the available data is a subject of a vast literature in diverse disciplines including statistics, information theory, and, machine learning [4–7]. In this paper we will be interested in a specific problem where the unknown variables in a large dimension are related to a smaller number of variables whose joint probability distribution is known from measurements.

In the above example, parameters describing microbial interactions could represent such unknown variables, and their number could be substantially larger than the number of measurable variables, such as abundances of distinct microbial species. The distribution of abundances of microbial species in a host population can be calculated from measurements performed on a large number of individual subjects. The challenge is to estimate the distribution of microbial interaction parameters using the distribution of microbial abundances.

These inference problems can be dealt with by Maximum Entropy (MaxEnt)-based methods that maximize an entropy function subject to constraints provided by the expectation values calculated from measured data [4,5,7,8]. In standard applications of MaxEnt, usually, averages, covariances, and, sometimes, higher-order moments calculated from the data are used to infer such distributions [4,5,7]. Including larger number of constraints in the MaxEnt formalism involves calculating a large number of Lagrange multipliers by solving an equal number of nonlinear equations, which can pose a great computational challenge [9]. Here we propose a novel MaxEnt-based method to infer the distribution of the unknown variables. Our method uses the distribution of the measured variables and provides an elegant MaxEnt solution that bypasses direct calculation of the Lagrange multipliers. Instead, the inferred distribution is described in terms of a degeneracy factor, described by a closed form expression, which depends only on the symmetry properties of the relation between the measured and the unknown variables.

More generally, the above problem relates to the issue of calculating a probability function of X from the probability function of Y, where X and Y are both random variables, and Y and X have a functional relationship. This could involve either discrete or continuous random variables. Standard textbooks [10] in probability theory usually deal with cases where (a) variables X are related to variables of Y by a well-defined functional relationship ($x = g(y)$), with the distribution of the Y variables (y) known, and (b) X resides in a manifold (dimension n) of lower dimension than the Y manifold (dimension m). However, it is not clear how to extend the standard calculations pertaining to the above well-defined case when multiple values of X variables are associated with the same Y variable. This situation easily arises when n is greater than m. We address this problem here, where we

estimate Q(x) from P(y) when n > m. *i.e.*, we infer the higher-dimension variable from the lower-dimension one. We show that when the variables are discrete, no unique solution exists for Q(x), as the system is underdetermined. However, the MaxEnt-based method can provide a MaxEnt solution in this situation that is constrained only by the available information (P(y) in this case) and is free from any additional assumptions. We then extend the results for continuous variables.

2. The Problem

We state the problem, illustrating in this section with discrete random variables. Consider a case when n different random variables, x_1, .., x_n, are related to m (n > m) different variables, y_1, …, y_m, as $\{Y_i = f_i(x_1, ..., x_n)\}$ (f: $R^n \to R^m$). We know the probabilities for the y variables and want to reach some conclusion about the probabilities of the x variables.

We introduce a few terms and notations borrowed from physics that we will use to simplify the mathematical description [11]. A state in the x (or y) space refers to a particular set of values in the variables x_1, ..., x_n (or y_1, ..., y_m). We denote the set of these states as $\{x_1, ..., x_n\}$ or $\{y_1, ..., y_m\}$. The vector notations, $\vec{x} = (x_1, \cdots, x_n)$ and $\vec{y} = (y_1, \cdots, y_m)$, will be used to compactly describe expressions when required. For the same reason, when we use f without a subscript, it will refer to a vector of f values, *i.e.*, $\vec{y} = \vec{f}(\vec{x}) = \left(f_1(\vec{x}), ..., f_m(\vec{x}) \right)$. In standard textbook examples in elementary probability theory and physics, we are provided with the probability distribution function $P(\vec{y})$, where X is related to Y by a well-defined function, $\vec{x} = \vec{g}(\vec{y})$. Such cases are common when Y resides in a higher or equal dimension (m ≥ n) than X. Then $Q(\vec{x})$, with lower dimension n, is calculated using

$$Q(\vec{x}) = \sum_{y_1, \cdots, y_m | \vec{x} = \vec{g}(\vec{y})} P(\vec{y}) \tag{1a}$$

The summation in Equation (1a) is performed over only those states $\{y_1, .., y_m\}$ that correspond to the specified state $\vec{x}$. However, note that the above relation does not hold even when m ≥ n if multiple values of X variables are associated with the same values of the Y variables, e.g., $x^2 = y$, where $-\infty < x < \infty$ and $0 \leq y < \infty$. The MaxEnt formalism developed here can be used for estimating $Q(\vec{x})$ using $P(\vec{y})$ in such cases (see Appendix A1).

Here we are interested in the inverse problem: we are still provided with the probability distribution P(y_1, ..., y_m) and need to estimate the probability distribution Q(x_1, ..., x_n), but now m < n. In this situation, multiple values of the unknown variable X are associated with the same values of observable Y variables and no unique solution for Q(x_1, ..., x_n) exists as the system is underdetermined. Instead of Equation (1a), we use this equation:

$$P(\vec{y}) = \sum_{x_1, \cdots, x_n | \vec{f}(\vec{x}) = \vec{y}} Q(\vec{x}) = \sum_{x_1, \cdots, x_n} Q(\vec{x}) \prod_{i=1}^{m} \delta_{y_i, f_i(\vec{x})} \tag{1b}$$

The constraints imposed on the summation in the last term by the relations ($\vec{y} = \vec{f}(\vec{x})$) between the states in x and y are incorporated using the Kronecker delta function (δ_{ab}, where, $\delta_{a,b} = 1$ when a = b, and, $\delta_{a,b} = 0$ when a ≠ b). For pedagogical reasons we elucidate the problem of non-uniqueness in the solutions using a simple example. This example can be easily generalized.

Example 1. We start with a discrete random variable y, with known distribution $P(y) = 1/3$ for y = 0, 1, 2. Then assume that discrete random variables x_1 and x_2 are related to y, as,

$y = f(x_1,x_2) = x_1 + x_2$. We restrict x_1 and x_2 to being nonnegative integers; hence x_1, and x_2 can assume only three values, 0, 1, and 2.

It follows that $Q(x_1,x_2)$ are related to $P(y)$ following Equation (1b) as,

$$P(y) = \begin{cases} Q(0,0) \ for \ y = 0 \\ Q(0,1) + Q(1,0) \ for \ y = 1 \\ Q(0,2) + Q(1,1) + Q(2,0) \ for \ y = 2 \end{cases}$$

Hence

$$Q(0,0) = 1/3$$

$$Q(0,1) + Q(1,0) = 1/3$$

$$Q(0,2) + Q(1,1) + Q(2,0) = 1/3$$

(2)

The above relation provides three independent linear equations for determining six unknown variables, $Q(0,0)$, $Q(1,0)$, $Q(0,1)$, $Q(1,1)$, $Q(2,0)$, and, $Q(0,2)$. Note, the condition of $\sum_{x_1,x_2} Q(x_1,x_2) = \sum_y P(y) = 1$ is satisfied by the above linear equations, which also makes $Q(1,2) = Q(2,1)$ $= Q(2,2) = 0$. Therefore, the linear system in Equation (2) is underdetermined and $Q(x_1,x_2)$ cannot be found uniquely using these equations. (e.g., $Q(0,1)$ and $Q(1,0)$ could each equal 1/6; or $Q(0,1)$ could equal 1/12, with $Q(1,0) = 1/4$; *etc.*)

This issue of non-uniqueness is general and will hold as long as the number of constraints imposed by $P(y_1, ..., y_m)$ is smaller than that of the number of unknown $Q(x_1, ..., x_n)$. For example, when each direction in y (or x) can take L (or L_1) discrete values and all the states in x are mapped to all the states in y, then the system will be underdetermined as long as, $L^m < L_1^n$.

3. A MaxEnt Based Solution (Discrete)

In this section we propose a solution of this problem using a Maximum Entropy based principle, for discrete variables. We can define Shannon's entropy [4,5,7], *S*, given by

$$S = - \sum_{x_1,\cdots,x_n} Q(\vec{x}) \ln Q(\vec{x})$$

(3)

and then maximize *S* with the constraint that $Q(\vec{x})$ should generate the distribution $P(\vec{y})$ in Equation (1b).

Equation (1b) describes the set of constraints spanning the distinct states in the y space. For example, when each element in the y vector assumes binary values (+1 or −1) there are in total 2^m number of distinct states in the y space providing 2^m number of equations of constraints. We can introduce a Lagrange multiplier for each of the constraint equations, which we denote compactly as a function, $\lambda(y_1, ..., y_m)$ or $\lambda(\vec{y})$ describing a map from $R^n \to R$. That is, every possible y vector is associated with a unique value of λ. Also note, when $P(\vec{y})$ is normalized, $Q(\vec{x})$ is normalized due to Equation (1b), therefore, we will not use any additional Lagrange multiplier for the normalization condition of $Q(\vec{x})$. The distribution, $\hat{Q}(\vec{x})$, that optimizes *S*, subject to the constraints can be calculated as follows. $Q(\vec{x})$ is slightly perturbed from $\hat{Q}(\vec{x})$, *i.e.*, $Q(\vec{x}) = \hat{Q}(\vec{x}) + \delta Q(\vec{x})$. Then expanding *S* in Equation (3) and the constraints in Equation (1b) in terms of $\delta Q(\vec{x})$ and setting the

terms proportional to $\delta Q(\vec{x})$ zero (optimization condition) yields $\hat{Q}(\vec{x})$ in terms of the Lagrange multipliers, *i.e.*,

$$-\sum_{x_1,\cdots,x_n}\delta Q(\vec{x})\ln\hat{Q}(\vec{x})-\sum_{x_1,\cdots,x_n}\delta Q(\vec{x})+\sum_{y_1,\cdots,y_m}\sum_{x_1,\cdots,x_n}\lambda(\vec{y})\delta Q(\vec{x})\prod_{i=1}^{m}\delta_{y_i,f_i(\vec{x})}=0 \qquad (4)$$

One can indeed confirm that the terms in the expansion of *S* and the constraints proportional to $(\delta Q)^2$ at $Q(\vec{x})=\hat{Q}(\vec{x})$ is $-1/\hat{Q}(\vec{x})$, thus, $\hat{Q}(\vec{x})$ maximizes *S*. The method used here for maximizing S subject to the constraints is a standard one [4,11].

The solution for $\hat{Q}(\vec{x})$ from the above Equation (4) is given by,

$$\hat{Q}(\vec{x})=e^{\sum_{y_1,\cdots,y_m}\lambda(\vec{y})\prod_{i=1}^{m}\delta_{y_i,f_i(\vec{x})}-1}=e^{\lambda(\vec{f}(\vec{x}))-1} \qquad (5)$$

Note the partition function (usually denoted as Z in textbooks [4,11]) does not arise in the above solution as the normalization condition for Q(x) is incorporated in the constraint equations in Equation (1b). We show the derivation of Equation (5) for Example 1 in Appendix A2 for pedagogical reasons. From the above solution (Equation (5)) we immediately observe the two main features that $\hat{Q}(\vec{x})$ exhibits:

(i) The values of $\hat{Q}(\vec{x})$ for the states {x₁, …, xₙ} that map to the same state y₁, …, yₘ via {$f_i(\vec{x})$} are equal to each other. In the simple example above, this implies Q(1,0) = Q(0,1), and, Q(1,1) = Q(0,2) = Q(2,0).

(ii) $\hat{Q}(\vec{x})$ contains all the symmetry properties present in the relation {yᵢ = fᵢ(x₁, …, xₙ)}. In the simple example, the relation between y and x was symmetric in permutation of x₁ and x₂, implying, Q(x₁,x₂) = Q(x₂,x₁).

We will take advantage of the above properties to avoid direct calculation of the Lagrange multipliers in Equation (4): For the states {$\tilde{x}_1,\cdots,\tilde{x}_n$} in the x space that map to the same state, $\tilde{y}_1,\cdots,\tilde{y}_m$, in the y space, Equation (1b) can rewritten as

$$P(\vec{\tilde{y}})=\sum_{\tilde{x}_1,\cdots,\tilde{x}_n}\hat{Q}(\vec{\tilde{x}})=k(\vec{\tilde{y}})\hat{Q}(\vec{\tilde{x}'}) \qquad (6a)$$

$$\Rightarrow \hat{Q}(\vec{\tilde{x}'})=P(\vec{\tilde{y}})/k(\vec{\tilde{y}}) \qquad (6b)$$

where $k(\tilde{y}_1,\cdots,\tilde{y}_m)$ gives the total number of distinct states {$\tilde{x}_1,\cdots,\tilde{x}_n$} in the x space that correspond to the state, $\tilde{y}_1,\cdots,\tilde{y}_m$ or $\vec{\tilde{y}}$. Since, all the states in {$\tilde{x}_1,\cdots,\tilde{x}_n$} will have the same probability, in the second step in Equation (6a) we replace the summation with $k(\vec{\tilde{y}})$, multiplied by the probability of any state $(\tilde{x}'_1,\cdots,\tilde{x}'_n)$ or $\vec{\tilde{x}'}$ in {$\tilde{x}_1,\cdots,\tilde{x}_n$}. We designate $k(\vec{\tilde{y}})$ as the degeneracy factor, borrowing a similar terminology in physics. $k(\vec{\tilde{y}})$ can be expressed in terms of the Kronecker delta functions as,

$$k(\vec{\tilde{y}})=\sum_{x_1,\cdots,x_n}\left[\prod_{i=1}^{m}\delta_{\tilde{y}_i,f_i(\vec{x})}\right] \qquad (7)$$

Note, the degeneracy factor in Equation (7) only depends on the relationship between {$\vec{x}$} and {$\vec{y}$}, and, does not depend on the probability distributions, P and Q. In our simple example above, since y = x₁ + x₂, Q(0,1) and Q(1,0) both correspond to y = 1, therefore $k(\tilde{y}=1)=2$. Equation (6b) is the

main result of this section, which describes the inferred distribution $\hat{Q}(\vec{x})$ in terms of the known probability distribution $P(\vec{y})$, and, $k(\vec{y})$, which can be calculated from the given relation between y and x. Thus, the calculation of $\hat{Q}(\vec{x})$, as shown in Equation (6b), does not involve direct evaluation of the Lagrange multipliers, $\lambda(\vec{y})$. These two quantities are related to $P(\vec{y})$, and, $k(\vec{y})$, following Equations (5), (6b) and (7), as,

$$e^{\lambda(\vec{f}(\vec{x}'))-1} = \frac{P(\tilde{\vec{y}})}{k(\tilde{\vec{y}})} \qquad (8)$$

Example 1, continued. We provide a solution for Example 1 presented above. By simple counting, we see the degeneracy factors are

$$k(\tilde{y}=0)=1, k(\tilde{y}=1)=2, k(\tilde{y}=2)=3$$

Thus following Equation (2), Q(0,0) = P(0) = 1/3, Q(0,1) = Q(1,0) = P(1)/2 = 1/6, and, Q(2,0) = Q(1,1) = Q(0,2) = P(2)/3 = 1/9. For more complex problems, the degeneracy factors can be calculated numerically. Maximizing the entropy, S, is what made all the Qs be equal for any one y value.

4. Results for Continuous Variables

The above results can be extended when {X_i} and {Y_i} are continuous variables. However, there is an issue that makes a straightforward extension of the calculations shown in the discrete case in the continuum limit difficult. The issue is related to the continuum limit of the entropy function *S* in Equation (3). Replacing the summation in Equation (3) with an integral in the limit of large number of states as the step size separating the adjacent states is decreased to zero creates an entropy expression which is negative and unbounded [12]. This problem can be ameliorated by defining a relative entropy, RE, defined as,

$$RE = \int dx_1 \cdots dx_n q(x_1, \cdots, x_n) \ln \left[\frac{q(x_1, \cdots, x_n)}{u(x_1, \cdots, x_n)} \right] \qquad (9)$$

where, u is a uniform probability density function defined on the same domain as q. RE always remains positive with a lower bound at zero. The results obtained by maximizing S in the previous section can be derived by minimizing a relative entropy (RE) defined above with the discrete distributions, Q and a uniform distribution, U, where the integral in Equation (9) is replaced by a summation over the states in the x space. RE in Equation (9) quantifies the difference between the distribution q(x_1, …, x_n) and the corresponding uniform distribution.

The definition of RE in Equation (9) still has an issue of defining the uniform distribution when the x variables are unbounded. In some cases, it may be possible to solve the problem by introducing finite upper and lower bounds and then analyzing the results in the limit where the upper (or lower) bound approaches ∞ (or −∞). We will illustrate this approach in Example 4, below. Also, see Example 3 for a comparison.

In the continuum limit, the constraints on q(x_1, ..., x_n) or $q(\vec{x})$, imposed by the probability density function (pdf) p(y_1,..,y_m) or $p(\vec{y})$ are given by,

$$p(\vec{y}) = \int dx_1 \cdots dx_n q(\vec{x}) \prod_{i=1}^{m} \delta_D (y_i - f_i(\vec{x})) \tag{10}$$

The Dirac delta function for a single variable x is defined as,

$$\int_R dx \, \delta_D (x) = 1 \tag{11}$$

where the region R contains the point $x = 0$.

Since, the pdf $p(\vec{y})$ resides in a lower dimension compared to $q(\vec{x})$, estimation of $q(\vec{x})$ in terms of $p(\vec{y})$ requires solution of an underdetermined system.

For continuous variables we can proceed with minimizing the relative entropy using functional calculus [13,14]. The calculation follows the same logic as in the discrete case, we show the steps explicitly for clarity and pedagogy.

The relative entropy (RE) in Equation (9) is a functional of $q(\vec{x})$. As in the discrete case, if $p(\vec{y})$ is normalized, *i.e.*, $\int dy_1 \cdots dy_m p(\vec{y}) = 1$, then Equations (10) and (11) imply $q(\vec{x})$ is normalized as well, *i.e.*,

$$\int dx_1 \cdots dx_n q(\vec{x}) = 1 \tag{12}$$

We introduce a Lagrange multiplier function, $\lambda(\vec{y})$, and generate a functional, $S_\lambda[q]$, that we need to minimize in order to minimize Equation (9) along with the constraints in Equation (10). Since, the normalization condition in Equation (12) follows from Equation (10) we do not treat Equation (12) as a separate constraint.

$S_\lambda[q]$ is given by,

$$\begin{aligned}
S_\lambda[q(\vec{x})] &= \int dx_1 \cdots dx_n \left[q(\vec{x}) \ln \left[\frac{q(\vec{x})}{u(\vec{x})} \right] \right] \\
&\quad - \int dy_1 \cdots dy_m \lambda(\vec{y}) \left[p(\{y_j\}) - \int dx_1 \cdots dx_n q(\vec{x}) \prod_{k=1}^{m} \delta_D (y_k - f_k(\vec{x})) \right] \\
&= \int dx_1 \cdots dx_n \left[q(\vec{x}) \ln \left[\frac{q(\vec{x})}{u(\vec{x})} \right] \right] - \int dy_1 \cdots dy_m \lambda(\vec{y}) p(\vec{y}) \\
&\quad - \int dx_1 \cdots dx_n \lambda(f(\vec{x})) q(\vec{x})
\end{aligned} \tag{13}$$

We can take the functional derivative to minimize S as,

$$\frac{\delta S_\lambda[q(\vec{x})]}{\delta q(\vec{x})} = \ln\left[q(\vec{x}) \right] + 1 - \ln[u(\vec{x})] - \lambda(\vec{f}(\vec{x})) = 0 \tag{14}$$

In deriving Equation (14) we used the standard relation $\frac{\delta f[x]}{\delta f[x']} = \delta_D(x - x')$. For multiple dimensions this generalizes to, $\frac{\delta f[\vec{x}]}{\delta f[\vec{x}']} = \prod_{j=1}^{n} \delta_D(x_j - x_j')$. The chain rule for derivatives of functions can be easily generalized for functional derivatives [14]. Equation (14) provides us with the solution that minimizes Equation (13):

$$\hat{q}(\vec{x}) = u(\vec{x}) e^{\lambda(\{f_i(\vec{x})\})-1} = u_0 e^{\lambda(\{f_i(\vec{x})\})-1} = \tilde{q}(\{f_i(\vec{x})\}) \tag{15}$$

where, u_0 is a constant related to the density of the uniform distribution. Note the $\{x_i\}$ dependence in the solution, $\hat{q}(\vec{x})$, arises only though $\vec{f}(\vec{x})$.

Substituting Equation (15) in Equation (10),

$$
\begin{aligned}
p(\vec{y}) &= \int dx_1 \cdots dx_n \hat{q}(\vec{x}) \prod_{k=1}^{m} \delta_D (y_k - f_k(\vec{x})) \\
&= \int dx_1 \cdots dx_n \tilde{q}(f(\vec{x})) \prod_{k=1}^{m} \delta_D (y_k - f_k(\vec{x})) \\
&= \tilde{q}(\vec{y}) \kappa(\vec{y}) \\
\Rightarrow \tilde{q}(\vec{y}) &= \frac{p(\vec{y})}{\kappa(\vec{y})}
\end{aligned}
\tag{16}
$$

where,

$$
\kappa(\vec{y}) = \int dx_1 \cdots dx_n \prod_{k=1}^{m} \delta_D (y_k - f_k(\vec{x}))
\tag{17}
$$

The second derivative gives,

$$
\frac{\delta^2 S_\lambda [q(\vec{x})]}{\delta q(\vec{x}') \delta q(\vec{x}'')} = 1 / q(\vec{x}'') \delta_D (x' - x'')
\tag{18}
$$

The second derivative of S_λ in Equation (18) is always positive, since q is positive. Therefore, $\hat{q}(\vec{x})$, minimizes the relative entropy in Equation (9). Equations (16) and (17) are the main results of this section, which are the counterparts for the Equations (6b) and (7) in discrete case.

We apply the above results for two examples below.

Example 2. Consider a linear relationship between y and x, e.g., $y = x_1 + x_2$, where, $0 \leq y \leq \infty$ and $0 \leq x_1 \leq \infty$, $0 \leq x_2 \leq \infty$. If the pdf in y is known as, $p(y) = 1/\mu \exp(-y/\mu)$, we would like to know the pdf corresponding $q(x_1, x_2)$, where, the pdfs p and q are related by Equation (10), *i.e.*,

$$
p(y) = \int_0^\infty dx_1 \int_0^\infty dx_2 q(x_1, x_2) \delta_D (y - x_1 - x_2)
$$

The degeneracy factor in the continuous case, according to Equation (17), in this case is,

$$
\begin{aligned}
\kappa(y) &= \int_0^\infty dx_1 \int_0^\infty dx_2 \delta_D (y - x_1 - x_2) = \int_0^y dx_1 \int_0^y dx_2 \delta_D (y - x_1 - x_2) \\
&= \int_0^y dx_1 \int_0^y dx_2 \delta_D (x_2 - (y - x_1)) = \int_0^y dx_1 = y
\end{aligned}
$$

The second equality results from the fact that the Dirac delta function is zero outside that region. The fourth equality uses the property of the Delta function,

$$
\int_0^y dx_2 \delta_D (x_2 - a) \text{ (where, } a = y - x_1 = const \text{ for this integration)} = 1
$$

Therefore, $\hat{q}(x_1, x_2) = \dfrac{e^{-(x_1+x_2)/\mu}}{\mu(x_1 + x_2)}$.

Example 3. Let $y = f(x_1, x_2) = x_1^2 + x_2^2$, $0 \leq y \leq \infty$ *and* $0 \leq (x_1, x_2) \leq \infty$. Then κ(y), as given by Equation (17), is,

$$\kappa(y) = \int_0^\infty dx_1 \int_0^\infty dx_2 \, \delta_D(y - x_1^2 - x_2^2)$$

$$= \int_0^\infty dr r \, \delta_D(y - r^2) \int_0^{\pi/2} d\phi$$

$$= \int_0^\infty d(r^2) \, \delta_D(y - r^2) \frac{\pi}{4}$$

$$= \frac{\pi}{4}$$

Therefore, according to Equation (17),

$$\hat{q}(y) = \frac{4p(y)}{\pi} \Rightarrow \hat{q}(x_1, x_2) = \frac{4p(f(x_1, x_2))}{\pi}$$

In our final example, we illustrate solving the problem by taking the limit when the upper and/or lower bound(s) approach $\pm \infty$, as mentioned near the beginning of this section of the paper.

Example 4. Let $y = x_1^2 + x_2^2$, $0 \leq y \leq 2L^2$ and $0 \leq (x_1, x_2) \leq L$. First we calculate κ(y) as given in Equation (17). Therefore, we need to evaluate the integral,

$$\kappa(y) = \int_0^L dx_1 \int_0^L dx_2 \delta_D(y - x_1^2 - x_2^2)$$

We divide the region of integration ($0 \leq (x_1, x_2) \leq L$) into two parts, region I (lighter shade) and II (darker shade) as shown in the Figure 1. Region I contains x_1 and x_2 values, where, $x_1^2 + x_2^2 = y^2 \leq L^2$, and, region II contains the remaining of the part of the domain ($0 \leq (x_1, x_2) \leq L$) of integration. The integrals in these regions are given by the first and the second term after the second equality sign in the equation below.

$$\kappa(y) = \int_{region\ I} \int dx_1 \, dx_2 \delta_D(y - x_1^2 - x_2^2)) + \int_{region\ II} \int dx_1 \, dx_2 \delta_D(y - x_1^2 - x_2^2))$$

$$= \int_0^L dr r \, \delta_D(y - r^2) \int_0^{\pi/2} d\phi + \int_0^L dr r \int_{\cos^{-1}(r/L)}^{\pi/2 - \cos^{-1}(r/L)} d\phi \delta_D(y - r^2)$$

In region I, where $y \leq L^2$,

$$\kappa_I(y) = \int_0^L dr r \, \delta_D(y - r^2) \int_0^{\pi/2} d\phi = \frac{\pi}{2} \frac{\sqrt{y}}{2\sqrt{y}} = \frac{\pi}{4}$$

In region II, where $L^2 \leq y \leq 2L^2$,

$$\kappa_{II}(y) = \int_0^L dr\, r \int_{\cos^{-1}(L/r)}^{\pi/2-\cos^{-1}(L/r)} d\phi\, \delta_D(y-r^2)$$

$$= \int_0^L dr\, r\, (\pi/2 - 2\cos^{-1}(L/r))\delta_D(y-r^2)$$

$$= \frac{\sqrt{y}}{2\sqrt{y}}(\pi/2 - 2\cos^{-1}(L/\sqrt{y}))$$

$$= \frac{\pi}{4} - \cos^{-1}(L/\sqrt{y})$$

$\cos^{-1}(L/\sqrt{y})$ varies between 0 (on the line $x_1^2 + x_2^2 = L^2$) and $\pi/4$ (at $x_1 = x_2 = L$). Note, $\kappa(y) = 0$ when $x_1 = x_2 = L$, which does have any degeneracy. Therefore, Equation (16) is not valid at this point. Thus, as in Example 2 and 3,

$$\hat{q}(x_1,x_2) = \begin{cases} \dfrac{4p(f(x_1,x_2))}{\pi}, & \text{when, } x_1^2 + x_2^2 \le L^2 \\[2mm] \dfrac{p(f(x_1,x_2))}{\left(\pi/4 - \cos^{-1}\left(L/\sqrt{x_1^2+x_2^2}\right)\right)}, & \text{when, } L^2 \le x_1^2 + x_2^2 < 2L^2 \end{cases}$$

Limit L→∞: When $y \le L^2$, $\kappa(y) = \pi/4$. Thus, as L→∞, as long as y remains in region I we correctly recover the result in example III. If y is in region II, then we can expand $\kappa(y)$ in a series of a small parameter $\varepsilon = (y - L^2)/L^2$ as $\kappa(y) = \pi/4 - \sqrt{\varepsilon} + O(\varepsilon^{3/2})$. This result follows from the expansion of $L/\sqrt{y} < 1$ in region II. We can write,

$$L/\sqrt{y} = 1/\sqrt{1+\varepsilon} = (1+\varepsilon)^{-1/2} = 1 - \varepsilon/2 + O(\varepsilon^2)$$

where, $0 < \varepsilon[= (y - L^2)/L^2] \le 1$. Using series expansion of $\cos^{-1}(x)$ [15] we find, $\cos^{-1}(L/\sqrt{y}) = \sin^{-1}(\sqrt{1-(L/\sqrt{y})^2}) = \sin^{-1}(\sqrt{\varepsilon/(1+\varepsilon)}) = \sqrt{\varepsilon} + O(\varepsilon^{3/2})$, and thus, $\kappa(y) = \pi/4 - \sqrt{\varepsilon} + O(\varepsilon^{3/2})$.
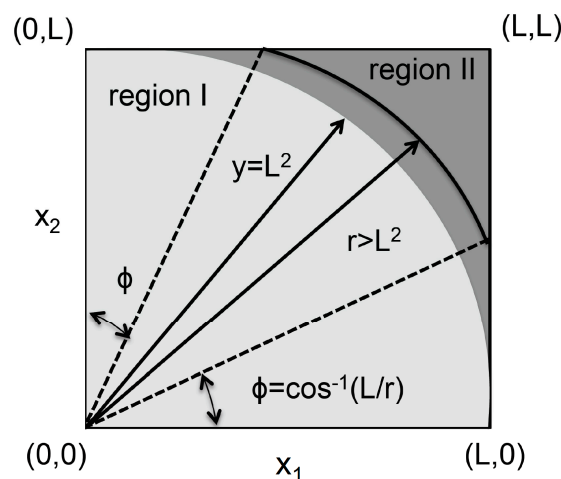

Figure 1. Shows the different regions used in calculating the integral for $\kappa(y)$ in Example 4.

5. Discussion

The problem we have attacked here arose from our work with microbial communities [3], but it also has broader statistical applications. For example, the responses of immune cells to external stimuli involve protein interaction networks, where protein-protein interactions, described by biochemical reaction rates, are not directly accessible for measurement *in vivo*. Recent developments in single cell measurement techniques allow for measuring many protein abundances in single cells, making it possible to evaluate distribution of protein abundances in a cell population [16]. However, it is a challenge to characterize protein-protein interactions underlying a cellular response because the number of these interactions could be substantially larger than the number of measured protein species [17]. These problems involve determining the distribution of a random variable x, where y is another random variable, and X and Y have a functional relationship. In the more common situation, x has dimensionality less than or equal to that of y, and there is often a unique solution. In contrast, we considered here the case where x's dimensionality is greater than that of y, so there is no unique solution to the problem.

Since there is no unique solution, we propose taking a MaxEnt approach, as a way of "spreading out the uncertainty" as evenly as possible. In the discrete case, intuition would suggest that if k values of Q sum to a given value of P, then the solution that makes the least additional assumptions is for each Q to equal P/k. This intuition is confirmed by our MaxEnt results for the discrete case. In the continuous case, the intuition is not as obvious. However, the MaxEnt solution does capture the same intuitive idea. Instead of dividing P by $k(\vec{y})$ (an integer), we divide p by $\kappa(\vec{y})$, where $\kappa(y) = \int_{\vec{x}} d\vec{x}\, \delta_D\left(y - f(\vec{x})\right)$ when *y* has dimension 1, or more generally by Equation (17). This use of the Dirac delta function has the similar effect of spreading out the uncertainty evenly.

Estimating the distribution Q(x) does not require explicit calculation of the Lagrange multipliers and the partition sum. Rather, Q(x) is directly evaluated following Equation (6b) (or Equation (17) in the continuous case), using the measured P(y) (or p(y)), and, k(y) (or κ(y)), which depends only on the relationship y = f(x). In standard MaxEnt applications, where constraints are imposed by the average values and other moments of the data, inference of probability distributions requires evaluation of the Lagrange multipliers and the partition sum Z. This involves solving a set of nonlinear equations and the relation between the Z and the Lagrange multipliers. Calculating these quantities, which is usually carried out numerically, can pose a technical challenge when the variables reside in large dimensions. In our case, we avoid these calculations and provide a solution for Q(x) in terms of a closed analytical expression, which is general and thus applicable to any well-behaved example. A limitation is that calculation of the degeneracy factor k(y) (or κ(y) in the continuous case) can present a challenge in higher dimensions and for complicated relations between y and x. Monte Carlo sampling techniques[18] and discretization schemes for Dirac delta functions[19] can be helpful in that regard.

Acknowledgments

Author Contributions

Jayajit Das, Sayak Mukherjee and Susan E. Hodge planned the research, carried out the calculations, and wrote the paper. All the authors have read and approved the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Appendix

A1. An example for MaxEnt for y = f(x) when n≤ m.

Consider a relation, $y = x^2$, where, y and x are integers, and, $0 \leq y \leq 1$ and $-1 \leq x \leq 1$. Thus, both $x = \pm 1$ are associated with $y = 1$. The pdf of X, Q(x), is related to the pdf of Y, P(y), as,

$$P(y = 0) = Q(x = 0) \tag{A1a}$$

$$P(y = 1) = Q(x = +1) + Q(x = -1) \tag{A1b}$$

Therefore, if P(y) is known, Equation (A1) cannot be used to uniquely determine Q(x) as the above system is underdetermined. We can use the MaxEnt scheme developed here to estimate Q(x). Equation (1) determines Q(0) and using Equation (6): $Q(1) = Q(-1) = P(1)/2$.

A2. Details of the MaxEnt Calculations for Example 1

$$S = -Q(0,0)\ln Q(0,0) - Q(0,1)\ln Q(0,1) - Q(1,0)\ln Q(1,0) - Q(0,2)\ln Q(0,2)$$
$$- Q(1,1)\ln Q(1,1) - Q(2,0)\ln Q(2,0)$$

The three constraints corresponding to P(y) at y = 0, y = 1, and, y = 2, are denoted by λ_1, λ_2, and λ_3, respectively. Therefore, Equation (4) in this case is given by,

$$\delta S = 0 = -\delta Q(0,0)(\ln \hat{Q}(0,0) + 1) - \delta Q(0,1)(\ln \hat{Q}(0,1) + 1) - \delta Q(1,0)(\ln \hat{Q}(1,0) + 1)$$
$$- \delta Q(0,2)(\ln \hat{Q}(0,2) + 1) - \delta Q(1,1)(\ln \hat{Q}(1,1) + 1) - \delta Q(2,0)(\ln \hat{Q}(2,0) + 1) + \lambda_1 \delta Q(0,0)$$
$$+ \lambda_2 (\delta Q(1,0) + \delta Q(0,1)) + \lambda_3 (\delta Q(2,0) + \delta Q(1,1) + \delta Q(0,2))$$

The solution for $\hat{Q}$ can be found by equating the coefficients of each of the δQ to zero since δQ is arbitrary.

$$\hat{Q}(0,0) = e^{\lambda_1 - 1}, \ \hat{Q}(1,0) = \hat{Q}(0,1) = e^{\lambda_2 - 1}, \ \hat{Q}(2,0) = \hat{Q}(0,2) = \hat{Q}(1,1) = e^{\lambda_3 - 1}$$

Using the above solution and Equation (2) we can easily find

$$\hat{Q}(0,0) = 1/3, \ \hat{Q}(1,0) = \hat{Q}(0,1) = 1/6, \ \hat{Q}(2,0) = \hat{Q}(0,2) = \hat{Q}(1,1) = 1/9$$

$\hat{Q}$ is normalized as expected.

Substituting the above equations in the constraint equation in Equation (2) (or Equation (8)) provides the values for the Lagrange multipliers, *i.e.*,

$$e^{\lambda_1 - 1} = 1/3 \Rightarrow \lambda_1 = -\ln(3) + 1; \ e^{\lambda_2 - 1} = 1/6 \Rightarrow \lambda_2 = -\ln(6) + 1; e^{\lambda_3 - 1} = 1/9 \Rightarrow \lambda_3 = -\ln(9) + 1$$

References

1. Human Microbiome Project Consortium. Structure, Function and Diversity of the Healthy Human Microbiome. *Nature* **2012**; *486*, 207–214.
2. Ley, R.E.; Hamady, M.; Lozupone, C.; Turnbaugh, P.J.; Ramey, R.R.; Bircher, J.S.; Schlegel, M.L.; Tucker, T.A.; Schrenzel, M.D.; Gordon, J.I. Evolution of Mammals and Their Gut Microbes. *Science* **2008**, *320*, 1647–1651.
3. Mukherjee, S.; Weimer, K.E.; Seok, S.C.; Ray, W.C.; Jayaprakash, C.; Vieland, V.J.; Edward Swords, W.; Das, J. Host-to-Host Variation of Ecological Interactions in Polymicrobial Infections. *Phys. Biol.* **2015**, *12*, 016003.
4. Bialek, W.S. *Biophysics: Searching for Principles*; Princeton University Press: Princeton, NJ, USA, 2012.
5. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
6. MacKay, D.J.C. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
7. Presse, S.; Ghosh, K.; Lee, J.; Dill, K.A. Principles of Maximum Entropy and Maximum Caliber in Statistical Physics. *Rev. Mod. Phys.* **2013**, *85*, 1115–1141.
8. Caticha, A. Towards an Informational Pragmatic Realism. *Mind Mach.* **2014**, *24*, 37–70.
9. Mora, T.; Bialek, W. Are Biological Systems Poised at Criticality? *J. Stat. Phys.* **2011**, *144*, 268–302.
10. Rényi, A. *Probability Theory*; Dover Books on Mathematics; Dover: Mineola, NY, USA, 2007.
11. Reif, F. *Fundamentals of Statistical and Thermal Physics*; Waveland Press: Long Grove, IL, USA, 2008.
12. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed; Wiley: Hoboken, NJ, USA, 2006.
13. Ryder, L.H. *Quantum Field Theory*, 2nd ed.; Cambridge University Press: New York, NY, USA, 1996.
14. Greiner, W.; Reinhardt, J; Bromley, D.A. *Field Quantization*; Springer: Berlin, Germany, 1996.
15. Abramowitz, M.; Stegun, I.A. Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables; Dover: Mineola, NY, USA, 1965.

16. Krishnaswamy, S.; Spitzer, M.H.; Mingueneau, M.; Bendall, S.C.; Litvin, O.; Stone, E.; Pe'er, D.; Nolan, G.P. Systems Biology. Conditional Density-Based Analysis of T Cell Signaling in Single-Cell Data. *Science* **2014**, *346*, 1250689.

17. Eydgahi, H.; Chen, W.W.; Muhlich, J.L.; Vitkup, D.; Tsitsiklis, J.N.; Sorger, P.K. Properties of Cell Death Models Calibrated and Compared Using Bayesian Approaches. *Mol. Syst. Biol.* **2013**, *9*, 644.

18. Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. *Numerical Recipes: The Art of Scientific Computing*, 3rd ed.; Cambridge University Press: New York, NY, USA, 2007.

19. Smereka, P. The Numerical Approximation of a Delta Function with Application to Level Set Methods. *J. Comput. Phys.* **2006**, *211*, 77–90.