JOURNAL OF Evolutionary Biology

doi: 10.1111/j.1420-9101.2010.02204.x

REVIEW

# A simple derivation and classification of common probability distributions based on information symmetry and measurement scale

S. A. FRANK*† & E. SMITH†

*Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, USA
†Santa Fe Institute, Santa Fe, NM, USA

*Keywords:*

population genetics;
theory.

## Abstract

Commonly observed patterns typically follow a few distinct families of probability distributions. Over one hundred years ago, Karl Pearson provided a systematic derivation and classification of the common continuous distributions. His approach was phenomenological: a differential equation that generated common distributions without any underlying conceptual basis for why common distributions have particular forms and what explains the familial relations. Pearson's system and its descendants remain the most popular systematic classification of probability distributions. Here, we unify the disparate forms of common distributions into a single system based on two meaningful and justifiable propositions. First, distributions follow maximum entropy subject to constraints, where maximum entropy is equivalent to minimum information. Second, different problems associate magnitude to information in different ways, an association we describe in terms of the relation between information invariance and measurement scale. Our framework relates the different continuous probability distributions through the variations in measurement scale that change each family of maximum entropy distributions into a distinct family. From our framework, future work in biology can consider the genesis of common patterns in a new and more general way. Particular biological processes set the relation between the information in observations and magnitude, the basis for information invariance, symmetry and measurement scale. The measurement scale, in turn, determines the most likely probability distributions and observed patterns associated with particular processes. This view presents a fundamentally derived alternative to the largely unproductive debates about neutrality in ecology and evolution.

## Introduction

Commonly observed patterns follow a few families of probability distributions. For example, Gaussian patterns often arise from measures of height or weight, and gamma patterns often arise from measures of waiting times. These common patterns lead to two questions. How are the different families of distributions related? Why are there so few families, when the possible patterns are essentially infinite?

These questions are important, because one can hardly begin to study nature without some sense of the fundamental contours of pattern and why those contours arise. For example, no one observing a Gaussian distribution of weights in a population would feel a need to give a special explanation for that pattern. The central limit theorem tells us that a Gaussian distribution is a natural and widely expected pattern that arises from measuring aggregates in a certain way.

With other common patterns, such as neutral distributions in biology or power laws in physical phenomena,

*Correspondence:* Steven A. Frank, Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697–2525, USA, and Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA. Tel.: +1 949 824 2244; fax: +1 949 824 2181; e-mail: safrank@uci.edu

the current standard of interpretation is much more variable. That variability arises because we do not have a comprehensive theory of how measurement and information shape the commonly observed patterns. Without a clear notion of what is expected in different situations, common and relatively uninformative patterns frequently motivate unnecessarily complex explanations, and surprising and informative patterns may be overlooked (Frank, 2009).

Currently, the differences between families of common probability distributions often seem arbitrary. Thus, little understanding exists with regard to how changes in process or in methods of observation may cause observed pattern to change from one common form into another.

We argue that measurement, described by the relation between magnitude and information, unifies the distinct families of common probability distributions. Variations in measurement scale may, for example, arise from varying precision in observations at different magnitudes or from the way that information is lost when measurements are made on aggregates. Our unified explanation of the different commonly observed distributions in terms of measurement points the way to a deeper understanding of the relations between pattern and process.

We develop the role of measurement through maximum entropy expressions for probability distributions. We first note that all probability distributions can be expressed by maximization of entropy subject to constraint. Maximization of entropy is equivalent to minimizing total information while retaining all the particular information known to constrain underlying pattern (Jaynes, 1957a,b, 2003). To obtain a probability distribution of a given form, one simply chooses the informational constraints such that maximization of entropy yields the desired distribution. However, constraints chosen to match a particular distribution only describe the sufficient information for that distribution. To obtain deeper insight into the causes of particular distributions and each distribution's position among related families of distributions, we derive the related forms of constraints through variations in measurement scale.

Measurement scale expresses information through the invariant transformations of measurements that leave the form of the associated probability distribution unchanged (Frank & Smith, 2010). Each problem has a characteristic form of information invariance and symmetry that sets the measurement scale (Hand, 2004; Luce & Narens, 2008; Narens & Luce, 2008) and the most likely probability distribution associated with that particular scale (Frank & Smith, 2010). We show that measurement scales and the symmetries of information invariances form a natural hierarchy that generates the common families of probability distributions. We use *invariance* and *symmetry* interchangeably, in the sense that symmetry arises when an invariant transformation leaves an object unchanged (Weyl, 1952).

The measurement hierarchy arises from two processes. First, we express the forms of information invariance and measurement scale through a continuous group of transformations, showing the relations between different types of information invariance. Second, the types of aggregation and measurement that minimize information and maximize entropy fall into two classes, each class setting a different basis for information invariance and measurement scale.

The two types of aggregation correspond to the two major families of stable distributions that generalize the process leading to the central limit theorem: the Lévy family that includes the Gaussian distribution as a special case and the Fisher-Tippett family of extreme value distributions. By expressing measurement scale in a general way, we obtain a wider interpretation of the families of stable distributions and a broader classification of the common distributions.

Our derivation of probability distributions and their familial relations supersedes the Pearson and similar classifications of continuous distributions (Johnson *et al.*, 1994). Our system derives from a natural description of varying information in measurements under different conditions (Frank & Smith, 2010), whereas the Pearson and related systems derive from phenomenological descriptions that generate distributions without clear grounding in fundamental principles such as measurement and information.

Some recent systems of probability distributions, such as the unification by Morris (1982; Morris & Lock, 2009), provide great insight into the relations between families of distributions. However, Morris's system and other common classifications do not derive from what we regard as fundamental principles, instead arising from descriptions of structural similarities among distributions. We provide a detailed analysis of Morris's system in relation to ours in Appendix C.

We favour our system because it derives the relations between distributions from fundamental principles, such as maximum entropy and the invariances that define measurement scale. Although the notion of what is fundamental will certainly attract controversy, our favoured principles of entropy, symmetries defined by invariances, and measurement scale certainly deserve consideration. Our purpose is to show what one can accomplish by starting solely with these principles.

## Maximum entropy and measurement scale

This section reviews our prior work on the roles of information invariance and measurement scale in setting observed pattern (Frank & Smith, 2010). The following sections extend this prior work by expressing measurement in terms of the scale of aggregation and the continuous group transformations of information invariance.

### Maximum entropy

The method of maximum entropy defines the most likely probability distribution as the distribution that maximizes a measure of entropy (randomness) subject to various information constraints (Jaynes, 2003). We write the quantity to be maximized as

$$\Phi = \mathscr{E} - \kappa C_0 - \sum_{i=1}^{n} \lambda_i C_i, \tag{1}$$

where $\mathscr{E}$ measures entropy, the $C_i$ are the constraints to be satisfied, and $\kappa$ and the $\lambda_i$ are the Lagrange multipliers to be found by satisfying the constraints. Let $C_0 = \int p_y \mathrm{d}y - 1$ be the constraint that the probabilities must total one, where $p_y$ is the probability distribution function of $y$. The other constraints are usually written as $C_i = \int p_y f_i(y) \mathrm{d}y - \bar{f}_i$, where the $f_i(y)$ are various transformed measurements of $y$, and the overbar denotes mean value. A mean value is either the average of some function applied to each of a sample of observed values, or an a priori assumption about the average value of some function with respect to a candidate set of probability laws. If $f_i(y) = y^i$, then $\bar{f}_i$ are the moments of the distribution – either the moments estimated from observations or a priori values of the moments set by assumption. The moments are often regarded as 'standard' constraints, although from a mathematical point of view, any properly formed constraint can be used.

Here, we confine ourselves to a single constraint of measurement. We express that constraint with a more general notation, $C_1 = \int p_y T(f_y) \mathrm{d}y - \bar{T}_f$, where $f_y \equiv f(y)$, and $T(f_y) \equiv T_f$ is a transformation of $f_y$. We could, of course, express the constraining function for $y$ directly through $f_y$. However, we wish to distinguish between an initial function $f_y$ that can be regarded as a standard measurement, in any sense in which one chooses to interpret the meaning of standard, and a transformation of standard measurements denoted by $T_f$ that arises from information about the measurement scale.

The maximum entropy distribution is obtained by solving the set of equations

$$\frac{\partial \Phi}{\partial p_y} = \frac{\partial \mathscr{E}}{\partial p_y} - \kappa - \lambda T_f = 0, \tag{2}$$

where one checks the candidate solution for a maximum and obtains $\kappa$ and $\lambda$ by satisfying the constraint on total probability and the constraint on $\bar{T}_f$. We assume that we can treat the entropy measures and the maximization procedure by the continuous limit of the discrete case.

In the standard approach, we define entropy by extension of Shannon information

$$\mathscr{E} = -\int p_y \log\left(\frac{p_y}{m_y}\right) \mathrm{d}y, \tag{3}$$

in which this expression may be called Jaynes's differential entropy (Jaynes, 2003), which is equivalent in form to the continuous expression of relative entropy or the Kullback–Leibler divergence (Cover & Thomas, 2006). Here, we will interpret $m_y$ by information invariance and measurement scale as discussed elsewhere. With these definitions, the solution of eqn (2) is

$$p_y \propto m_y e^{-\lambda T_f}, \tag{4}$$

where $\lambda$ satisfies the constraint $C_1$, and the proportionality is adjusted so that the total probability is one by choosing the parameter $\kappa$ to satisfy the constraint $C_0$.

### Information invariance and measurement scale

Maximum entropy must capture all of the available information about a particular problem. One form of information concerns transformations to the measurement scale that leave the most likely probability distribution unchanged (Jaynes, 2003; Frank, 2009; Frank & Smith, 2010). Here, it is important to distinguish between measurements and measurement scale. In our notation, we start with measurements, $f_y$, made on the measurement scale $y$. For example, we may have measures of squared deviations about zero, $f_y = y^2$, with respect to the measurement scale $y$, such that $\bar{f}_y$ is the second moment of the measurements with respect to the underlying measurement scale.

Suppose that we obtain the same information about the underlying probability distribution from measurements of $f_y$ or transformed measurements, $G(f_y)$. Put another way, if one has access only to measurements $G(f_y)$, one has the same information that would be obtained if the measurements were reported as $f_y$. We say that the measurements $f_y$ and $G(f_y)$ are equivalent with respect to information, or that the transformation $f_y \to G(f_y)$ is an information invariance that describes a symmetry of the measurement scale.

To capture this information invariance in maximum entropy, we must express our measurements so that

$$T(f_y) = \delta + \phi T[G(f_y)] \tag{5}$$

for some arbitrary constants $\delta$ and $\phi$ (Frank & Smith, 2010). Putting this definition of $T(f_y) \equiv T_f$ into eqn (4) shows that we obtain the same maximum entropy solution whether we use the observations $f_y$ or the transformed observations, $G(f_y)$, because the $\kappa$ and $\lambda$ constants will adjust to the constants $\delta$ and $\phi$ so that the distribution remains unchanged.

## Deriving probability distributions

The prior section established two key steps. First, maximum entropy probability distributions have the form given in eqn (4) as $p_y \propto m_y e^{-\lambda T_f}$. Second, the expression of $T(f_y)$ for each problem comes from the particular information invariance $G(f_y)$ associated with that

particular problem. To derive specific probability distributions, we must pass three further steps, which we develop in the following sections.

First, we turn the abstract notions of information invariance and measurement scale into specific expressions for the measurement scale function, $T(f_y)$. We accomplish this by developing the continuous group transformations for information invariance. Those continuous transformations provide an abstract hierarchy of forms for probability distributions based on the scale factor, $m_y$, the specific measured attribute, $f_y$, and how the information and precision of measurements change with magnitude expressed by the measurement scale $T(f_y)$.

Second, we define $m_y$ as the relation between the scale of information invariance and the scale on which we express probability. To use the maximization of entropy and the associated minimization of information, we must relate the information invariance of measurement to the scale on which underlying processes dissipate information. We consider alternative interpretations of scale that may be associated with the dissipation of information by aggregation of random perturbations and by measurements of extreme values. We also consider measurements on a scale that differs from the basis for dissipation of information.

Third, we consider how to interpret $f_y$, which is the value used to describe the informational constraint in relation to the measurement scale $T(f_y)$, leading to the constraint $\bar{T}_f$. We discuss $f_y$ as a reduction in the dimensionality of information to a single sufficient dimension. That sufficient dimension sets the form of probability under the various processes of information dissipation that lead to the common probability distributions.

## Continuous group transformations of measurement

The transformation in eqn (5) sets the relation between information invariance and measurement scale. However, that expression does not show in a simple way the relations between information and measurement.

To understand commonly observed patterns in relation to the families of probability distributions, it is helpful to express in a general way the underlying symmetry that determines information invariance and measurement scale. From that underlying symmetry, we may see more clearly the associated relations between the forms of probability distributions.

### The affine structure

The relation between information invariance and measurement scale in eqn (5) arises directly from the form of maximum entropy solutions in eqn (4), in which probability distributions are exponentials of the transformed constraint measures, $T_f$. In particular, the probability distribution associated with a constraint is invariant to an additive shift of the constraint and a multiplicative change in the scale of the constraint, given by the parameters $\delta$ and $\phi$ in eqn (5). It is that symmetry in the affine structure of invariant transformation that ultimately sets the underlying relations between information, measurement and familial forms of the common probability distributions.

To understand the affine structure of the invariant transformation in eqn (5) more clearly, we can express that invariant transformation as a continuous operator. First, rearrange eqn (5) as an equivalent expression

$$T[G(f_y)] = a + bT(f_y) \tag{6}$$

with new parameters $a$ and $b$ that are easily calculated from eqn (5). We show in Appendix A that we can express the same information invariance of $G(f_y)$ by the differential operator defined as

$$v_w = (\alpha + \beta T)\frac{\mathrm{d}}{\mathrm{d}T} \tag{7}$$

that can be applied to $T$ as

$$v_w(T) = \alpha + \beta T. \tag{8}$$

Recursive application of $v_w$ preserves the affine structure and so keeps the successive transformations within the group of admissible invariance relations.

We can express $v_w$ as

$$v_w = \frac{\mathrm{d}}{\mathrm{d}w}, \tag{9}$$

where $w \equiv w(f_y)$ is some function of $f_y$. We then have a differential equation for $T$ as

$$\frac{\mathrm{d}T}{\mathrm{d}w} - \beta T = \alpha, \tag{10}$$

which has solutions of the general form

$$T(f_y) = T_0 e^{\beta w} + \frac{\alpha}{\beta}(e^{\beta w} - 1), \tag{11}$$

which as $\beta \to 0$ goes to $T(f_y) \to T_0 + \alpha w$. eqn (11) gives the most general class of measurement functions, $T(f_y)$, for which the associated transformations generated by $v_w$ preserve information invariance.

The operator $v_w$ can be applied repeatedly, creating a recursively generated sequence of deformations that all satisfy the fundamental relation between deformations of measurement and information invariance. By thinking of $w(f_y)$ as a parameter that expresses the deformation of measurement associated with a measurement scale, $T(f_y)$, we can create a sequence in which each successive deformation corresponds to a successive class of probability distributions with familial relations to each other defined by the structure of the sequence of deformations to $w(f_y)$.

### The general form of probability distributions

From eqn (4), the maximum entropy solution is

$$p_y \propto m_y e^{-\lambda T_f}. \quad (12)$$

From eqn (11), we can now express the maximum entropy solution as

$$p_y \propto m_y e^{-\Lambda e^{\beta w}}, \quad (13)$$

where $\Lambda = \lambda(T_0 + \alpha/\beta)$, and $w \equiv w(f_y)$. In the limit $\beta \to 0$, this becomes

$$p_y \propto m_y e^{-\gamma w}$$

where $\gamma = \lambda\alpha$.

In Appendix B, we describe the case of extreme values, for which we will use $m_y = \mathrm{d}T(f_y)/\mathrm{d}y$. When $f_y = y$ and $m_y = \mathrm{d}T(y)/\mathrm{d}y = T'$, it will be convenient to write

$$T' \propto w' e^{\beta w}, \quad (14)$$

where $w' = \mathrm{d}w(y)/\mathrm{d}y$, and as $\beta \to 0$, $T' \propto w'$.

## Intuitive description of measurement and probability

Intuitively, one can think of the symmetry of information invariance and measurement scale in the following way. On a linear scale, each incremental change of fixed length yields the same amount of information or surprise independently of magnitude. Thus, if we change the scale by multiplying all magnitudes by a constant, we obtain the same pattern of information relative to magnitude. In other words, the linear scale is invariant to multiplication by a constant factor so that, within the framework of maximum entropy subject to constraint, we obtain the same information about probability distributions from an observation $y$ or $G(y) = cy$. In this section, we use $f_y = y$ to isolate the symmetry expressed by particular choices of $T$ and $G$.

On a logarithmic scale, each incremental change in proportion to the current magnitude yields the same amount of information or surprise. Information is scale dependent. We obtain the same information at any point on the scale by comparing ratios. For example, we gain the same information from the increment $\mathrm{d}y/y = \mathrm{d}\log(y)$ independently of the magnitude of $y$. Thus, we achieve information invariance with respect to ratios by measuring increments on a logarithmic scale. Within the framework of maximum entropy subject to constraint, we obtain the same information about probability distributions from an observation $y$ or $G(y) = y^c$, corresponding to informationally equivalent measurements $T(y) = \log(y)$ and $T(y^c) = c\log(y)$ (see Frank & Smith, 2010).

The form of a probability distribution under maximum entropy can be read directly as an expression of how the measurement scale changes with magnitude. From the general solution in eqn (4), linear scales $T(y) \propto y$ yield distributions that are exponential in $y$, whereas logarithmic scales $T(y) \propto c\log(y)$ yield distributions that are linear in $y^c$. Exponential distributions of the form $e^{-\lambda y}$ arise from underlying linear scales, whereas power law distributions of the form $y^{-c}$ arise from underlying logarithmic scales.

Many common distributions have compound form, in which one can read directly how the underlying measurement scale changes with magnitude. For example, the gamma distribution has form $y^{-c}e^{-\lambda y}$. When the magnitude of $y$ is small, the shape of the distribution is dominated by the power law component, $y^{-c}$. As the magnitude of $y$ increases, the shape of the distribution is dominated by the exponential component, $e^{-\lambda y}$. Thus, the underlying measurement scale grades from logarithmic at small magnitudes to linear at large magnitudes. Indeed, the gamma distribution is exactly the expression of an underlying measurement scale that grades from logarithmic to linear as magnitude increases. Nearly, every common probability distribution can be read directly as a simple expression of the change in the underlying measurement scale with magnitude.

## Hierarchies of common probability distributions

Given a particular form for the function $w(f_y)$, the measurement scale $T(f_y)$ follows from eqn (11) and the associated probability distribution follows from eqn (13). Although we can choose $w$ in any way that we wish, certain measurement scales and information invariances are likely to be common. We discussed in our earlier paper the importance of two scales (Frank & Smith, 2010). The first scale grades from linear to logarithmic as magnitude increases, which we call the linear-log scale. The second scale inverts the linear-log scale to be logarithmic at small magnitudes and linear at large magnitudes, giving the log-linear scale. The inversion relating the two scales can be expressed by a Laplace transform, showing the natural duality of the scales and a connection to recent studies on superstatistics (Frank & Smith, 2010).

### The linear-log scale

In terms of the notation in the present paper, we can define $w$ to establish a hierarchy of measurement deformations, in which each level in the hierarchy arises from successive application of the linear-log scaling to the scale in the previous level in the hierarchy.

To define the linear-log measurement function in terms of $w$, note from eqn (11) that, as $\beta \to 0$, the forms of $w$ and the measurement function $T$ become equivalent with respect to setting the associated probability distribution. Thus, by setting $w$, we are defining the limiting form of the measurement function. With these issues in mind, define

**Table 1** The logarithmic measurement hierarchy.*

| $w(f_y)$ | $p_y$ | $p_{y\|\beta\to 0}$ |
|---|---|---|
| $f_y$ | $m_y e^{-\Lambda e^{\beta f_y}}$ | $m_y e^{-\gamma f_y}$ |
| $\log f_y$ | $m_y e^{-\Lambda f_y^{\beta}}$ | $m_y f_y^{-\gamma}$ |
| $\log \log f_y$ | $m_y e^{-\Lambda(\log f_y)^{\beta}}$ | $m_y(\log f_y)^{-\gamma}$ |

*$p_y$ is the form of the probability distribution function from eqn (13). Note that $\beta \to 0$ of each line corresponds to $\beta = 1$ of the following line.

$$w^{(i)} = \log(c_i + w^{(i-1)}),$$

with $w^{(0)} = f_y$. The constant $c_i$ sets the transition between linear and logarithmic scaling: the scale is linear when $w^{(i-1)}$ is small relative to $c_i$ and logarithmic when $w^{(i-1)}$ is large relative to $c_i$. As $c_i \to 0$, we can use $w^{(i)} = \log(w^{(i-1)})$.

It is easiest to see the abstract structure of the measurement hierarchy and the associated forms of probability distributions in the limiting case $c_i \to 0$, leading to purely logarithmic deformations. The first row of Table 1 begins with the base measurement $w^{(0)} = f_y$. The following two rows show the first two deformations for the sequence $i = 0,1,2$.

This table gives the hierarchy of probability distributions that arise from successive logarithmic deformations. With this structure in mind, we give the full expansion with $c_i \neq 0$ in Table 2.

We discuss the interpretation of $m_y$ and $f_y$ below. The different interpretations of those values lead directly to specific forms for probability distributions. Before interpreting $m_y$ and $f_y$, we present an alternative measurement scale.

### The log-linear scale

We obtain the log-linear measurement deformation hierarchy (Frank & Smith, 2010) from

$$w^{(i)} = c_i w^{(i-1)} + \log(w^{(i-1)}),$$

from which we obtain the probability distributions in Table 3. The log-linear scale changes logarithmically at small magnitudes and linearly at large magnitudes.

### Other scales

The linear-log and log-linear scales describe common forms of measurement functions. In this section, we briefly mention some other scales listed in Table 4. These additional scales illustrate the ways in which measurement relates to the patterns of probability.

The first line of Table 4 shows a log-linear-log scale for a measure on the interval $(c_1,c_2)$. That scale changes logarithmically near the boundaries and linearly near the middle of the range, in which $\log b$ describes the skew in the scaling pattern.

**Table 2** The linear-log measurement hierarchy.

| $w(f_y)$ | $p_y$ | $p_{y\|\beta\to 0}$ |
|---|---|---|
| $f_y$ | $m_y e^{-\Lambda e^{\beta f_y}}$ | $m_y e^{-\gamma f_y}$ |
| $\log\,(c_1 + f_y)$ | $m_y e^{-\Lambda(c_1+f_y)^{\beta}}$ | $m_y(c_1+f_y)^{-\gamma}$ |
| $\log(c_2 + \log(c_1 + f_y))$ | $m_y e^{-\Lambda(c_2+\log(c_1+f_y))^{\beta}}$ | $m_y(c_2+\log(c_1+f_y))^{-\gamma}$ |

**Table 3** The log-linear measurement hierarchy.*

| $w(f_y)$ | $p_y$ | $p_{y\|\beta\to 0}$ |
|---|---|---|
| $f_y$ | $m_y e^{-\Lambda e^{\beta f_y}}$ | $m_y e^{-\gamma f_y}$ |
| $c_1 f_y + \log f_y$ | $m_y e^{-\Lambda f_y^{\beta} e^{c_1 \beta f_y}}$ | $m_y f_y^{-\gamma} e^{-c_1 \gamma f_y}$ |
| $c_2(c_1 f_y + \log f_y) + \log(c_1 f_y + \log f_y)$ | $m_y e^{-\Lambda e^{\beta w}}$ | $m_y e^{-\gamma w}$ |

*In the last line of the table, we use $w \equiv w(f_y)$ to shorten the expression.

**Table 4** Alternative measurement scales.*

| $w(f_y)$ | $p_y$ | $p_{y\|\beta\to 0}$ |
|---|---|---|
| $\log((c_2 - f_y)(f_y - c_1)^b)$ | $m_y e^{-\Lambda(c_2-f_y)^{\beta}(f_y-c_1)^{b\beta}}$ | $m_y(c_2-f_y)^{-\gamma}(f_y-c_1)^{-b\gamma}$ |
| $c_2 f_y + b\,\log(c_1 + f_y)$ | $m_y e^{-\Lambda(c_1+f_y)^{b\beta} e^{c_2\beta f_y}}$ | $m_y(c_1+f_y)^{-b\gamma} e^{-c_2\gamma f_y}$ |

*As $\beta \to 0$, line 1 is a log-linear-log measurement scale, and line 2 is a linear-log-linear measurement scale.

The second line of Table 4 shows a linear-log-linear scale for $f_y > 0$. That scale changes linearly near the lower boundary of zero, linearly at large magnitudes and logarithmically at intermediate values.

## The scale of information

The prior section presented probability distributions in terms of $m_y$ and $f_y$. This section develops the interpretation of $m_y$, which arises from the relation between the scale of information invariance and the scale on which we express probability.

The key issue is that maximum entropy requires some underlying process to dissipate information. With regard to deriving probability distributions, we may consider three aspects of scale in relation to the dissipation of information. First, we may measure an outcome that arises from the aggregation of a series of random perturbations. Second, we may measure only the extreme values of some underlying process thereby throwing away all information about the underlying process except the form of the upper or lower tail of the underlying distribution. Third, the dissipation of information may occur on one scale, but we may wish to make our measurements with respect to another scale.

Each of these three aspects of the scale of information dissipation leads to a simple interpretation of probability measure in maximum entropy analysis. We give a brief

description of each scale of information dissipation in relation to calculating $m_y$.

### Aggregation of perturbations

In the standard application of maximum entropy, accumulation of random perturbations without constraint leads to a uniform probability measure, which has maximum entropy and minimum information. Thus, the scale at which information dissipates is the same as the scale of the probability measure. In this case, our formulation of maximum entropy has $m_y \equiv 1$, because any information that arises from deformation of measurement relative to the uniform default is included in our expression of measurement scale, $T(f_y)$.

### Extreme values

The distribution of extreme values depends only on the total (integral) of the probability measure in the tail of an underlying probability distribution (Embrechts *et al.*, 1997). Because extreme value distributions arise from integrals of probability measures, the dissipation of information and the associated measurement scale for extreme values are expressed in terms of the cumulative distribution function (see Appendix B). To obtain the associated form of the probability measure with respect to the probability distribution function, $p_y$, we must transform the invariant measurement scale originally expressed with respect to the integral of the underlying probability measure.

To change from the integral scale of the cumulative distribution to the scale of the probability measure associated with the probability density function, we simply differentiate the initial measurement scale, $T(f_y)$, from the cumulative distribution scale to obtain the associated change in probability measure (Appendix B). For $f_y = y$, we obtain $m_y = \mathrm{d}T(y)/\mathrm{d}y = T'$. We gave the general form of $m_y = T'$ in eqn (14).

### Change of variable

In some cases, information may dissipate on one scale, but we choose to express probability on another scale. The log-normal distribution is the classic example. Using Table 1, we may consider measurements that lead to the Normal or Gaussian distribution by either analysing squared deviations from a central value, $f_y = (y - \mu)^2$ in line one of Table 1 with $\beta \to 0$ or, equivalently, linear perturbations of $f_y = (y - \mu)$ in line two of Table 1 with $\beta = 2$. In these cases, the perturbations are direct measures rather than the tail probabilities of extreme values, so $m_y = 1$, and we have the standard form of the Gaussian as $p_y \propto e^{-\gamma(y-\mu)^2}$.

If we prefer to analyse values on a logarithmic scale, then we make the transformation $y \to \log y$. This case does not arise from invariant information and the associated measurement transformation, but rather from a change of variable to a different scale. So we must change our measure, as in any standard change of variable. In this case, the change of measure is $m_y \mathrm{d}y = \mathrm{d}\log y = \mathrm{d}y/y$, thus $m_y = y^{-1}$ and we obtain the log-normal distribution $p_y \propto y^{-1} e^{-\tilde{\gamma}(\log y - \tilde{\mu})^2}$, where $\tilde{\gamma}$ and $\tilde{\mu}$ are transformed appropriately.

## Sufficiency: reduction of information

The algorithm of maximum entropy allows us to choose any constraint $T(f_y)$. However, one of our main goals is to provide a clear rationale for the choice of constraint, so that maximum entropy is more than a tautological description of probability distributions. We have expressed the choice of the measurement scale, $T$, in terms of information invariance set by the underlying problem. Although information invariance may take various forms, we followed our earlier paper (Frank & Smith, 2010) in which we defended the linear-log and log-linear scales as likely to be common scales associated with common information invariances.

Once we have set the transformation $T(f_y)$ by these common information invariances, many widely observed probability distributions follow. In some cases, deriving probability distributions requires using an observable, $f_y \neq y$, that differs from the scale $y$ of the underlying probability measure. For example, we may use the squared deviations from a central location, or a fractional moment $f_y = y^\alpha$, where $\alpha$ is not an integer (Frank, 2009). Use of $f_y = y$ or of squared deviations $f_y = (y - \mu)^2$ is widely accepted. Such choices lead to $f_y$ being a sufficient reduction of all of the information in observations in order to express common probability distributions.

For our purposes in this paper, we simply note that we can derive many common distributions by the widely accepted use of $f_y = y$ or $f_y$ as a squared deviation. But the reasons that particular choices of $f_y$ are sufficient have not been fully explained with regard to maximum entropy, particularly fractional moments such as $f_y = y^\alpha$ (Frank, 2009). Those reasons probably have to do with the sort of analysis described by large deviation theory (Touchette, 2009), in which the retained information arises from the minimal descriptions of location and scale that remain when one normalizes the consequences of a sequence of perturbations so that one obtains a stable limiting form.

## Conclusion

Table 5 shows many of the commonly observed probability distributions. Those distributions arise directly from maximum entropy applied to various natural measurement scales. The measurement scales express information invariances associated with particular types of problems and the scale on which information dissipation occurs. We confined ourselves to various combinations

**Table 5** Some common probability distributions.*

| Distribution | $p_y$ | T.L.C | $m_y$ | $f_y$ | Notes and alternative names |
|---|---|---|---|---|---|
| Gumbel | $e^{\beta y-\Lambda e^{\beta y}}$ | 1.1.2 | $T'$ | $y$ | |
| Gibbs/Exponential | $e^{-\gamma y}$ | 1.1.3 | $T'$,1 | $y$ | |
| Gauss/Normal | $e^{-\gamma y^2}$ | 1.1.3 | 1 | $y^2$ | |
| Log-Normal | $y^{-1}e^{-\gamma(\log y)^2}$ | 1.1.3 | $y^{-1}$ | $y^2$ | Change of variable $y \to \log y$ |
| Fréchet/Weibull | $y^{\beta-1}e^{-\Lambda y^\beta}$ | 1.2.2 | $T'$ | $y$ | |
| Stretched exponential | $e^{-\Lambda y^\beta}$ | 1.2.2 | 1 | $y$ | Gauss with $\beta = 2$ |
| Symmetric Lévy | $e^{-\Lambda\lvert y\rvert^\beta}$ (Fourier domain) | 1.2.2 | 1 | $\lvert y\rvert$ | $\beta \le 2$; Gauss ($\beta = 2$), Cauchy ($\beta = 1$)† |
| Pareto type I | $y^{-\gamma}$ | 1.2.3 | $T'$,1 | $y$ | |
| Log-Fréchet | $y^{-1}(\log y)^{\beta-1}e^{-\Lambda(\log y)^\beta}$ | 1.3.2 | $T'$ | $y$ | Also from Fréchet: $y \to \log y$, $m_y = y^{-1}T'(y)$ |
| ?? | $e^{-\Lambda(\log y)^\beta}$ | 1.3.2 | 1 | $y$ | Also stretched exponential with $f_y = \log y$ |
| Log-Pareto type I | $y^{-1}(\log y)^{-\gamma-1}$ | 1.3.3 | $T'$ | $y$ | Log-gamma; Pareto I: $y \to \log y$, $m_y = y^{-1}$ |
| ?? | $(\log y)^{-\gamma}$ | 1.3.3 | 1 | $y$ | Also from Pareto I with $f_y = \log y$ |
| Pareto type II | $(c_1 + y)^{-\gamma}$ | 2.2.3 | 1 | $y$ | Lomax |
| Generalized Student's | $(c_1 + y^2)^{-\gamma}$ | 2.2.3 | 1 | $y^2$ | Pearson type VII, Kappa; includes Cauchy |
| ?? | $(\log(c_1 + y))^{-\gamma}$ | 2.3.3 | 1 | $y$ | $c_2 = 0$; also Pareto I with $f_y = \log(c_1+y)$ |
| Gamma | $y^{-\gamma}e^{-c_1\gamma y}$ | 3.2.3 | 1 | $y$ | Pearson type III, includes chi-square |
| Generalized gamma | $y^{-k\gamma}e^{-c_1\gamma y^k}$ | 3.2.3 | 1 | $y^k$ | Chi with $k = 2$ and $c_1\gamma = 1/2$ |
| Beta | $(c_2 - y)^{-\gamma}(y - c_1)^{-b\gamma}$ | 4.1.3 | 1 | $y$ | Pearson type I; log-linear-log on $(c_1,c_2)$ |
| Beta prime/F | $y^{-b\gamma}(1 + y)^{(b+1)\gamma}$ | 4.1.3 | 1 | $\frac{y}{1+y}$ | Pearson type VI, $y > 0$ |
| Gamma variant | $(c_1+y)^{-b\gamma}e^{-c_2\gamma y}$ | 4.2.3 | 1 | $y$ | Linear-log-linear pattern as $y$ rises from zero |

*The column T.L.C gives the table, line and column of the underlying form presented in the earlier tables of abstract distributions. For example, 1.1.2 refers to Table 1, first line, second column. The measurement adjustment is given as either $m_y = 1$ for direct scales, or $m_y = T'$ for extreme values as in eqn (14), along with any consequences from a change of variable such as $y \to \log y$. Cases in which the same structural form arises for either $m_y = T'$ or $m_y = 1$ are shown as $T'$,1, without adjusting parameters for trivial differences. The value of $f_y$ gives the reduction of data to sufficient summary form. Direct values $y$, possibly corrected by displacement from a central location, $y - \mu$, are shown here as $y$ without correction. Squared deviations $(y - \mu)^2$ from a central location are shown here as $y^2$. Listings of distributions can be found in various texts (Johnson *et al.*, 1994, 1995; Kleiber & Kotz, 2003). Many additional forms can be generated by varying the measurement function. In the first column, the question marks denote a distribution for which we did not find a commonly used name.
†See Frank (2009).

of linear and logarithmic scaling, which were sufficient to express many common distributions. Our method readily extends to other types of information invariance and measurement scale and their associated probability distributions.

## Acknowledgments

SAF's research is supported by National Science Foundation grant EF-0822399, National Institute of General Medical Sciences MIDAS Program grant U01-GM-76499, and a grant from the James S. McDonnell Foundation. ES's work is supported by Insight Venture Partners.

## References

Cover, T.M. & Thomas, J.A. 2006. *Elements of Information Theory*, 2nd edn. Wiley, Hoboken, NJ.

Embrechts, P., Kluppelberg, C. & Mikosch, T. 1997. *Modeling Extremal Events: For Insurance and Finance*. Springer Verlag, Heidelberg.

Frank, S.A. 2009. The common patterns of nature. *J. Evol. Biol.* **22**: 1563–1585.

Frank, S.A. & Smith, D.E. 2010. Measurement invariance, entropy, and probability. *Entropy* **12**: 289–303.

Hand, D. 2004. *Measurement Theory and Practice*. Arnold, London.

Jaynes, E.T. 1957a. Information theory and statistical mechanics. *Phys. Rev.* **106**: 620–630.

Jaynes, E.T. 1957b. Information theory and statistical mechanics II. *Phys. Rev.* **108**: 171–190.

Jaynes, E.T. 1968. Prior probabilities. *IEEE Trans. Syst. Sci. Cybern.* **4**: 227–241.

Jaynes, E.T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press, New York.

Jeffreys, H. 1957. *Scientific Inference*, 2nd edn. Cambridge University Press, London.

Johnson, N.L., Kotz, S. & Balakrishnan, N. 1994. *Continuous Univariate Distributions*, 2nd edn, Vol. 1. Wiley, New York.

Johnson, N.L., Kotz, S. & Balakrishnan, N. 1995. *Continuous Univariate Distributions*, 2nd edn, Vol. 2. Wiley, New York.

Kleiber, C. & Kotz, S. 2003. *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley, New York.

Kotz, S. & Nadarajah, S. 2000. *Extreme Value Distributions: Theory and Applications*. World Scientific, Singapore.

Luce, R.D. & Narens, L. 2008. Measurement, theory of. In: *The New Palgrave Dictionary of Economics* (S.N. Durlauf & L.E. Blume, eds) Palgrave Macmillan, Basingstoke. Available from: http://www.dictionaryofeconomics.com/extract?id=pde2008_M000128.

Mahan, G.D. 2000. *Many Particle Physics*, 3rd edn. Springer, New York.

Morris, C.N. 1982. Natural exponential families with quadratic variance functions. *Ann. Stat.* **10**: 65–80.

Morris, C.N. 1983. Natural exponential families with quadratic variance functions: statistical theory. *Ann. Stat.* **11**: 515–529.
Morris, C.N. & Lock, K.F. 2009. Unifying the named natural exponential families and their relatives. *Am. Stat.* **63**: 247–253.
Narens, L. & Luce, R.D. 2008. Meaningfulness and invariance. In: *The New Palgrave Dictionary of Economics* (S.N. Durlauf & L.E. Blume, eds), Palgrave Macmillan, Basingstoke. Available from: http://www.dictionaryofeconomics.com/extract?id=pde2008_M000121.
Sato, K. 2001. Basic results on Lévy processes. In: *Lévy Processes: Theory and Applications* (O.E. Barndorff-Nielsen, T. Mikosch & S.I. Resnick, eds), pp. 3–37. Birkäuser, Boston.
Seidenfeld, T. 1979. Why I am not an objective Bayesian: some reflections prompted by Rosenkrantz. *Theory Decis.* **11**: 413–440.
Touchette, H. 2009. The large deviation approach to statistical mechanics. *Phys. Rep.* **478**: 1–69.
Weyl, H. 1952. *Symmetry*. Princeton University Press, Princeton.

## Appendix A: On the association between measurement functions and classes of scale transformations

If the transformation $f_y \to G(f_y)$ is an invariance of a measurement function $T$, it is clear that repeated applications of $G$, expressed as $G\circ G, G\circ G\circ G, \ldots$, are also invariances of $T$. It is the larger group of invariances that we wish to identify with the measurement scale that defines $T$, and not only a single transformation. To simplify notation in this Appendix, we use $f_y = y$. The same analysis applies to $f_y$.

In general, making a unique association between a transformation $G$ and a measurement function $T$ is inconvenient for finite transformations, because $G$ combines a magnitude and a direction of deformation. The magnitude is added under compositions $G\circ G\ldots$, while the direction remains invariant. As we will derive below, the relevant measure of the magnitude of a transformation as in eqn (6) will be $\sim \log b$, and the relevant measure of direction will be $a/(b-1)$. To isolate the direction of $G$ that may be associated with a measurement function $T$, we work with infinitesimal rather than finite affine transformations.

Infinitesimal transformations are constructed from eqn (6) in the text by writing $a \equiv \epsilon\alpha$, $(b-1) \equiv \epsilon\beta$ and then taking $\epsilon \to 0$ for fixed $\alpha$ and $\beta$. An infinitesimal transformation $G^\epsilon$ then satisfies eqn (6) in the form

$$T[G^\epsilon(y)] = T(y) + \epsilon[\alpha + \beta T(y)]. \tag{15}$$

$G$ itself must therefore also be infinitesimally different from the identity and must have the form

$$G^\epsilon(y) = y + \epsilon v(y). \tag{16}$$

for some function $v(y)$.

We introduce a quantity $\hat{v}$ called the *generator* of the deformation, such that the operator $e^{\epsilon\hat{v}}$ generates the infinitesimal transformations eqns (15, 16), and such that finite transformations $G$ or affine transformations eqn (6) are produced by the exponential operation of $\hat{v}$ with noninfinitesimal $\epsilon$. Compounding a function corresponds to addition of parameters $\epsilon$, as may be checked from the power-series definition of $e^{\epsilon\hat{v}}$ within its radius of convergence.

We define a *representation* of the generator $\hat{v}$ as an explicit differential operator that produces the correct transformation on the argument $y$ or $T(y)$, as appropriate. The two representations of the generators are related as

$$\begin{aligned} T[y + \epsilon v(y)] &= \left[1 + \epsilon v(y)\frac{\mathrm{d}}{\mathrm{d}y}\right] T(y) \\ &= \left[1 + \epsilon(\alpha + \beta T)\frac{\mathrm{d}}{\mathrm{d}T}\right] T. \end{aligned} \tag{17}$$

From the requirement that the two expressions produce the same result, we may assign the representations

$$\begin{aligned} \hat{v} &\leftrightarrow v(y)\frac{\mathrm{d}}{\mathrm{d}y} \equiv \frac{\mathrm{d}}{\mathrm{d}w} \\ &\leftrightarrow (\alpha + \beta T)\frac{\mathrm{d}}{\mathrm{d}T} \end{aligned} \tag{18}$$

for some function $w(y) = \int^y \mathrm{d}y'\, 1/v(y')$.

Regarding $T$ as a function of argument $w$ rather than $y$, and setting equal the two coefficients of $\epsilon$ in eqn (17), we obtain a relation between any function $w(y)$, coefficients $\alpha$ and $\beta$ and the function $T$ in the form

$$\frac{\mathrm{d}T}{\mathrm{d}w} = \alpha + \beta T. \tag{19}$$

This is rearranged to produce eqn (10).

From the solutions to eqn (19), we may readily check that the action of the transformation $e^{\epsilon\hat{v}}$ for arbitrary $\epsilon$ (not necessarily small) is

$$T[G(y)] = e^{\epsilon\hat{v}}T(y) = \frac{\alpha}{\beta}(e^{\epsilon\beta} - 1) + e^{\epsilon\beta}T(y), \tag{20}$$

from which we recover expressions for the coefficients $a$ and $b$ in eqn (6). Under composition $G \to G\circ G$, the parameter $\epsilon \to 2\epsilon$. The composition rules for $a$ and $b$ under composition of $G$ may be worked out easily, but depending on the function $w(y)$, the direct composition of finite transformations $G$ on $y$ may be quite complicated.

## Appendix B: Information measures for cumulative distributions

The presence of the measure $m_y$ in the probability density function in eqn (12) complicates the discussion of measurement invariance, because in the general case $m_y$ is not required to obey any prescribed transformation when $f_y \to G(f_y)$. In general, $y$ need not even be a numerical index, whereas $T(f_y)$ is necessarily numerical because it is proportional to an information measure $-\log(p_y/m_y)$.

The class of cases in which the measurement function, $T$, completely controls the properties of $p_y$ are

those in which measurement constrains the *cumulative* probability distribution function rather than the probability density function. For these cases, $m_y$ is not independent, but is given in terms of $T$ and $f_y$, as we now show.

Relative entropy is ordinarily defined for the probability density. However, if we set

$$m_y = \frac{\mathrm{d}T(f_y)}{\mathrm{d}y} = \frac{\mathrm{d}f_y}{\mathrm{d}y} T'(f_y), \tag{21}$$

then $m_y$ becomes a Lebesgue measure on $y$ with respect to the increment $\mathrm{d}T$. The probability density from eqn (12) becomes

$$p_y \propto \frac{\mathrm{d}}{\mathrm{d}y} e^{-\lambda T(f_y)}. \tag{22}$$

Equation (22) defines the relation between a probability density and its cumulative distribution, meaning that under a suitable ordering of $y$, we may take $e^{-\lambda T(f_y)}$ to be the cumulative distribution.

With this choice of measure, the relative entropy $\mathscr{E}$ from eqn (3) becomes

$$\begin{aligned} -\int \mathrm{d}y\, p_y \log\left(\frac{p_y}{m_y}\right) &= -\int \mathrm{d}y \frac{\mathrm{d}T}{\mathrm{d}y}\left(\frac{p_y}{\mathrm{d}T/\mathrm{d}y}\right) \log\left(\frac{p_y}{\mathrm{d}T/\mathrm{d}y}\right) \\ &= -\int \mathrm{d}T p_T \log p_T, \end{aligned} \tag{23}$$

in which $p_T$ is the probability density defined on the variable $T$. Since the maximum entropy solution is always exponential in $T$, the relative entropy of eqn (23) is effectively an information function for the cumulative distribution.

An application in which constraints under aggregation apply by construction to the cumulative distribution is the computation of extreme-value statistics (Kotz & Nadarajah, 2000). The cumulative probability distribution for the maximizer or minimizer of a sample of $n$ realizations of a random variable is the product of $n$ factors of the cumulative distribution for a single realization.

It was also noted in Frank (2009) that the relative entropy may be evaluated on the characteristic function (Fourier or Laplace transform) of a distribution and that the maximum entropy solutions in the transformed domain are the Lévy stable distributions. The characteristic function at frequency argument $k = 0$ always takes value unity. Therefore it, like a cumulative distribution, has a reference normalization of unity, and indeed, the symmetric Lévy-stable distributions (Sato, 2001) correspond in form to the Weibull family of extreme value distributions. Both are obtained within our classification for $m_y$ defined by eqn (21), for suitable reductions $f_y$.

## Appendix C: The Morris Natural Exponential Families in relation to entropy-maximizing distributions

### Symmetry-based approaches to select or to classify probability distributions

Many systems, since Pearson's, for either selecting or classifying probability distributions, have been based on symmetry groups, as our method is. (Pearson's system may be seen as one based on the analytic structure of the log-probability, a criterion that we will return to consider in a moment.) The systems differ in generality, depending on the space in which the symmetry group acts and depending on whether it constrains a single distribution or a family. Two methods based on symmetry (ours and that of Carl Morris, described below) have interpretations in terms of *scale invariance* of observables. Both systems collect probability distributions into families, whose members differ only by a scale factor. A third approach (known as *Objective Bayesian* methods) applies symmetry to the underlying measure space, which may be very different from the space of observed magnitudes. This approach is concerned not directly with families of distributions, but with the particular distribution defined by a reference measure. We will briefly summarize the overlaps and differences of these methods.

Objective Bayesian methods, initiated by Jeffreys (1957) but given the interpretation of objectivity largely by Jaynes (1968), recognize that the reference measure $m_y$ in a relative entropy – beyond being needed to make logarithms well-defined and independent of change of variables – may reflect information about measurement scales. By ensuring that the reference measure is consistent with known symmetries of the phenomenon under study (which are not generally expressed within particular sample observations), Objective Bayesian methods seek to systematize the entire maximum entropy procedure. This use of the reference measure is consistent with our treatment of measurement, although by itself it is more limited, as we discuss in Frank & Smith (2010), and it may also be misleading in cases (Seidenfeld, 1979). In the context of the present discussion, the most important limitation of Objective Bayesian methods is that they select properties of a single distribution $m_y$, rather than properties of a family.

Our approach broadens the class of symmetries that can be considered, beyond those available to Objective Bayesian methods, as discussed in Sec. 7 of Frank & Smith (2010). Through the measurement function, it relates a potentially nonlinear contour of deformations of measured magnitudes to a linear transformation within the affine group that exists for general maximum entropy problems. We have embedded distributions within a hierarchy by using the two-parameter freedom of the affine group to provide a range of

responses of information to the change in the scale of measurement.

### The Morris classification of distributions in relation to maximum entropy

In a pair of papers, Carl Morris (1982, 1983) proposed another classification system for probability distributions, which overlaps both with Objective Bayesian methods and with our approach. Like our method, Morris's concerns families of probability distributions generated by a change in constraint or measurement scale. Like all of the approaches we have mentioned, Morris's system uses relative entropy in a conventional maximization framework. That system differs from ours in using only a linear constraint on what Morris terms the *natural observation* and obtaining nonlinear dependence on that constraint through a second boundary condition placed on entropy.

The Morris system blends interesting elements of Pearson's restrictions on analytic structure, our use of symmetry and the Objective Bayesian concern with the reference measure, as follows: Morris considers distribution families that are invariant under offset and rescaling of the natural observation, which Morris labels $X$, and which is analogous to using a coordinate system that is always linear in our $f_y$. His classification therefore does not invoke any explicit representation of the symmetries inherent in differing measurement systems. In order to encompass distributions that are not simply exponential in the values $x$ (taken by the observation $X$), he instead restricts the form of the reference measure in a relative entropy, analogous to our $m_y$. Unlike Objective Bayesian methods, however, this restriction does not come from the direct action of a symmetry on the reference measure, but rather from the form of the relative entropy across the family of distributions produced by scale change.

The classification system of Morris (1982) derives from the cumulant-generating function, and the relation between the variance and the mean as the parameter in this generating function is shifted. The distributions that define the cumulant-generating function constitute what Morris calls *natural exponential families* (NEF), and the dependence of variance on mean within these families is restricted in his system to be an exact quadratic polynomial. The resulting subclass of distributions within the NEF class is termed QVF (for *quadratic variance function*). The mean–variance relation that defines the NEF-QVF distributions is preserved under offset and rescaling of the natural observation, and under convolution. Therefore, the distributions in this class would be expected to arise frequently in problems of aggregation. We show in this appendix that the QVF condition is equivalent to the requirement that the relative entropy over a family has the form (up to analytic continuation) of a Kullback–Leibler divergence. The analytic continuation is determined by the roots of the quadratic variance polynomial, and these roots in turn have a relation to the roots for log-probability in the Pearson system.

The distributions selected by Morris's criterion are bounded or have exponential or faster decay in their tails. We show that, when they are classified according to their analytic structure, they are in fact either interior members or degenerate limits of only two families of distributions: One family of continuous-valued distributions is associated with complex-conjugate roots of the variance function, and a complex analytic continuation of the Kullback–Leibler form for relative entropy. A second family of discrete-valued distributions is associated with real-valued roots and real-valued continuations of the Kullback–Leibler relative entropy. In this sense, the Morris classification shows that six important distribution families are in fact selected by a single set of invariances – of these, the offset and scale invariances are instances of our linear measurement rescaling. These selected families are therefore very commonly observed, but also rather tightly restricted. Preservation of a functional class under convolution is similar to the criterion leading to the extreme value or Lévy distributions, as we have discussed in the main text, and is therefore one of many forms of measurement invariance that may be considered.

Here, we will re-formulate the Morris criterion and its solutions within a standard framework of maximum entropy. We will show that the role of the reference measure in a relative entropy is equivalent to that of a *second* observed quantity, which will generally be linearly independent of the natural observation $X$. Scale change of the natural observation defines what is known as an expansion path, which consists of the distributions within an exponential family. The second observed quantity, associated with the reference measure, is given a gradient constraint rather than a value constraint. It is through the interaction of these two constraints that nonlinear dependence on $x$ is obtained in the log-probability. At the end of the Appendix, we mention a relation between the Morris system and the Pearson system based on the log-probability. When the Morris QVF criterion is expressed as a formal constraint on entropy, this form is imposed on the leading terms of log-probability by the large-deviations property of cumulant-generating functions.

### Definition of the natural exponential families

The NEF distributions are defined in relation to the cumulant-generating function, which arises naturally in the method of maximum entropy. The most direct way to re-formulate the original presentation of Morris (1982, 1983) in terms of maximum entropy is to assume a (Shannon-type) entropy in a higher-dimensional state space than the univariate space of the natural observation $X$. The high-dimensional states have nonuniform density

when they are projected onto the one dimension in which the probability distribution varies. Once a Lagrangian is defined from this initial re-formulation, it becomes easy to re-interpret the density of states as a reference measure in a relative entropy (and the latter interpretation is more general). The cumulant-generating function is then the Legendre transform of this relative entropy. We develop the two interpretations in order, to connect the derivations of Morris (1982, 1983) systematically to the formulation we use in the main text.

*The Stieltjes measure as a density of states*

Morris (1982) introduces a Stieltjes measure $\mathrm{d}F(x)$, and an initial probability distribution $P_0$ associated with this measure, defined by

$$P_0(X \in A) = \int_A \mathrm{d}F(x), \tag{24}$$

for an arbitrary set $A$ in the range of $x$. With respect to this original probability measure, Morris introduces the exponential families in terms of a probability mass function

$$\phi(x \mid \theta) = e^{x\theta - \psi(\theta)}, \tag{25}$$

which multiplicatively weights the original measure $\mathrm{d}F(x)$. The normalizing constant $\psi(\theta)$ in eqn (25) is the cumulant-generating function, given by

$$e^{\psi(\theta)} \equiv \int \mathrm{d}F(x) e^{x\theta}. \tag{26}$$

The NEF distributions are the normalized versions of the distributions that define the cumulant-generating function.

In the original Stieltjes measure, the probabilities defined from these distributions are

$$P(X \in A) = \int_A \mathrm{d}F(x)\, e^{x\theta - \psi(\theta)}. \tag{27}$$

With respect to the measure $\mathrm{d}F(x)$, we may obtain the solutions (25) by extremizing the Lagrangian

$$\begin{aligned} \mathscr{L} = &- \int \mathrm{d}F(x)\, \phi(x) \log \phi(x) \\ &+ \theta\left( \int \mathrm{d}F(x)\, \phi(x) x - \mu \right) \\ &- \kappa\left( \int \mathrm{d}F(x)\, \phi(x) - 1 \right) \end{aligned} \tag{28}$$

over its natural argument $\phi(x)$ and the Lagrange multipliers $\theta$ and $\kappa$. Here, we have replaced the notation $\lambda$ from the text with Morris's $\theta$ for ease of reference. From its role as a normalization constant, the multiplier $\kappa$ must evaluate to the cumulant-generating function $\psi(\theta)$ on solutions.

Lagrangian problems of this form arise frequently in systems where a high-dimensional state space is projected down onto a single coordinate $x$, which is the only observed property on which distributions depend. The Lagrangian (28) effectively treats $\phi(x)$ as the ratio of a probability density to a *uniform* reference measure on the original high-dimensional space. The Stieltjes measure $\mathrm{d}F(x)$ is the marginal projection of the original measure onto the coordinate $x$, and the derivative $\mathrm{d}F/\mathrm{d}x$ is known as the *density of states*. ($\mathrm{d}F(x)$ need not be smooth, and $\mathrm{d}F/\mathrm{d}x$ may readily be a noncontinuous distribution, such as a sum of Dirac $\delta$-functions, representing a discrete rather than continuous probability density).

The entropy in this formulation appears as a standard Shannon entropy (equivalent to a relative entropy with a uniform reference measure) in the high-dimensional coordinates. It evaluates to the Legendre transform of the cumulant-generating function,

$$\begin{aligned} S(\mu(\theta)) &\equiv -\int \mathrm{d}F(x)\, p(x \mid \theta) \log p(x \mid \theta) \\ &= -\theta\mu(\theta) + \psi(\theta), \end{aligned} \tag{29}$$

in which $\mu(\theta)$ is the mean value in the distribution $p(x \mid \theta)$. $\theta$ is the natural argument of $\psi$, while $\mu$ from the variational problem is the natural argument of $S$. Therefore, it is usual to write this Legendre transform pair as

$$\begin{aligned} \psi(\theta) &= S(\mu) - \mu \left.\frac{\mathrm{d}S}{\mathrm{d}\mu}\right|_{\mu=\mu(\theta)} \\ S(\mu) &= \psi(\theta) - \theta \left.\frac{\mathrm{d}\psi}{\mathrm{d}\theta}\right|_{\theta=\theta(\mu)} \end{aligned} \tag{30}$$

In the second line, $\theta(\mu)$ is the inverse function to $\mu(\theta)$. (In statistical mechanics, where $-\theta$ is the inverse temperature if $x$ is the energy, $\psi$ arises as $\theta$ times the *Helmholtz Free Energy*.)

We note several properties of these functions that will be useful in understanding Morris's NEF-QVF families. When $\theta = 0$ no correction to the normalization is needed in $P(X \in A)$, so we have immediately that $\psi(0) = 0$ as well. If we denote by $\mu_0 \equiv \mu(0)$, then it follows that $S(\mu_0) = 0$ also. The definition of the Legendre transform pair (30) gives the important dual relations

$$\begin{aligned} \frac{\mathrm{d}\psi(\theta)}{\mathrm{d}\theta} &= \mu(\theta) \\ \frac{\mathrm{d}S(\mu)}{\mathrm{d}\mu} &= -\theta(\mu). \end{aligned} \tag{31}$$

It follows that $\mathrm{d}S/\mathrm{d}\mu|_{\mu_0} = 0$. With these two constants of integration, $S(\mu)$ will be completely specified by the form of its second derivative.

*Replacing the density of states with a reference measure in relative entropy*

For the univariate distributions, whether continuous or discrete, we may define a shorthand for eqn (27) by identifying the probability density function on $x$ as

$$p_{x|\theta} \equiv \frac{\mathrm{d}F}{\mathrm{d}x} e^{x\theta - \psi(\theta)}. \tag{32}$$

The Lagrangian (28) becomes, under this change of variable,

$$\mathscr{L} = -\int \mathrm{d}x\, p_x \log\left(\frac{p_x}{\mathrm{d}F/\mathrm{d}x}\right) + \theta\left(\int \mathrm{d}x\, p_x x - \mu\right) - \kappa\left(\int \mathrm{d}x\, p_x - 1\right). \quad (33)$$

The constraint terms are unchanged, but the entropy is now manifestly a *relative entropy* for the density $p_x$ with reference measure $\mathrm{d}F/\mathrm{d}x$.

### *Arriving at nonlinear expansion paths through mixed boundary conditions*

The Morris families, like the Pearson families and like our classes based on measurement, include distributions that are nonlinear in the values $x$ taken by the natural observation $X$. Both Morris's families and ours are based on affine transformation, so that their distributions form what are known as *expansion paths* [this term is used also in economics for constrained maximization problems, in which $\mu$ generally describes a budget constraint; the original usage, in statistics, is mentioned in Morris (1982)]. Whereas we achieve nonlinear dependence on $x$ by considering the symmetries of measurement, the Morris system achieves nonlinearity through the use of mixed boundary conditions, when this system is described in terms of entropy maximization. By using two constraints – one to specify the family and the other to fix a point on the expansion path – Morris is able to apply a fixed-gradient condition with respect to one constraint, and a fixed-value condition for the natural observation. Because we specify distributions from the affine transformation of a single observable, we must incorporate nonlinearities into the measurement function itself.

Here, we derive the NEF criterion by converting the relative entropy to a form in which the reference measure may be interpreted as a second observable. The ubiquitous use, in statistical physics and thermodynamics, of cumulant-generating functions and their Legendre transforms under mixed boundary conditions, provides intuition from familiar systems for the meaning of the resulting expansion paths. In the next section, we derive the way in which the QVF condition of Morris then places constraints on the reference measure, which plays the role of the secondary observation.

The Lagrangian (33) is an instance of a more general class of maximum entropy problems in which the relative entropy has uniform measure (and therefore has the form of a Shannon entropy), and the reference measure appears as an additional constraint term,

$$\mathscr{L} = -\int \mathrm{d}x\, p_x \log p_x + \theta\left(\int \mathrm{d}x\, p_x x - \mu\right) + \lambda \int \mathrm{d}x\, p_x \log\left(\frac{\mathrm{d}F}{\mathrm{d}x}\right) - \kappa\left(\int \mathrm{d}x\, p_x - 1\right). \quad (34)$$

Here, a variable $\lambda$ has been added as a *parameter* in the variational problem, parallel to the parameter $\mu$ in the constraint on $\int \mathrm{d}x p_x x$. When $\lambda = 1$, eqn (34) reduces to eqn (33), and the choice of reference measure does not matter because it cancels in the two logarithms. For more general $\lambda$, a uniform reference measure is explicitly required to make the logarithms well defined. The distribution solving eqn (34) is

$$p_x = e^{\lambda \log(\mathrm{d}F/\mathrm{d}x) + \theta x - \kappa}. \quad (35)$$

The Shannon entropy of eqn (34) is maximized subject to mixed constraints, which may be seen as follows. The entropy with two constraint terms is a function of two arguments $S(\mu,\xi)$, where $\xi = <\log(\mathrm{d}F/\mathrm{d}x)>$ at the given values of $\lambda$ and $\mu$. Then, $\lambda = -\partial S/\partial\xi$, just as $\theta = -\partial S/\partial\mu$ from eqn (31). Because $\mu$ is an argument to the entropy, whereas $\lambda$ is a gradient, problems of this sort resemble solutions to differential equations under mixed Dirichlet and Neumann boundary conditions.

The set of distributions (35), as $\lambda$ is held fixed and $\mu$ is varied, make up the expansion path for the entropy with respect to constraint $\int \mathrm{d}x\, p_x x$. The natural exponential families are the distributions on this expansion path, given a gradient constraint with respect to the observable $\int \mathrm{d}x\, p_x \log(\mathrm{d}F/\mathrm{d}x)$.

## The subset of natural exponential families with quadratic variation

Any reference measure may in principle form the basis for an expansion path with mixed constraints. In contrast to Objective Bayesian methods, in which $\log(\mathrm{d}F/\mathrm{d}x)$ is constrained by symmetry, the Morris system constrains reference measures by restricting the form of the variance function – equivalent to restricting the form of the *entropy* – along the nonlinear expansion path.

### *The QVF family and Kullback–Leibler entropies*

The definition of the cumulant-generating function is that, not only does $\mathrm{d}\psi/\mathrm{d}\theta = \mu$, but $\mathrm{d}^2\psi/\mathrm{d}\theta^2$ is the variance of the observation $X$. Morris defines its relation to the mean $\mu$ as a *variance function* $V(\mu)$. The *quadratic variance relation* is the dependence

$$\frac{\mathrm{d}\mu}{\mathrm{d}\theta} = v_0 + v_1\mu + v_2\mu^2. \quad (36)$$

By definition of $\theta(\mu)$ and $\mu(\theta)$ as inverse functions, it follows that the variance is also the (geometric and algebraic) inverse of the curvature of the relative entropy. We differentiate the second line in eqn (30) twice and substitute eqn (36), to produce

$$\frac{\mathrm{d}^2 S}{\mathrm{d}\mu^2} = -\frac{\mathrm{d}\theta}{\mathrm{d}\mu} = \frac{-1}{v_0 + v_1\mu + v_2\mu^2}. \quad (37)$$

Because we have first and second constants of integration from the relations following eqn (30), eqn (37) has an unambiguous integral. To assign meaning to this integral, however, and in the process to expose a relation between the Morris and Pearson approaches to classifi-

cation, we first factor the variance function into an overall normalization and the roots of the polynomial. Write

$$v_0 + v_1\mu + v_2\mu^2 \equiv v_2(\mu - \mu_1)(\mu - \mu_2), \tag{38}$$

with the solutions

$$\mu_{1,2} = -\frac{v_1}{2v_2} \mp \sqrt{\left(\frac{v_1}{2v_2}\right)^2 - \frac{v_0}{v_2}}. \tag{39}$$

Then, the integral of eqn (37) becomes

$$\begin{aligned} v_2 S = &\left(\frac{\mu_2 - \mu}{\mu_2 - \mu_1}\right) \log\left(\frac{\mu_2 - \mu}{\mu_2 - \mu_0}\right) \\ &+ \left(\frac{\mu - \mu_1}{\mu_2 - \mu_1}\right) \log\left(\frac{\mu - \mu_1}{\mu_0 - \mu_1}\right). \end{aligned} \tag{40}$$

If we denote by $\varphi \equiv (\mu - \mu_1)/(\mu_2 - \mu_1)$, the analytic continuation of a partition of the unit interval, we may write eqn (40) as

$$\begin{aligned} v_2 S &= (1 - \varphi) \log\left(\frac{1 - \varphi}{1 - \varphi_0}\right) + \varphi \log\frac{\varphi}{\varphi_0} \\ &= D(\vec{\varphi}\,||\vec{\varphi}_0). \end{aligned} \tag{41}$$

In the second line, we use $\vec{\varphi}$ to stand for the 'probability distribution' $(\varphi, 1 - \varphi)$ on two atoms, and likewise for $\vec{\varphi}_0$. $D(\vec{\varphi}\,||\vec{\varphi}_0)$ is the Kullback–Leibler divergence of $\vec{\varphi}$ from the distribution $\vec{\varphi}_0$ defined by the equilibrium mean $\mu_0$ and the variance function. The standard form for the curvature of a Kullback–Leibler divergence $S$ may be written

$$v_2(\mu_2 - \mu_1)^2 \frac{\mathrm{d}^2 S}{\mathrm{d}\mu^2} = \frac{1}{\varphi(1 - \varphi)}. \tag{42}$$

A slight variation on the formula (40), making use of forms (39) for the roots, the Legendre transform relations (30) and the constants of integration, reads

$$\begin{aligned} 2v_2\psi(\theta) + v_1\theta &= \log\left(\frac{(\mu_2 - \mu)(\mu - \mu_1)}{(\mu_2 - \mu_0)(\mu_0 - \mu_1)}\right) \\ &= \log\left(\frac{\varphi(1 - \varphi)}{\varphi_0(1 - \varphi_0)}\right). \end{aligned} \tag{43}$$

This integral relation between the cumulant-generating function and the variance function appears as eqn 3.7 in Morris (1982).

*Two fundamental NEF-QVF families and various limits*

Working in terms of the signs and magnitudes of the coefficients $v_0$, $v_1$, $v_2$, Morris identifies exactly six inequivalent natural exponential families with quadratic variance functions. Three are continuous (Gaussian, gamma, and hyperbolic-cosecant probability density functions), and three are discrete (binomial, negative binomial, and Poisson probability mass functions), up to offset and scaling of the natural observation $X$. We will see here that, working in terms of the analytic structure of the entropy (40), and a simple classification of the roots $\mu_{1,2}$, we may identify two main classes, corresponding to the continuous and discrete distributions, and various limiting forms of these, which complete Morris's families.

The quantity that distinguishes the continuous from the discrete NEF-QVF families is the discriminant $d \equiv v_1^2 - 4v_0v_2 = 4v_2^2(\mu_2 - \mu_1)^2$ (which is unchanged by offset of $X$). In the case where $d > 0$, the variance function (36) has two real roots, while if $d < 0$, it has two complex-conjugate roots. By choice of offset and scale, we may obtain Morris's canonical families by making the complex-conjugate roots purely imaginary when $d < 0$, or by taking one of the two real roots to lie at the origin if $d > 0$.

We begin with the imaginary roots, which select the continuous-valued NEF-QVF distributions. The canonical form for these is obtained when $v_1 \equiv 0$, and $v_0, v_2 > 0$. We may then define

$$\mu_{1,2} \equiv \mp i\Lambda, \tag{44}$$

with $\Lambda \equiv \sqrt{v_0/v_2}$.

The relative entropy, about a distribution $p_{x|0}$ in the NEF-QVF family with mean $\mu_0$, must have the form

$$\begin{aligned} v_2 S &= \frac{1}{2}\log\left(\frac{\Lambda^2 + \mu^2}{\Lambda^2 + \mu_0^2}\right) + \frac{\mu}{\Lambda}\left[\tan^{-1}\left(\frac{\mu_0}{\Lambda}\right) - \tan^{-1}\left(\frac{\mu}{\Lambda}\right)\right] \\ &= \frac{1}{2}\log\left(\frac{\Lambda^2 + \mu^2}{\Lambda^2 + \mu_0^2}\right) + \frac{\mu}{\Lambda}\left[\tan^{-1}\left(\frac{\Lambda}{\mu}\right) - \tan^{-1}\left(\frac{\Lambda}{\mu_0}\right)\right]. \end{aligned} \tag{45}$$

The relation of $\theta$ to $\mu$ and $\mu_0$ is

$$v_2\theta = \frac{1}{\Lambda}\left[\tan^{-1}\left(\frac{\mu}{\Lambda}\right) - \tan^{-1}\left(\frac{\mu_0}{\Lambda}\right)\right]. \tag{46}$$

If we choose a background in which $\mu_0 = 0$ (by freedom to offset $X$), it follows that we may write the cumulant-generating function as

$$v_2\psi = \frac{1}{2}\log\left(1 + \tan^2(v_2\Lambda\theta)\right). \tag{47}$$

The canonical normalization for this family of distributions is given by $v_2 = 1$. One may check directly that they are produced by the family of hyperbolic-cosecant density functions

$$p_{x|0} = \frac{1}{\Lambda}\frac{1}{e^{\pi x/2\Lambda} + e^{-\pi x/2\Lambda}} \tag{48}$$

(The proof is by contour integral. Check that

$$\begin{aligned} \cos(\Lambda\theta)e^{\psi(\theta)} &= \frac{1}{\pi}\int_0^\infty \frac{\mathrm{d}u}{1 + u^2}\left((iu)^{\tilde{\theta}} + (-iu)^{\tilde{\theta}}\right) \\ &= \frac{1}{\pi}\int_{-\infty}^\infty \frac{\mathrm{d}u(iu)^{\tilde{\theta}}}{1 + u^2} = 1, \end{aligned} \tag{49}$$

with integration variable $u \equiv e^{\pi x/2\Lambda}$ and shifted parameter $\tilde{\theta} \equiv 2\Lambda\theta/\pi$. The contour that avoids branch cuts, in the log-

transform to variables $u$, closes in the negative-imaginary half-plane, encircling the pole $u = -i$.) The distributions at $\Lambda = 1$ are the canonical densities given in Morris (1982), eqn 4.2.

It is straightforward to check that, as $\Lambda \to \infty$, the relative entropy (45) reduces to the form

$$S \to -\frac{(\mu - \mu_0)^2}{2v_0}, \tag{50}$$

for a Gaussian distribution

$$p_{x|0} = \frac{1}{\sqrt{2\pi v_0}} e^{-(x-\mu_0)^2/2v_0} \tag{51}$$

with arbitrary mean. We have used $v_2\Lambda^2 \equiv v_0$ as $v_2 \to 0$.

In the other limit, as $\Lambda \to 0$, it is convenient to take $v_2 = 1/q \equiv 1/\mu_0$, in which case we recover the relative entropy

$$S \to \mu_0 - \mu + \mu_0 \log\left(\frac{\mu}{\mu_0}\right), \tag{52}$$

appropriate to the standard gamma distribution

$$p_{x|0} = \frac{1}{\Gamma(q)} x^{(q-1)} e^{-x}. \tag{53}$$

Two of the three continuous-valued NEF-QVF families, therefore, are degenerate limits of the hyperbolic-cosecant distribution, which represents the generic case.

The discrete-valued families, following when the variance function has real roots, may be handled in similar fashion. We choose canonical forms by offsetting $x$ to set $\mu_1 = 0$ and attain this in the variance function by taking $v_0 \to 0$. The canonical scale for $x$ is then given by taking $v_1 = 1$.

For the discrete distributions, there are two 'interior' families of solutions (the binomial and negative binomial) and one limiting family (the Poisson) that may be reached from either of them. The root $\mu_2 = -v_1/v_2$ in all cases. To obtain the binomial distribution on $N$ samples with mean $\mu_0 = pN$,

$$p_{x|0} = \binom{N}{x} p^x (1-p)^{N-x}, \tag{54}$$

we take $\mu_2 = N$, corresponding to $v_2 = -1/N$. For this distribution only, the range is finite, $0 \le x \le N$. The relative entropy takes the standard form of a Kullback–Leibler divergence without extending the definition of $\varphi$ by analytic continuation,

$$\begin{aligned} S &\to -N\left\{\left(1-\frac{\mu}{N}\right)\log\left(\frac{1-\mu/N}{1-p}\right) + \frac{\mu}{N}\log\left(\frac{\mu/N}{p}\right)\right\} \\ &= -N\left\{\left(1-\frac{\mu}{N}\right)\log\left(\frac{1-\mu/N}{1-\mu_0/N}\right) + \frac{\mu}{N}\log\left(\frac{\mu}{\mu_0}\right)\right\} \\ &= -ND(\vec{\mu}/N||\vec{p}). \end{aligned} \tag{55}$$

The negative binomial distribution is immediately obtained by taking $N \to -N$ in the second line of eqn (55) while holding $\mu_0$ fixed. The corresponding distribution is

$$p_{x|0} = \binom{N-1+x}{x} p^x (1-p)^N, \tag{56}$$

with $p = \mu_0/(N + \mu_0)$. This is the other 'interior' solution, with $\mu_2 = -N$ and therefore $v_2 = 1/N$.

The Poisson distribution is the limit of either of the previous two forms as $v_2 \to 0$, so $\mu_2 \to \pm\infty$, at $p = \mu_0$ fixed. The distribution is

$$p_{x|0} = e^{-\mu_0} \frac{\mu_0^x}{x!}, \tag{57}$$

and the entropy becomes

$$S \to \mu - \mu_0 - \mu \log\left(\frac{\mu}{\mu_0}\right), \tag{58}$$

For either of the negative binomial or the Poisson, the range of $x$ is unbounded, $x \ge 0$.

The relative entropy expressions (52, 58) for the gamma and the Poisson distributions are the same functional form, under exchange of the reference mean $\mu_0$ with the distribution mean $\mu$. Their respective distributions are likewise interchanged under exchange of $x$ with $\mu_0$, except that in the gamma case (53), a further shift $\mu_0 \to \mu_0 - 1$ must be performed as well. We will return to integer shifts of this form in the next section.

(We note that the association of imaginary roots with continuous-valued distributions, and of real roots with discrete-valued distributions, is a defining structural feature of quantum mechanical distributions for particles with finite temperature but continuous time-dependence (Mahan, 2000). This is one of many interesting connections to the NEF-QVF families that it will not be possible to explore in this publication.)

## Relations to the Pearson system through large-deviations formulae

It is instructive to compare the forms for the entropies of the distributions in the NEF-QVF families to the logarithms of the probability densities or mass functions themselves. By virtue of the entropy as a large-deviations measure (Touchette, 2009), it and the log-probability will coincide to leading exponential order for sufficiently sharply peaked distributions.

The entropy is defined in the Morris system as a second integral of a rational function with two poles. The $\log p_{x|0}$ is defined in the Pearson system similarly, except that it is a first-integral of a rational function with two poles (Johnson *et al.*, 1994). The difference between these two degrees of integration leads to noncoincidence of the two families, although in many parameter limits they overlap.

We begin by comparing the continuous distributions. For the Gaussian, the two functions are identical up to a constant

$$\begin{aligned} \log p_{x|0} &= -\frac{(x-\mu_0)^2}{2v_0} - \frac{1}{2}\log(2\pi v_0) \\ S &= -\frac{(\mu-\mu_0)^2}{2v_0}. \end{aligned} \tag{59}$$

For the standard gamma with mean $\mu_0 = q$,

$$\begin{aligned} \log p_{x|0} &= q - 1 - x + (q-1)\log\left(\frac{x}{q-1}\right) \\ S &\approx q - x + q\log\left(\frac{x}{q}\right), \end{aligned} \tag{60}$$

in which the $\approx$ in the second line keeps the first two terms in Stirling's formula for $\log\Gamma(q)$. The functions are identical in form but differ by an offset $q \to q-1$.

The hyperbolic-cosecant density shows the least similarity in its domain of small argument. However, at small $\Lambda$, where it is sharply peaked, and at fixed $x$ or $\mu$, the following expansion becomes informative,

$$\begin{aligned} \log p_{x|0} &= -\frac{\pi|x|}{2\Lambda} - \log\Lambda - \log\left(1 + e^{-\pi|x|/\Lambda}\right) \\ S &= -\frac{\mu}{\Lambda}\tan^{-1}\left(\frac{\mu}{\Lambda}\right) - \log\Lambda + \frac{1}{2}\log\left(\mu^2 + \Lambda^2\right). \end{aligned} \tag{61}$$

For $\mu/\Lambda \gg 1$, $\tan^{-1}(\mu/\Lambda) \to \mathrm{sgn}(\mu)\pi/2$, giving the same two leading terms for $x$ and for $\mu$.

The discrete distributions behave similarly. For the binomial,

$$\begin{aligned} \log p_{x|0} &\approx -ND\left(\frac{\vec{x}}{N}\middle\|\frac{\vec{\mu}_0}{N}\right) \\ S &= -ND\left(\frac{\vec{\mu}}{N}\middle\|\frac{\vec{\mu}_0}{N}\right), \end{aligned} \tag{62}$$

and for the Poisson

$$\begin{aligned} \log p_{x|0} &\approx x - \mu_0 - x\log\left(\frac{x}{\mu_0}\right) \\ S &= \mu - \mu_0 - \mu\log\left(\frac{\mu}{\mu_0}\right), \end{aligned} \tag{63}$$

where again $\approx$ stands for the first two terms in Stirling's formula for factorials. Within these approximations, the two functions are identical. The negative binomial differs by terms at $\mathcal{O}(x/N)$, but within a similar Stirling approximation, it may be written

$$\begin{aligned} \log p_{x|0} &\approx (N+x)\log\left(\frac{N+x}{N+\mu_0}\right) - x\log\left(\frac{x}{\mu_0}\right) \\ &\quad + (N+x)\log\left(1 - \frac{1}{N+x}\right) - N\log\left(1 - \frac{1}{N}\right) \\ &\quad - \log\left(1 + \frac{x}{N-1}\right) \\ &\approx (N+x)\log\left(\frac{N+x}{N+\mu_0}\right) - x\log\left(\frac{x}{\mu_0}\right) \\ &\quad - \mathcal{O}\left(\frac{x}{N-1}\right) \\ S &= (N+\mu)\log\left(\frac{N+\mu}{N+\mu_0}\right) - \mu\log\left(\frac{\mu}{\mu_0}\right). \end{aligned} \tag{64}$$

The leading terms, corresponding to the analytic continuation of the Kullback–Leibler form, again coincide. The only differences arise from shifts $N \to N-1$ in a subset of terms, similar to the shift $q \to q-1$ in eqn (60).

The equivalence of $\log p_{x|0}$ and $S$ to leading exponential order is a consequence of the *large-deviations property* (Touchette, 2009) for these distributions. The cumulant-generating function is the integral of the shifted density,

$$e^{\psi(\theta)} = \int \mathrm{d}x\, p_{x|0} e^{\theta x}. \tag{65}$$

The exponential of the entropy cancels the absolute magnitude of the inserted weight factor $e^{\theta x}$ near the maximum of the shifted distribution, because for sharply peaked distributions the maximum is near $x \approx \mu$,

$$e^{S(\mu)} = \int \mathrm{d}x\, p_{x|0} e^{\theta(\mu)(x-\mu)}. \tag{66}$$

$S(\mu)$ is therefore approximately equal to $p_{x|0}$, evaluated at $x \approx \mu$. Thus, the Morris restriction to quadratic variance functions implies that $\log p_{x|0}$, at leading order, will equal the analytic continuation of a function of Kullback–Leibler form.

*Received 12 November 2010; accepted 16 November 2010*