Entropy, information and inference

Wences Gouveia, Fernando S. de Moraes and John A. Scales
Department of Geophysics and Center for Wave Phenomena
Colorado School of Mines
Golden, Colorado 80401, USA

ABSTRACT
The goal of inversion is to make quantitative inferences about physical systems from observations. Since such inferences are invariably uncertain, owing to uncertainties in the data, in the specification of the model and in the accuracy of the forward modeling operator, a conservative guiding principle would seem to be: draw the least informative conclusions that are consistent with the available information. This is especially important when it comes to the specification of those prejudices we hold as to what makes a model reasonable or not. These prejudices can be based on theory or observation but must enter the inference process since otherwise we may just as easily compute unreasonable models that are consistent with the data. A possible strategy for conservatively quantifying prior information is to find that distribution which has the least information content (or maximum entropy) among all distributions that agree with the available information. In order to carefully lay out the logic of this approach, and its potential pitfalls, we will give an overview of the foundations of Bayesian entropy-based methods in inference. Starting with the thermodynamic definition of entropy, which is fundamentally linked to the extent to which one can specify the state of a system, we show how this leads naturally to statements about the uncertainty of certain experiments involving the state of thermodynamic systems. It is a short leap from this to the information-theoretic use of entropy developed by Shannon. We will see that a certain set of postulates, which we believe any reasonable measure of information must satisfy, leads inevitably to Shannon's definition of entropy, but that in order make the concept rigorous and useful in inversion we must abandon the notion of absolute states of information and instead focus on measuring the information of one state relative to another (reference) state. We compute Bayesian probability distributions by maximizing the entropy of the distribution subject to constraints on some number of estimated moments. Using only first and second moments of the prior information we get a Gaussian approximation to the prior distribution. Using higher-order moments affords a better approximation to some data, especially data that have an asymmetrical distribution. We will show examples of one-dimensional calculations using moments up to fourth order. The application of these methods to n-dimensional problems is a significant computational challenge; we show an approach that eliminates the need of doing multi-dimensional quadrature.

Key words: Entropy, Information, Statistical inference, Bayesian prior distributions, Inverse theory

Introduction

The goal of inversion is to make quantitative inferences about physical systems from observations. An *inference* is a conclusion drawn from certain premises. Inasmuch as today's conclusions form tomorrow's premises, conclusions and premises can both be viewed as relative states of information. *Information* is the unifying concept behind all inverse calculations. When we contemplate such a calculation we must begin by surveying the information

available to us from which we can draw conclusions. This information comes in many forms and describes what we know about the errors in the data, the distribution of reasonable model parameters and the accuracy of the forward modeling procedure. Some of this information is absolutely certain and results from purely logical argument. The fact that density is positive is not a claim subject to experimental verification. A second level of information comes as the result of a theoretical argument. The fact that the speed of light *in vacuo* must be constant, for example. But by far, the preponderance of information we have at our disposal is the result of direct observation and measurement of physical systems. These observations can result in subjective information–expertise–or quantitative information–data, sometimes with no clear distinction between the two. We believe that these two forms of information should enter inference calculations in a unified fashion.

Thus the motivation for quantifying states of information is clear. We need to be able to assess the amount of information contained in different premises, which premises are usually expressed as functionals on the space of models and data. The statement that, for instance, the P-wave velocity V_p of a certain piece of the earth is positive is less informative than the statement that the V_p velocity lies within some finite range $[0 < x \leq V_p \leq y < \infty]$. The first is absolutely true but conveys little information, while the second is rather more informative, but possibly wrong. When we attempt to quantify the state of information that results from an inference or inverse calculation, we are, in effect, measuring the *resolution* of the calculation. This quantification takes the form of probabilistic statements about hypotheses, expressing a degree of belief in the their truth. In the above example, we may wish to know the smallest range of P-wave velocity that contains the true value with a given degree of confidence.

The modern theory of information can be traced to the pioneering work of Shannon who defined the information of a communication channel in terms of the probabilities associated with the different symbols being transmitted. Actually Shannon dealt not with information directly, but with entropy, the negative of information: the more information, the less entropy, and vise versa. Unfortunately, in attempting to understand these ideas and, perhaps, make use of them in our work on inversion, we quickly discovered a number of troubling technical issues. E.g., most of the textbooks define entropy (or information) in a way that is unacceptable for our purposes since the definition fails to satisfy a basic invariance property. (It cannot matter what coordinate system we use to make our inferences.) Further, it appears that in using a limiting argument to pass from a discrete to a continuous measure of information, certain infinities crop up that are swept under the rug. As we shall see, fixing the first problem also fixes the second. Additionally, in so doing we achieve a measure of information that is finite for finite (and normalizeable) states of information. We will also show how reasonable results can be achieved in the case of infinite states of information by a suitable limiting procedure.

The price we pay for this mathematical consistency is the replacement of the notion of information (or entropy) with one of *relative* information: we cannot speak of the content of a given state of information; we can only speak of the information of one state of information relative to another. But this is, after all, what we need for inverse calculations since we always begin with some prior state of information and attempt to refine it.

This point of view is hardly new. Here is Kolmogorov writing 40 years ago:

> Furthermore, I insist that the fundamental concept, which admits of generalization to perfectly arbitrary continuous information and signals, is not directly the entropy concept but the concept of the quantity of information $I(\eta\,\xi)$ in the random object ξ relative to the object η... It is well known that the quantity $h(\xi)$ [continuous entropy] has no direct real interpretation and is not even invariant with respect to coordinate transformations in the space of the x's. For an infinitely-dimensional distribution, the analog of $h(\xi)$ is nonexistent, in general (Kolmogorov, 1956).

And yet this principle of relative information does not seem to be well understood in practice, Tarantola's book (1987), being a notable exception.

Our original motivation for looking into entropy was the ability to use the principle of maximum entropy to calculate conservative Bayesian priors from observations. For example, if we take *in-situ* measurements of some property and attempt to use those measurements to build a Bayesian *a priori* distribution, then we are faced with the task of constructing the joint n-dimensional distribution of a process of unknown complexity. If we are willing to assume that the process is Gaussian, then computing the prior reduces to the classical problem of estimating the mean and covariance from sampled data. But if we wish to get beyond the Gaussian assumption, then we must somehow take into account higher order information. In principle, one can estimate some number of higher order marginal distributions via histograms, but without some underlying model of the process, we need an external criterion to decide when to stop this procedure. In Scales & Tarantola (1994) we considered using statistical hypothesis testing to answer this question, but that approach is not without limitations too, and could be prohibitively expensive beyond second or third order

marginals without additional approximations. Thus we were lead to consider whether a properly conservative estimation of prior distributions could be achieved via the principle of maximum entropy. In other words, given sampled data, estimate some number of sample moments of the unknown distribution and then calculate a prior distribution that is least informative subject to these sample moments as constraints. Once we have the prior distribution we can proceed to solve the full Bayesian inference calculation. In theory this estimation of the prior distribution is a straightforward application of the maximum entropy method. Indeed we show examples of this approach applied to one-dimensional random processes. However, the challenges in higher dimensions are formidable.

But we are getting ahead of ourselves. To lay a proper foundation for the study of entropy-based methods we begin at the beginning. In the first section we discuss the *other* entropy, i.e., that used in thermodynamics and statistical mechanics. By considering a simple example of an irreversible thermodynamic process we will see that even the classical entropy can be thought of in terms of the uncertainty of the microscopic state of a system. This connection between the macroscopic (thermodynamic) and the microscopic (statistical mechanics) view of a system allows one to derive an expression for the entropy in terms of the probability distribution for the coordinates and momenta of a microscopic state of the system in phase space Γ. It was Shannon's great insight that this quantitative measure of uncertainty could be applied generally to problems involving information. In the second section we discuss Shannon's definition and show by a number of simple examples that it does indeed capture our qualitative understanding as a measure of information. This alone may seem hardly compelling. However, it can be shown that by asking what properties *any* reasonable measure of information should have, we will be drawn inexorably to Shannon's definition, or something very like it. The hedge in the last sentence has to do with the fact that, as we have already mentioned, Shannon's definition of entropy fails to be invariant with respect to coordinate transformations. The remedy for this is mathematically quite straightforward, but the interpretation requires that we alter fundamentally our interpretation of the procedures by which we obtain and measure states of information. Using this revised definition of information clears up a number of related mathematical difficulties and allows a single theoretical framework for continuous and discrete states of information.

In an appendix we give a review of these claims and consider several different derivations of the basic equations. In addition, we show a number of worked examples of how to do these calculations. For example, it makes no sense to speak of the entropy of a distribution (even a normalizable one) relative to a constant of information on an unbounded domain since no such reference state of information is normalizable. Fortunately we get the *right* answer via a limiting procedure, treating only normalizable distributions at each stage.

With this theoretical foundation, we are then in a position to discuss the principle of maximum entropy. Since entropy is a measure of uncertainty, maximizing the entropy, subject to certain constraints determined by our information, represents a fundamental statement of conservatism: in the presence of uncertainty we should make the least informative inferences (i.e., the safest ones) consistent with our information. So the maximum entropy calculation is actually a constrained optimization calculation for an unknown probability distribution. In principle this can be solved by the method of Lagrange multipliers. And for one-dimensional problems this is straightforward. The situation is much less clear in higher dimensions. We will conclude with some recent work we have been doing on the numerical extension of these ideas to $N-$dimensional calculations and show how these techniques might be used to incorporate complicated prior information in the Bayesian inverse calculations.

Entropy in Thermodynamics and Statistical Mechanics

The concept of entropy in thermodynamics is intimately related to the question of reversibility. The second law of thermodynamics says that a process whose only net result is to take heat from a reservoir and convert it to work is impossible. Carnot used this principle to show that the amount of work obtained for any two reversible engines extracting an amount of heat Q_1 from a reservoir at temperature T_1 and delivering it to a reservoir at temperature T_2 must be the same *independent of the design of the engine.*

Carnot considered an idealized kind of heat engine which operates using a gas-filled piston to extract heat Q_1 from a reservoir at temperature T_1 during an isothermal expansion. The gas is then allowed to expand further but now isolated from the heat reservoir so that the temperature drops to T_2 with no heat being transferred. Next the gas is put in contact with a heat reservoir of temperature T_2 and compressed isothermally delivering a heat Q_2 to the reservoir. Finally the gas is removed from contact with the T_2 reservoir and compressed adiabatically (no heat transfer) raising the temperature from T_2 to T_1. If the piston is frictionless and all the operations are carried on very slowly, then this idealized engine is

reversible. The gas ends up in exactly the state it began with some useful work having been delivered.

Since Carnot's results are independent of the type of reversible engine used, it is possible to study a particular kind of engine and derive universal laws. For Carnot's engine using an ideal gas (i.e., a gas whose equation of state is $PV = NkT$, where P is the pressure, V is the volume, N is the number of particles, k is Boltzmann's constant, and T is the temperature) it is possible to evaluate the expressions for Q_1 and Q_2 in terms of T_1, T_2 and the volumes of the piston at various stages of the engine's cycle. It then follows that

$$\frac{Q_1}{T_1} = \frac{Q_2}{T_2}. \tag{1}$$

The derivation can be found in most thermodynamics books, but an especially clear one is given by Feynman (Feynman *et al.* (1963), Volume I, section 44). The universality of Carnot's result means that although Equation (1) has been proved for an ideal gas, it must be true *for any reversible engine.* (Feynman calls this purely logical deduction of a universal principle one of the most beautiful pieces of reasoning in physics.) Since there can be no net change in $\frac{Q}{T}$ during any reversible cycle, this ratio is special enough to warrant a new name: *entropy.* So in a reversible cycle the total entropy of the system does not change. In an irreversible cycle the total entropy of the system always increases.

Let us consider a specific example of an irreversible transformation. Suppose we have a box divided in half by a partition with a quantify of gas in one half of the box. We remove the partition and let the gas expand to fill the whole box while keeping the temperature constant. Clearly this is an irreversible change since the gas molecules will never arrange themselves so as to spontaneously occupy only half of the box. The entropy change in this transformation is

$$\Delta S = \int \frac{dQ}{T}. \tag{2}$$

Since the temperature does not change we must add a little heat dQ to equal the work done by the gas in the expansion PdV. So

$$\Delta S = \int_{V_1}^{V_2} P\frac{dV}{T} \tag{3}$$

which, for an ideal gas, equals

$$\int_{V_1}^{V_2} \frac{NkT}{V}\frac{dV}{T}. \tag{4}$$

Thus,

$$\Delta S = Nk \ln \frac{V_2}{V_1}. \tag{5}$$

The entropy has increased by an amount $Nk \ln \frac{V_2}{V_1}$ even though the temperature and energy of the system are the same. All we have done is given the molecules more room to move around. Or, anticipating a bit, we have increased the number of possible states of the system; that is, the number of ways in which the molecules could be arranged. The entropy is the logarithm of this number of states.

Now that we have slipped into thinking about our system microscopically, it is useful to imagine the state of any system of N particles as being specified by a point in a $6N$-dimensional phase space Γ comprising the values of all $3N$ coordinates $\mathbf{q}$ and $3N$ momenta $\mathbf{p}$. The time evolution of the system is then described by a trajectory $\{\mathbf{q}(t), \mathbf{p}(t)\}$ in phase space which is uniquely determined by $6N$ initial values of the coordinates and momenta $\{\mathbf{q}(t_0), \mathbf{p}(t_0)\}$. So since it is clear from the gas expansion example that the density of states available to a system is closely related to the classical entropy, it will be useful to define a probability density ρ on the phase space telling us the probability that our system will be found in any element $d\mathbf{q}d\mathbf{p}$ by

$$\mathrm{P}\left[(\mathbf{q}, \mathbf{p}) \in d\mathbf{q}d\mathbf{p}\right] \equiv \rho(\mathbf{q}, \mathbf{p})\ d\mathbf{q}d\mathbf{p} \tag{6}$$

The precise form of the *ensemble* function ρ depends on the the type of problem being treated. If the system is isolated, then the Hamiltonian H is constant and any state on the surface $H = E$ is equally likely. Thus this *micro-canonical* ensemble must be described by a delta function: $\rho = \delta(H(\mathbf{q}, \mathbf{p}) - E)$. However, for purposes of Carnot's theory, we need the ensemble associated with a system at constant constant temperature, not constant energy. This is called the *canonical* ensemble and the expression for ρ in this case turns out to be the Boltzmann distribution (Weiner, 1983)

$$\rho(\mathbf{q}, \mathbf{p}) = \exp(-H(\mathbf{q}, \mathbf{p})/kT).$$

The key idea is this: Our expression above for the change of entropy (Equation (2)) involves the change in heat or energy at constant temperature. In the canonical ensemble this would just be the time rate of change of the expected value of the Hamiltonian. The details can be found in (Weiner, 1983). The upshot is that by making this connection between the macroscopic thermodynamic picture and the microscopic statistical one, it is possible to deduce the following expression for the entropy of a system in the canonical ensemble:

$$S = -k \int_{\Gamma} \rho(\mathbf{q}, \mathbf{p}) \ln \rho(\mathbf{q}, \mathbf{p})\ d\mathbf{q}d\mathbf{p} \tag{7}$$

Anticipating further our journey into information theory, we can imagine the thought experiment of observing our system at any instant of time. The possible outcomes of this experiment are the points in the phase

space accessible to the system. And so the entropy of the system must be a measure of the uncertainty associated with this experiment. If the trajectory of the system is confined to a relatively small volume of Γ it will have a relatively small entropy. If the trajectory wanders widely in Γ it will have a correspondingly greater entropy.

Classical Measures of Information

Consider an experiment with N possible outcomes each occuring with a probability p_i. In analogy with the statistical mechanical definition of entropy just discussed, (Shannon, 1948) introduced the following definition of the entropy for such discrete probabilities:

$$H(p) = -\sum_i p_i \log p_i. \tag{8}$$

Following Shannon, three postulates should be satisfied by $H(p)$ or any other measure of information. Those are:

(i) Continuity;
(ii) Monotonicity, and
(iii) Composition Law.

We elaborate on these postulates in Appendix A, for now a qualitative understanding is sufficient. The first postulate requires that we should not gain a large amount of information by making a small change to the probabilities. The second postulate, monotonicity, refers to the information associated with a collection of independent, equally likely events. It is clear that in such a case the uncertainty must increase monotonically with the number of possible outcomes. The third postulate requires that it should not matter how one regroups the events of a given set. The entropy of the set should stay the same.

Later we will have to modify the definition of H given in Equation 8 in an important way, but for now let us just consider its basic meaning. If one of the outcomes of the experiment is absolutely certain, then we represent the probability density by a Kronecker delta: $p_i = \delta_{iq}$ where q is the certain event. In this case there is no uncertainty and $H(p) = 0$. If there are two equally likely outcomes then $H(p) = -1/2\log(1/2) - 1/2\log(1/2) = \log 2$. Whereas if one of the events has a probability $1/10$ and one has probability $9/10$ then the entropy is $H(p) = -1/10\log(1/10) - 9/10\log(9/10) = \log 10 - .9\log 9$, which is about half that in the equally likely case. For M equally likely events $H(p) = \log M$. And in the limit that M goes to infinity, then the uncertainty must too.

The usefulness of the definition 8 depends on the definition of the probability p. If the p is a 1D distribution associated with the frequency of outcomes of the possible events, then p is not affected by the correlation of the points. E.g., if we sample 10 points pseudo-randomly from a probability distribution with two equally likely outcomes 0 and 1, we might see something like the following

0010110110.

As luck would have it there are 5 0's and 5 1's. Now sort these outcomes in increasing order

0000011111.

There are still 5 0's and 5 1's but we certainly would not regard the latter experiment as representing the same degree of uncertainty as the former. Similarly, if we sample 1000 points independently from a Gaussian we'll see a nice bell-shaped curve. But the frequencies of the binned events is independent of their order. So, once again, sorting them into monotonic order will not change the entropy. Later we will see how it is possible to include the dependency of events in the definition of p and achieve a completely general definition of entropy.

Extension to Continuous Probabilities.

The extension of Equation (8) definition to *continuous* probabilities is necessary since in many scientific problems we deal with continuous variables. Several approaches can be found in the literature (Jaynes, 1963, Rietsch (1977)), and others). Here we describe the one given by Rietsch. Assume that x is a continuous variable that lies in some interval $[a, b]$. Let $p(x)dx$ denote the probability that a value of x be in the interval $[x, x+\Delta x]$. A "natural" way to define the entropy of x would be to subdivide the interval $[a, b]$ into n sub-intervals $[x_{i-1}, x_i]$ of length Δx_i; $p(x_i)\Delta x_i$ is the probability that x is in this interval. Therefore, the entropy of this discretized probability distribution reads

$$\begin{aligned} H[p(x)] = & -\sum_i p(x_i)\Delta x_i \log[p(x_i)\Delta x_i] = \\ & -\sum_i p(x_i)\Delta x_i \log p(x_i) \\ & -\sum_i p(x_i)\Delta x_i \log \Delta x_i. \end{aligned} \tag{9}$$

Introducing the variable $q(x_i)$ and the constant δ such that

$$\Delta x_i = \frac{\delta}{q(x_i)}, \ \sum_i q(x_i)\Delta x_i = n\delta = 1, \tag{10}$$

we have for Equation (9):

$$H[p(x)] = -\sum_i p(x_i)\Delta x_i \log p(x_i)$$

$$
\begin{aligned}
& -\sum_i p(x_i)\Delta x_i \log \frac{\delta}{q(x_i)} \\
= & -\sum_i p(x_i)\Delta x_i \log p(x_i) \\
& -\sum_i p(x_i)\Delta x_i \log \delta \\
& +\sum_i p(x_i)\Delta x_i \log q(x_i) \\
= & -\sum_i p(x_i)\Delta x_i \log \frac{p(x_i)}{q(x_i)} + \log n, \qquad (11)
\end{aligned}
$$

in which we used the fact that $\sum_i p(x_i)\Delta x_i = 1$. Notice that as $n \to \infty$, Equation (11) diverges because of $\log n$. Nonetheless this term is neglected in Rietsch's derivation, and in the limit, the final result reads:

$$H[p(x); q(x)] = -\int_a^b p(x) \log \frac{p(x)}{q(x)} dx. \qquad (12)$$

The above expression defines what is known as relative entropy. $q(x)$ is Jayne's *invariant measure*. It is not complicated to show that $q(x)$ is is sufficient to make the function $H[p(x); q(x)]$ invariant to coordinate transformations. To do so consider changing coordinates from x to y in Equation (12). This yields:

$$
\begin{aligned}
H[p(x); q(x)] &= -\int_a^b p(y)\frac{dy}{dx} \log \frac{p(y)\frac{dy}{dx}}{q(y)\frac{dy}{dx}} dx \\
&= -\int_{a'}^{b'} p(y) \log \frac{p(y)}{q(y)} dy, \qquad (13)
\end{aligned}
$$

where a' and b' are the corresponding limits of integration in the y coordinates. Equations (12) and (13) are of course identical. As it will be shown later, the quantity $q(x)$ can also be interpreted as a reference (or prior) state of information. In this view, $H[p(x); q(x)]$ quantifies the relative entropy of the probability function $p(x)$ with respect to $q(x)$.

However, we are not entirely comfortable with Rietsch's derivation. Other approaches to the extension from discrete to continuous entropy resort to similar arguments; and while the renormalization of infinite integrals by subtraction is not unheard of, we were motivated to see whether this was really necessary. Let us therefore consider the following definition for discrete entropy:

$$H[p_i; q_i] = -\sum_i p_i \log \frac{p_i}{q_i}. \qquad (14)$$

Here, q_i is a discrete probability characterizing a reference state of information. The extension of Equation (14) to the continuous case is clean and straightforward:

$$
\begin{aligned}
H[p(x); q(x)] &= -\sum_i p(x_i)\Delta x_i \log \frac{p(x_i)\Delta x_i}{q(x_i)\Delta x_i} \\
&= -\sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)} \Delta x_i \\
&= -\int_a^b p(x) \log \frac{p(x)}{q(x)} dx, \qquad (15) \\
&\quad \text{as } \Delta x_i \to 0 \text{ (or } n \to \infty)
\end{aligned}
$$

which is finite.

From here on we will adopt Equation (14) as the definition of relative entropy in the discrete case, and, as commonly done, the last expression of Equation (16) as the definition of relative entropy in the continuous case. $q(x)$, or q_i, represents a state of information against which we make comparisons. Finally it is worth mentioning that the negative of the quantity $H[p(x); q(x)]$, known as cross-entropy, was first defined by Kullback (1959) as the *directed divergence*. This quantity defines the amount of *information* of the probability density $p(x)$ with respect to $q(x)$. See also (Shore & Johnson, 1981).

Reference Priors

Let us look more closely at the reference state of information. Some argue that $q(x)$ represents a fundamental state of knowledge involving the variable x. Tarantola (1987) calls it *null state of information*, which implies that $q(x)$ must be completely *non-informative*. Jaynes (1968) refers to such distributions as *ignorance* priors.

The issue of assigning non-informative prior distributions goes back to the time of Bayes and Laplace. They both used uniform priors, but even then it was clearly unsatisfactory since that assumption leads to contradictions upon a change of variables. Specifically, if one considers the uniform distribution for x as the fundamental state of knowledge and changes the variables to a different system of coordinates y, the reference prior will not be the same. But if we begin with a uniform prior for y, it will yield a different reference prior for x, which makes no sense. Because of that, Jaynes (1968) proposed the use of group theory to define invariant priors to represent our ignorance about a given parameter.

This can be illustrated with a simple example from Jaynes (1968). If we are completely ignorant about a location and scale parameters (μ and σ, e.g., the mean and standard deviation), a change of variables

$$\mu^* = \mu + b \quad \text{and} \qquad (16)$$

$$\sigma^* = a\,\sigma \qquad (17)$$

takes our problem into a completely equivalent one. That is, a change in variable does not contain any information about the unknown quantities. By the usual transformation equation, any distribution $f(\mu, \sigma)$ transforms as

$$g(\mu^*, \sigma^*) = a^{-1} f(\mu, \sigma), \qquad (18)$$

where a^{-1} is the Jacobian of the transformation. So, if we require that the prior distribution be invariant with respect to a change of coordinates, then we must have

$$g = f. \tag{19}$$

By substituting this into (18), we get the functional equation

$$f(\mu, \sigma) = a\, f(\mu + b, a\sigma). \tag{20}$$

The solution for this is Jeffrey's prior, which is given by

$$f(\mu, \sigma) = \frac{c}{\sigma},$$

where c is an arbitrary constant.

Thus, we can outline a general procedure for determining the invariant measure as given by

(i) Find the relevant group of transformations for the problem. In the above example, it is given by all transformations if the form of Equations (16) and (17) with $0 < a < \infty$ and $-\infty < b < \infty$.

(ii) By the combination of Equations type (18) and (19) get the functional equation representing the invariance.

(iii) Solve the functional equation to find the invariant prior.

Usually, physical reasoning is used for defining an appropriate group of transformations relevant to the problem at hand and leading to a functional equation whose solution exists.

Example 19 in Chapter 1 of (Tarantola, 1987) provides an illustration of this sort of reasoning. This example describes the problem of assigning a non-informative distribution for the velocity of a non-relativistic particle. Then, if $\mathbf{r}$ is the Cartesian vector describing the particle's position at time t, the magnitude of the particle's velocity is given by

$$v = \left| \frac{d\mathbf{r}}{dt} \right| .$$

The same quantity can be represented through the Galilean transformation

$$\begin{aligned} \mathbf{r}' &= \mathbf{r}_0 + a\,\mathbf{r}\ , \\ t' &= t_0 + b\,t\ . \end{aligned}$$

Using the postulate of space-time homogeneity we have the invariance of mathematical form. Thus, the particle's velocity in the new coordinate system is

$$v' = \left| \frac{d\mathbf{r}'}{dt'} \right| ,$$

which leads to $v' = \alpha\, v$, where $\alpha = \left| \frac{a}{b} \right|$. To avoid the contradiction described in the previous section, a non-informative distribution for both systems also should be the same, which is expressed by Equation (19). Consequently, we find a functional equation similar to Equation (20), which is given by $f(v) = \alpha\, f(\alpha\, v)$, the solution of which is $f(v) = \frac{c}{v}$.

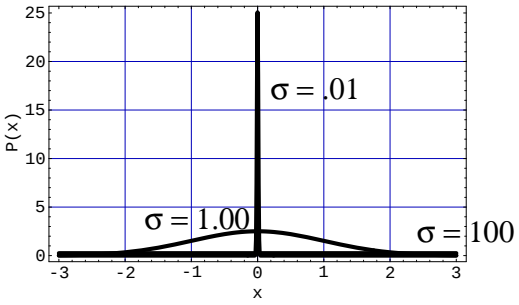


Figure 1. A 1D Gaussian distribution for three different values of the standard deviation. If σ is small, then the distribution is relatively sharp–informative. If σ is large, then the distribution is relatively broad–noninformative.

Some Examples of Entropy Calculations

In this section we digress briefly to show an analytic calculation of relative entropy. We shall calculate the relative entropy $H[p(\mathbf{x}); q(\mathbf{x})]$ of one uncorrelated Gaussian with respect to another. By examining the limiting cases of zero and infinite variance, we can derive the relative entropy associated with states of perfect knowledge and total uncertainty. Let p and q be two Gaussian distributions defined by

$$p(\mathbf{x}) = \sqrt{\pi^n \sigma_q^n} \exp\left(-\frac{1}{2} \frac{\|\mathbf{x}\|^2}{\sigma_p^2} \right) \tag{21}$$

$$q(\mathbf{x}) = \sqrt{\pi^n \sigma_p^n} \exp\left(-\frac{1}{2} \frac{\|\mathbf{x}\|^2}{\sigma_q^2} \right) , \tag{22}$$

where n is the dimension of the vector variable $\mathbf{x}$, and σ_p and σ_p are the respective standard deviations. In the limit that σ goes to zero, these distributions converge to delta functions, corresponding to states of perfect knowledge. In the limit that σ goes to infinity, they converge to (zero) uniform probability distributions, corresponding to states of complete uncertainty. This is illustrated in Figure (1).

Given the definition of relative entropy it is straightforward to show that:

$$\begin{aligned} &H[p(\mathbf{x}); q(\mathbf{x})] = \\ &= -\int_V p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} + \int_V p(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{x} \\ &= -\int_V p(\mathbf{x}) \left(\log \frac{1}{2\pi^{\frac{n}{2}} \sigma_p^n} - \frac{1}{2} \frac{\|\mathbf{x}\|^2}{\sigma_p^2} \right) d\mathbf{x} \end{aligned}$$

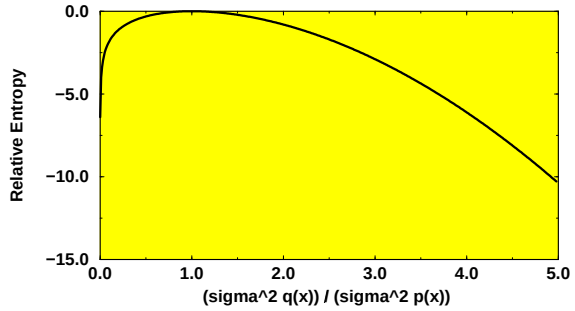


Figure 2. Relative entropy of one uncorrelated Gaussian $p(\mathbf{x})$ to another $q(\mathbf{x})$ (Equation (23)) as a function of the ratio of the variances.

$$
\begin{aligned}
&+\int_V p(\mathbf{x})\left(\log\frac{1}{2\pi^{\frac{n}{2}}\sigma_q^n}-\frac{1}{2}\frac{\|\mathbf{x}\|^2}{\sigma_q^2}\right)d\mathbf{x}\\
&= -n\log\frac{\sigma_q}{\sigma_p}+\frac{1}{2}\left(\frac{\sigma_q^2-\sigma_p^2}{\sigma_p^2\sigma_q^2}\right)\langle\|\mathbf{x}\|^2\rangle\\
&= -n\left[\log\frac{\sigma_q}{\sigma_p}-\frac{1}{2}\left(1-\frac{\sigma_p^2}{\sigma_q^2}\right)\right]. \qquad (23)
\end{aligned}
$$

Here we have used the fact that:

$$\langle\|\mathbf{x}\|^2\rangle=\int_V\|\mathbf{x}\|^2\,p(\mathbf{x})\,d\mathbf{x}=n\sigma_p^2. \qquad (24)$$

Equation 23 is the entropy of one uncorrelated Gaussian relative to another. Figure 2 shows a plot of this as a function of the ratio of the variances. We see immediately that the relative entropy is always negative (information always positive). When the ratio of the variances is one, then both distributions are equally informative and the relative entropy is 0. In other words, the state of information that we are measuring, p, affords no new information compared to the reference state q.

The relative entropy decreases (information increases) monotonically with increasing σ_q/σ_p. This is achieved by letting σ_p be small compared to σ_q, corresponding to knowing p with increasing relative certainty. In the limit that σ_p goes to zero, p becomes a Dirac delta function and for any finite σ_q the relative information is infinite.

On the other hand, the relative entropy also decreases monotonically (information increases) when σ_q/σ_p decreases below 1. This is a somewhat unexpected result and would correspond to the situation in which the reference state of information is more informative than the one we are trying to measure. It is not inconceivable that we could encounter an inverse calculation in which the reference state of information is wrong or incompatible with observations, in which case the posterior probability could be less informative than the prior.

The situation can be summarized as follows. The relative information of one Gaussian to another is always positive (relative entropy always negative) and that this information is a minimum when the two distributions are the same, and a maximum when the measured state p is infinitely precise, for a fixed reference state. When this state is more informative than the one defined by $p(\mathbf{x})$, the relative entropy provides inconsistent results, and an alternative information measure should be used in this case.

Principle of maximum entropy

Convinced that entropy is a suitable measure for the uncertainty of a probability distribution, Jaynes (1957) showed that a useful tool for conservatively assigning probabilities was to maximize the entropy of the unknown distribution subject to constraints on its moments.

Mathematically this variational problem can be expressed by maximizing Equation (14) subjected to the normalization of the distribution

$$\sum_{i=1}^{N}p(x_i)=1, \qquad (25)$$

and to other constraints given in the form of expectations $\langle w_k(x)\rangle$

$$\sum_{i=1}^{N}w_k(x_i)p(x_i)=\mu_k\ , \qquad k=1,...,K, \qquad (26)$$

where μ_n is a numerical value computed from the available data.

This is equivalent to the unconstrained problem, given by

$$
\begin{aligned}
S(p;\lambda,q) \;=\; & -\sum_{i=1}^{N}p\,(x_i)\ln\frac{p\,(x_i)}{q\,(x_i)}\\
& -(\lambda_0-1)\left[\sum_{i=1}^{N}p\,(x_i)-1\right]\\
& -\sum_{k=1}^{K}\lambda_k\left[\sum_{i=1}^{N}w_k(x_i)\,p\,(x_i)-\mu_k\right], \qquad (27)
\end{aligned}
$$

where the λ are the Lagrange multipliers associated with the constraints. Note that the term (λ_0-1) is just a redefinition of the zero-order Lagrange multiplier introduced for convenience. If we take the first variation of the functional $S(p;\lambda,q)$ with respect to the probabilities, we get that $\delta S(p;\lambda,q)$ equals

$$\sum_{i=1}^{N}\left[\frac{\partial H}{\partial p(x_i)}-(\lambda_0-1)-\sum_{k=1}^{K}\lambda_k\,w_k(x_i)\right]\delta p(x_i), \qquad (28)$$

with

$$\frac{\partial H}{\partial p(x_i)}=-\left[\ln\frac{p\,(x_i)}{q\,(x_i)}+1\right]. \qquad (29)$$

The solution to the problem can be found in the usual way by letting $\delta S(p) = 0$, which yields

$$p(x_i) = q(x_i) \exp\left[-\lambda_0 - \sum_{k=1}^{K} \lambda_k \, w_k(x_i)\right], \tag{30}$$

or

$$p(x_i) = Z^{-1} q(x_i) \exp\left[-\sum_{k=1}^{K} \lambda_k w_k(x_i)\right], \tag{31}$$

with

$$Z \equiv \exp(\lambda_0) = \sum_{i=1}^{N} q(x_i) \exp\left[-\sum_{k=1}^{K} \lambda_k \, w_k(x_i)\right]. \tag{32}$$

However, for a complete solution we still need to find the values for the other Lagrange multipliers and to specify the prior probability q.

Maximization of the continuous entropy functional is essentially the same as in the discrete case presented before. As usual we require the distribution be normalized

$$\int_R p(x)\, dx = 1. \tag{33}$$

And there are other constraints in the form of expectations $\langle w_k(x) \rangle$

$$\int_R w_k(x)\, p(x)\, dx = \mu_k, \qquad k = 1, ..., K\ , \tag{34}$$

where μ_n is a numerical value known from the available data. The solution for this problem obtained in the same way as before is given by

$$p(x) = q(x) \exp\left[-\lambda_0 - \sum_{k=1}^{K} \lambda_k \, w_k(x)\right]\ , \tag{35}$$

or

$$p(x) = Z^{-1} q(x) \exp\left[-\sum_{k=1}^{K} \lambda_k w_k(x)\right]\ , \tag{36}$$

with

$$Z \equiv \exp(\lambda_0) = \int_R q(x) \exp\left[-\sum_{k=1}^{K} \lambda_k \, w_k(x)\right] dx\ . \tag{37}$$

Like the discrete case, the complete solution requires the determination of the Lagrange multipliers and the reference prior $q(x)$, which we investigate next.

Computation of the Lagrange Multipliers

The specification of the maximum-entropy probability requires the determination of the Lagrange multipliers. In this section we present two approaches for such computation. The first one relies on optimization techniques for the determination of these parameters, and it is fully described by Mead and Papanicolau (1984). The second is an extension of the technique proposed in Jumarie (1990) for the multidimensional case.

The Optimization Approach of Mead and Papanicolau

Consider the one-dimensional case. As discussed before, the maximum-entropy distribution when the moments $\langle x^i \rangle = \mu_i, i = 1, ..., n$ and the prior distribution $q(x)$ are known, is given by:

$$p(x) = Z^{-1} q(x) \exp\left(-\sum_{i=1}^{n} \lambda_i x_i\right), \tag{38}$$

where

$$Z = \int q(x) \exp\left(-\sum_{i=1}^{n} \lambda_i x_i\right). \tag{39}$$

Mead and Papanicolau (1984) defined the following potential function:

$$\Gamma(\lambda_1, \lambda_2, ..., \lambda_n) = \log Z + \sum_{i=1}^{n} \lambda_i \mu_i. \tag{40}$$

The desired set of Lagrange multipliers are the stationary points of the potential Γ, being the solution of the linear system of equations:

$$\frac{\partial \Gamma}{\partial \lambda_i} = 0 \rightarrow \langle x_i \rangle = \mu_i,\ i = 1, ..., n. \tag{41}$$

Therefore, the computation of the Lagrange multipliers can be formulated as a optimization problem, which can be solved by Newton's method. If we denote the vector of the Lagrange multipliers by $\mathbf{l}$ and the gradient of Γ by $\mathbf{r}$, we can write the iteration equation for Newton's method as

$$\mathbf{l}^{(n+1)} = \mathbf{l}^{(n)} - \mathbf{H}^{-1}\mathbf{r}\ .$$

From Equation (41), each component of $\mathbf{r}$ is given by

$$r_i = \mu_i - \left\langle m^i \right\rangle\ ,$$

which is the residual between the input sample moment and the corresponding expected value over the estimated pdf at the n-th iteration.

An Alternative Approach

In the multidimensional case the maximum-entropy distribution is given by:

$$p(\mathbf{x}) = q(\mathbf{x}) \exp\left[-\lambda_0 - \sum_{i=1}^{n} \lambda_i w_i(\mathbf{x})\right], \tag{42}$$

once the prior distribution $q(\mathbf{x})$ and $\langle w_i(\mathbf{x})\rangle, i = 1, ..., n$ are available. $w(\mathbf{x})$ represents the expression for the moment used in the maximum-entropy computation. For example, if we are dealing with the $m-th$ order moment about zero, $w(\mathbf{x})$ is given by:

$$w_m(\mathbf{x}) = \prod_{i=1}^{m} x_i \tag{43}$$

The algorithm proposed here is a multidimensional version of the procedure described in Jumarie (1990), which is based solely on the integration by parts of the following moment integrals:

$$\langle w_i(\mathbf{x})\rangle = \int w_i(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}, \quad i = 1, ..., n. \tag{44}$$

To simplify the description of the procedure, consider that the problem at hand is to find the maximum-entropy distribution when $q(\mathbf{x})$ and moments up to the second-order are known, and given by:

$$\begin{aligned} \langle x_i\rangle &= \int_V x_i\, p(\mathbf{x})\, d\mathbf{x}, \quad w(\mathbf{x}) = x_i, \\ \langle x_i x_j\rangle &= \int_V x_i x_j\, p(\mathbf{x})\, d\mathbf{x}, \quad w(\mathbf{x}) = x_i x_j, \end{aligned} \tag{45}$$

valid for all $i, j = 1, \ldots n$. Cumbersome but straightforward algebra shows that if we perform an integration by parts in the first expression of Equation (45) in the variable x_{i+1} we obtain:

$$\begin{aligned} \langle x_i\rangle &= \lambda_{i+1}\langle x_i x_{i+1}\rangle \sum_{j=1}^{n} \lambda_{i+1\, j}\langle x_i x_{i+1} x_j\rangle \\ &+ \lambda_{i+1\, i+1}\langle x_i x_{i+1}^2\rangle, \quad i = 1, ..., n-1. \end{aligned} \tag{46}$$

If we do the same with the second expression of Equation (45), now however with the integration by parts being on the variable x_k, with $k \neq i$ and $k \neq j$, we obtain:

$$\begin{aligned} \langle x_i x_j\rangle &= \lambda_k\langle x_i x_j x_k\rangle + \sum_{l=1}^{n} \lambda_{lk}\langle x_i x_j x_k x_l\rangle \\ &+ \lambda_{kk}\langle x_i x_j x_k^2\rangle, \quad i = 1, ..., n. \end{aligned} \tag{47}$$

Equations (46) and (47) form a linear system of equations that in principle could be solved for the Lagrange multipliers. Notice that no numerical integration is required in this procedure as it is in the method previously described. The price we pay is the need for computation of higher moments. In this case, a second-order multidimensional moment problem, we have to compute up to the fourth-order moments. These moments should be obtained from the data available to the maximum-entropy problem. This is still an open topic, but our conclusion up to this point is that higher-order moments (higher than 2) are sensitive to the number of data samples (population size); therefore accurate estimates are difficult to obtain. Once one overcomes this complication, the algorithm formulated here can be used for the construction of the maximum-entropy multidimensional probability function.

Numerical Calculation of 1D Priors Via Maximum Entropy

To illustrate the algorithm of Mead and Papanicolau, we will describe several calculations involving different distributions. In these examples, various probability densities are used to generate data sets with one thousand samples each (Figure 3). Then, sample moments up to the fourth-order (Figure 4) are computed and input into a routine that computes the Lagrange multipliers. The resulting maximum entropy probability densities are shown in Figure 5. The reference prior for all examples is uniform on $[x_l, x_u]$, where x_l and x_u are respectively the minimum and maximum values computed from each set of samples. We stopped the iteration of the algorithm when the moments of the maximum entropy distribution agreed with the sample moments to 10^{-6} or better. In the computations performed, it took an average of six to seven iterations for the convergence of the algorithm. An alternative model was computed for comparison, using only the first two moments. This corresponds to truncated normal distributions, shown in Figure 6. When Figures 5 and 6 are compared, they show how the inclusion of the third and fourth sample moments into the computations improves the fit between the true and the maximum entropy distributions.

Conclusions

Maximizing the entropy of an unknown function subject to certain constraints is a powerful technique for finding the most featureless function consistent with those constraints. When applied to inference, this principle allows one to calculate the most conservative probabilistic description of prior information consistent with calculable features of that information such as sample moments. In order to make this procedure satisfy basic consistency requirements, it is necessary to abandon the notion of absolute states of information (or their negative, entropy), and instead think in terms of the information of one state relative to another. To some, the latter (reference) state of information has deep significance derived from physical invariance properties; while to others, this reference state is simply what we know about a problem before we have done an experiment to acquire new information.

We have set forth the basic principles of entropy and maximum entropy in a general way and shown that any

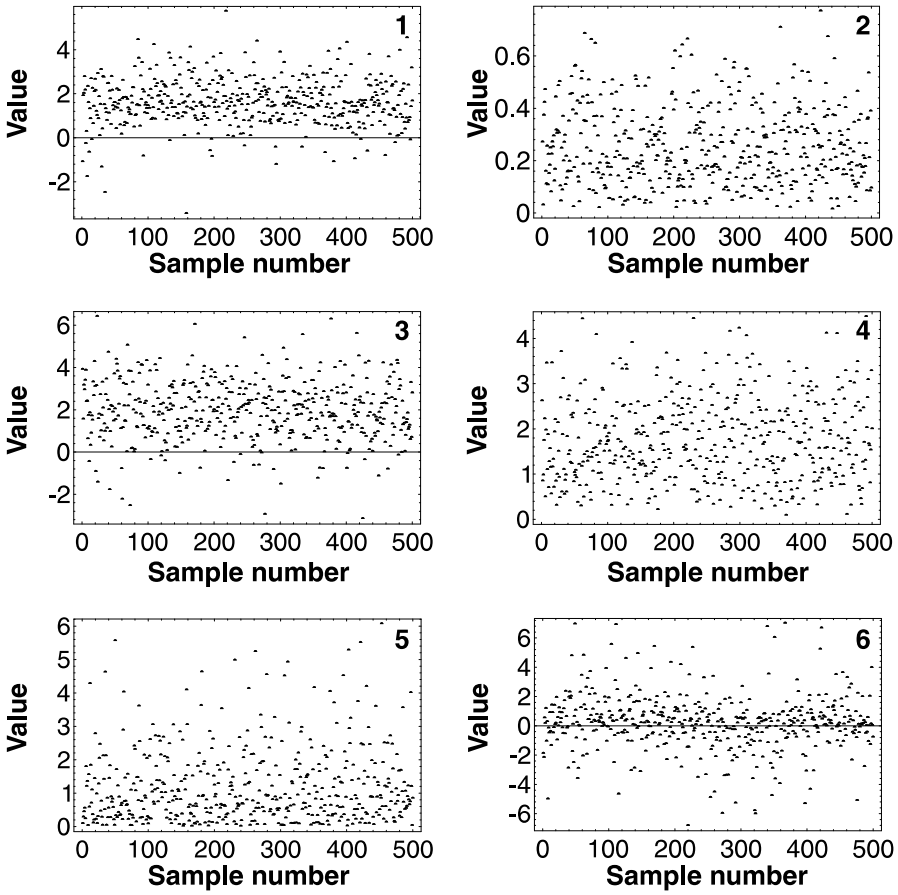


Figure 3. Random samples generated from a group of selected probability densities. Clockwise from upper right these are the beta, logistic, Rayleigh, exponential, Cauchy and Laplace. The group number is indicated in the upper right-hand side of each graph.

reasonable specification of an information measure will lead to essentially the same definition. We have shown that some of the renormalization procedures used to connect the discrete and continuous information measures are unnecessary and result from a failure to properly define the reference state of information. Finally, we have shown some applications of the principle of maximum entropy to the problem of computing *a priori* distributions from synthetic data as well as a technique for computing multi-dimensional maximum entropy distributions.

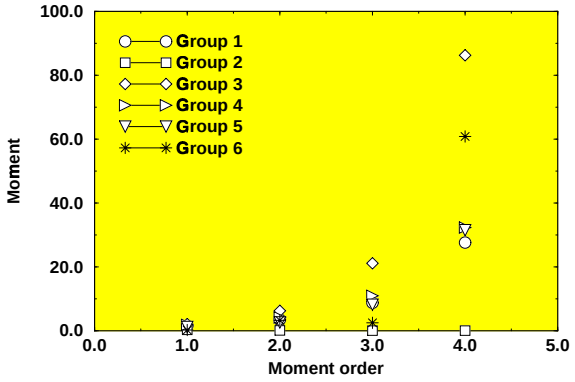


Figure 4. Sample moments up to the fourth-order computed for each group of one thousand samples as shown in Figure 3.

Acknowledgements

This work was partially supported by the sponsors of the Consortium Project on Seismic Inverse Methods for Complex Structures at the Center for Wave Phenomena, the sponsors of the Gravity and Magnetics Project, both at the Colorado School of Mines, the Shell Foundation and the Army Research Office. In addition, the second author acknowledges the support of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil).

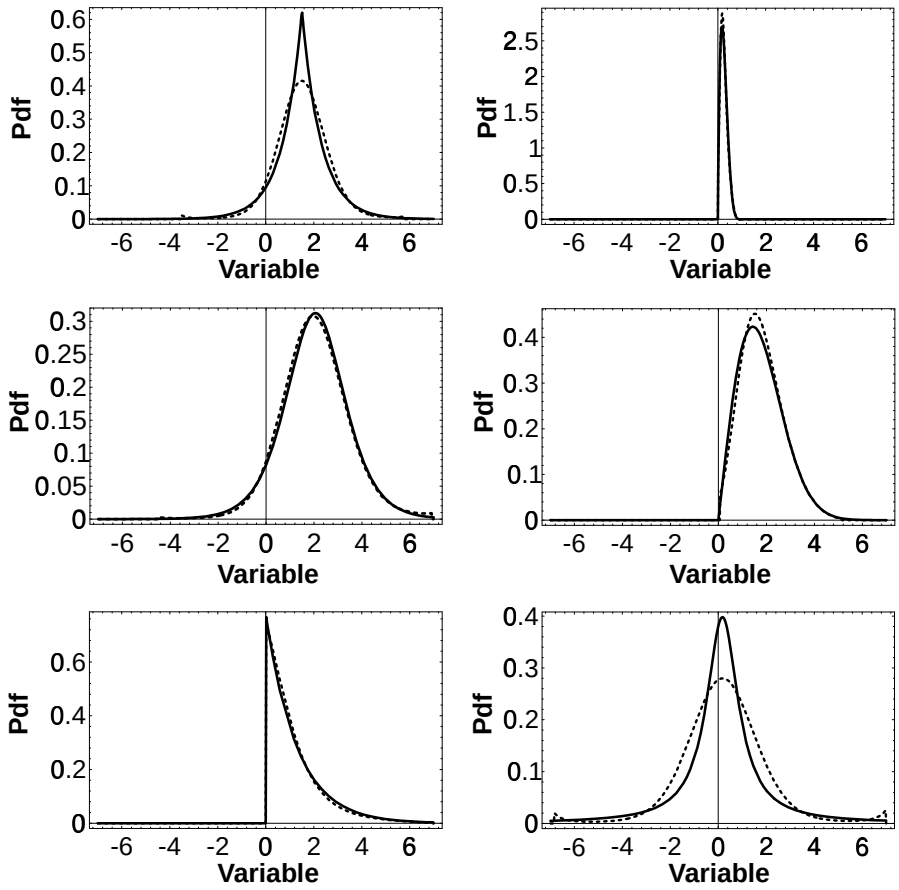


Figure 5. Comparison between the exact probability density (solid line) and the distribution computed using entropy maximization subjected to moment constraints up to the fourth-order (dashed line). The reference prior is uniform on [min, max], which are computed from the samples.

References

Feynman, R.P., Leighton, R.B., & Sands, M. 1963. *The Feynman Lectures on Physics.* Addison-Wesley.

Jaynes, E. T. 1957. Information theory and statistical mechanics. *Phys. Rev.*, **106**, 171–190.

Jumarie, G. 1990. Nonlinear filtering: A weighted mean squares approach and a Bayesian one via the maximum entropy principle. *Signal Processing*, **21**, 323–338.

Jaynes, E. T. 1968. Prior probabilities. *IEEE Transactions on Systems and Cybernetics*, **SSC-4**, 227–241.

Jaynes, E. T. 1982. On the rationale of maximum-entropy methods. *Proc. IEEE*, **70**, 939–952.

Khinchin, A. I. 1957. *Mathematical Foundations of Information Theory.* Dover Publications, Inc.

Kolmogorov, A. N. 1956. On the Shannon theory of information transmission in the case of continuous signals. *IEEE Trans. Inform. Theory*, **IT-2**, 102–108.

Kullback, S. 1959. *Information Theory and Statistics.* New York, N. Y.: Wiley. Plublished by Dover in 1968.

Mead, L. R., & Papanicolau, N. 1984. Maximum Entropy in the problem of moments. *J. Math. Phys.*, **25**, 2404–2417.

Rietsch, E. 1977. The maximum entropy approach to inverse problems. *J. Geophys.*, **42**, 489–506.

Scales, J. A., & Tarantola, A. 1994. *Bayesian inversion with realistic a priori information.* Tech. rept. 159. Center for Wave Phenomena, Colorado School of Mines, Golden, CO.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell Systems Technical Jour.*, **27**, 379–423,623–656.

Shore, J. E., & Johnson, R. W. 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross–entropy. *IEEE Trans. on Information Theory*, **IT-26**, 26–37.

Shore, J. E., & Johnson, R. W. 1981. Properties of

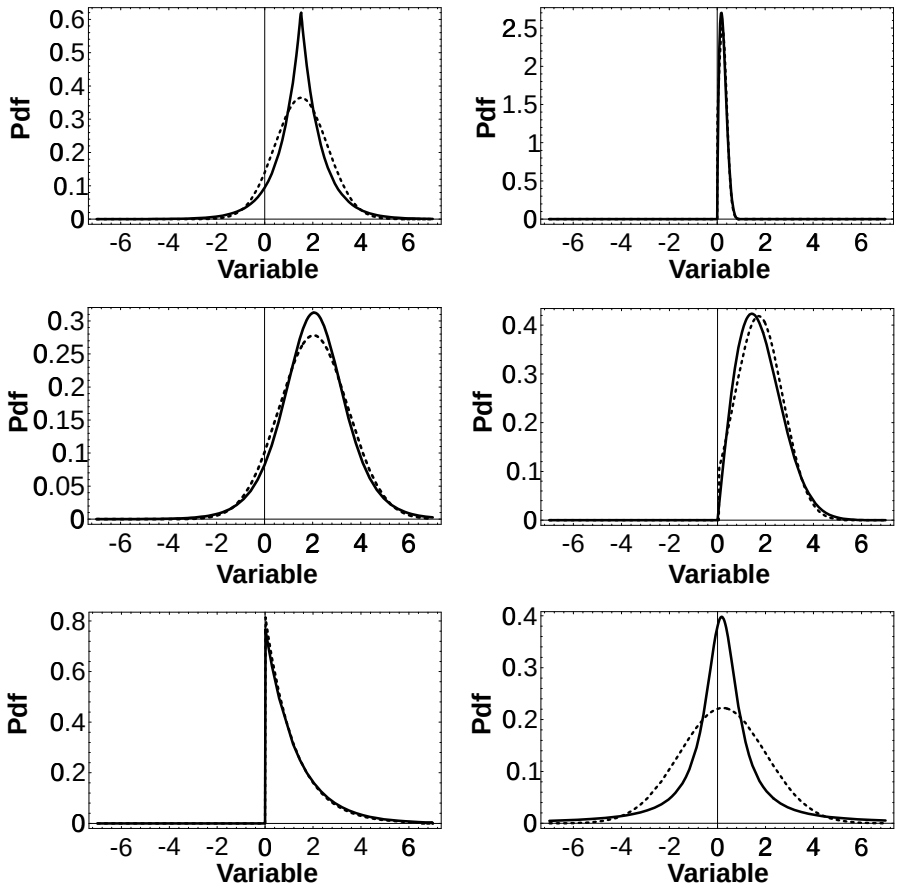


Figure 6. Comparison between the exact probability density (solid line) and the approximation computed using entropy maximization subjected to moment constraints up to the second-order (dashed line). The reference prior used is the uniform on $[min, max]$, which are computed from the samples.

cross–entropy minimization. *IEEE Trans. on Information Theory*, **IT-27**, 472–482.

Tarantola, A. 1987. *Inverse Problem Theory - Methods for Data Fitting and Model Parameter Estimation.* Elsevier.

Weiner, J.H. 1983. *Statistical Mechanics of Elasticity.* New York: Wiley-Interscience.

APPENDIX A: Postulates of Information

A1 Derivation of entropy

After Shannon's original derivation of the information entropy, several works were published giving alternative derivations. The majority of these derivations begin from a list of conditions that an adequate measure of information must satisfy, from which the entropy functional form was invariably achieved in a unique way. This rather informal approach led to criticisms that it could only indicate the plausibility of the entropy formula, but it did not rule out some other forms. Because of that, Shore & Johnson (1980) presented their derivation beginning from four axioms of consistency that any information-based method must satisfy. Their results pointed not only to the mathematical expression for entropy, but also to the principle of maximum entropy and minimum cross-entropy, which is the symmetric of relative entropy. In this section we will discuss the original arguments leading to the functional form of entropy given by Shannon (1948) .

Consider a set of propositions $A = \{A_1, \ldots, A_N\}$, with individual probabilities $P(A_i) = p_i$. Assuming that there exists a measure of total uncertainty H in a probability distribution, it must satisfy the following conditions:

(i) Continuity: H must be a continuous function of the individual probabilities p_i, which implies that a small change in the probability distribution corresponds to a small change in entropy;

(ii) Monotonicity: for the case when all propositions are equally likely ($p_i = \frac{1}{N}$), $H\left(\frac{1}{N}, \cdots, \frac{1}{N}\right)$ is a monotonic increasing function of N;

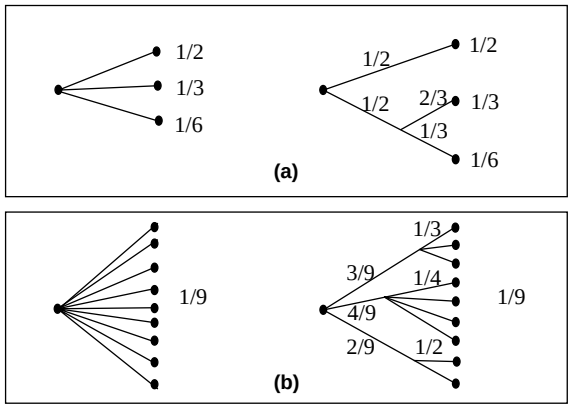


Figure A1. Schematic figure illustrating the composition law. (a) example given in Shannon (1948) and (b) example given in Jaynes (1957) for equally likely probabilities.

(iii) Composition law: consider the case where the propositions in A are grouped into M other propositions each with m_i elements, defining another set of propositions B. More precisely, $B = \{B_1, \ldots, B_M\}$, where $B_i = \{A_{p_i}, \ldots, A_{q_i}\}$ with $p_i = 1 + \sum_{j=0}^{i-1} m_j$ (with $m_0 = 0$) and $q_i = p_i + m_i - 1$. The measure H computed from the probabilities directly assigned to A will be the same as the weighted sum of individual measures H computed from probabilities assigned to B and to A conditional to B. Thus, if we denote $P(B_i) = q_i$, we have

$$H(p_1, \ldots, p_N) = H(q_1, \ldots, q_M) + q_1 H\left(\frac{p_1}{q_1}, \cdots, \frac{p_{m_1}}{q_1}\right) + \cdots + q_M H\left(\frac{p_k}{q_M}, \cdots, \frac{p_N}{q_M}\right) . \quad \text{(A1)}$$

The normalization of the probabilities in the right-hand side of Equation (A1) is considered since, after observing that B_j is true with probability q_j, the probability that $A_i \in B_j$ is true is given by p_i/q_j. One example of an application of this condition is illustrated in Figure A1 a. For this example

$$H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(1) + \frac{1}{2}H(\frac{2}{3}, \frac{1}{3})$$

but $H(1) = 0$.

To find the mathematical form for the measure H, consider the example depicted in Figure A1 b, which begins with a set A of equally likely propositions. When all the probabilities are equal, whatever turns out to be the final entropy function it will be completely defined by the number of possibilities (or propositions) N. Thus, to simplify the notation we may define $h(N) \equiv H\left(\frac{1}{N}, \cdots, \frac{1}{N}\right)$. Following the guidelines in Condition (iii), we can define the set B, in this example, as composed by three propositions $\{B_1, B_2, B_3\}$, where

$B_1 = \{A_1, A_2, A_3\}$,

$B_2 = \{A_4, A_5, A_6, A_7\}$ and

$B_3 = \{A_8, A_9\}$.

The probability for each B_i will be $q_1 = \frac{3}{9}$, $q_2 = \frac{4}{9}$ and $q_3 = \frac{2}{9}$, respectively. With the above determinations, we can just write the expression equivalent to Equation (A1), for the proposed example, as given by

$$h(9) = H\left(\frac{3}{9}, \frac{4}{9}, \frac{2}{9}\right) + \frac{3}{9}h(3) + \frac{4}{9}h(4) + \frac{2}{9}h(2) \quad \text{(A2)}$$

or

$$H\left(\frac{3}{9}, \frac{4}{9}, \frac{2}{9}\right) = h(9) - \left[\frac{3}{9}h(3) + \frac{4}{9}h(4) + \frac{2}{9}h(2)\right] \quad \text{(A3)}$$

The Equation (A3) is really the form that reflects the term that we are interested in, which is the entropy for general probabilities. To find an expression for the measure h, we can generalize Equation (A2) by letting m_i be the number of propositions A_i in each set B_i and the total number of propositions in B $(3 \to r)$ be arbitrary, which gives

$$h(\Sigma_i^r m_i) = H(p_1, \ldots, p_r) + \sum_{i=1}^{r} p_i\, h(m_i) . \quad \text{(A4)}$$

If we now let m_i be just a constant m we can reduce the above equation to

$$h(r\,m) = h(r) + h(m) , \quad \text{(A5)}$$

since $p_i = \frac{m_i}{M}$, where $M = \Sigma_i^r m_i$. This equation has a solution given by

$$h(m) = K \log m, \quad \text{(A6)}$$

where K is a multiplicative constant corresponding to the choice of basis for the logarithmic function. If we substitute (A6) back in Equation (A4), using $K = 1$, we get

$$\log M = H(p_1, \ldots, p_r) + \sum_{i=1}^{r} p_i \log m_i ,$$

$$H(p_1, \ldots, p_r) = \sum_{i=1}^{r} \frac{m_i}{M} \log M - \sum_{i=1}^{r} \frac{m_i}{M} \log m_i ,$$

$$H(p_1, \ldots, p_r) = -\sum_{i=1}^{r} \frac{m_i}{M} (\log m_i - \log M) ,$$

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i .$$

A2 Uniqueness

To show that (A6) is the unique solution of Equation (A5), Shannon (1948) used the fact that it can be extended by induction to yield

$$h(r\,m\,o\,p) = h(r) + h(m) + h(o) + h(p) + \cdots , \quad \text{(A7)}$$

or when all arguments are equal

$$h(r^{\,k}) = k\,h(r)\,. \tag{A8}$$

Now, using that for t and s integer, not less than 2, and n arbitrarily large, we can find an m such that

$$s^m \leq t^n < s^{m+1}\,.$$

Taking the logarithm of the above inequality and dividing by $n \log s$, we get

$$\frac{m}{n} \leq \frac{\log t}{\log s} < \frac{m+1}{n}$$

or

$$\left|\frac{m}{n} - \frac{\log t}{\log s}\right| < \epsilon\,,$$

where ϵ is an arbitrarily small positive constant. By the monotonicity condition of h we can write

$$h(s^m) \leq h(t^n) < h(s^{m+1})$$

or

$$m\,h(s) \leq n\,h(t) < (m+1)\,h(s)\,.$$

Similarly, we may divide last inequality by $n\,h(s)$ to get

$$\frac{m}{n} \leq \frac{h(t)}{h(s)} < \frac{m+1}{n} \qquad \text{or} \qquad \left|\frac{m}{n} - \frac{h(t)}{h(s)}\right| < \epsilon\,,$$

which imply that

$$\left|\frac{h(t)}{h(s)} - \frac{\log t}{\log s}\right| \leq 2\epsilon\,.$$

The above expression is uniquely satisfied for

$$h(t) = -K \log t. \tag{A9}$$

A3 Khinchin's derivation

Following Khinchin (1957), the entropy functional can also be uniquely defined from the following conditions

(i) The uncertainty associated with a finite complete scheme A takes its largest value if all the events are equally likely, which means

$$H(p_1, \ldots, p_n) \leq H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right).$$

(ii) Addition of an impossible event to a scheme does not change the amount of uncertainty.

(iii) The uncertainty in the product AB of two schemes A an B is equal to the uncertainty in scheme A increased by the uncertainty remaining in scheme B after the realization of scheme A. This means that

$$H(AB) = H(A) + H_A(B),$$

where H_A is the *conditional entropy*. Below we show that the conditional entropy corresponds to a weighted sum in Shannon's original derivation discussed in the previous section.

Next, Khinchin considers that schemes A and B are divided into n and m events respectively. Thus, the product AB will have nm events with probabilities given by

$$AB = \begin{pmatrix} A_1B_1 & A_1B_2 & \cdots & A_1B_m & \cdots & A_nB_m \\ p_1q_{11} & p_1q_{12} & \cdots & p_1q_{1m} & \cdots & p_nq_{nm} \end{pmatrix}. \tag{A10}$$

The components of the expression given in item (iii) above will be defined as follows:

$$\begin{aligned} H(AB) &= H(p_1q_{11}, p_1q_{12}, \ldots, p_nq_{nm}), \\ H(A) &= H(p_1, \ldots, p_n) \qquad \text{and} \\ H_A(B) &= \sum_{i=1}^{n} p_i\, H(q_{i1}, \ldots, q_{im}). \end{aligned} \tag{A11}$$

We can see that this will lead to the same equations of the previous section by explicitly writing the values for the probability for the events in (A10). In doing so, we get

$$A_1B = \begin{pmatrix} A_1B_1 & A_1B_2 & A_1B_3 & A_1B_4 \\ \frac{3}{9}\frac{1}{3} & \frac{3}{9}\frac{1}{3} & \frac{3}{9}\frac{1}{3} & \frac{3}{9}0 \\ A_1B_5 & A_1B_6 & A_1B_7 & A_1B_8 & A_1B_9 \\ \frac{3}{9}0 & \frac{3}{9}0 & \frac{3}{9}0 & \frac{3}{9}0 & \frac{3}{9}0 \end{pmatrix},$$

$$A_2B = \begin{pmatrix} A_2B_1 & A_2B_2 & A_2B_3 & A_2B_4 \\ \frac{4}{9}0 & \frac{4}{9}0 & \frac{4}{9}0 & \frac{4}{9}\frac{1}{4} \\ A_2B_5 & A_2B_6 & A_2B_7 & A_2B_8 & A_2B_9 \\ \frac{4}{9}\frac{1}{4} & \frac{4}{9}\frac{1}{4} & \frac{4}{9}\frac{1}{4} & \frac{4}{9}0 & \frac{4}{9}0 \end{pmatrix},$$

$$A_3B = \begin{pmatrix} A_3B_1 & A_3B_2 & A_3B_3 & A_3B_4 \\ \frac{2}{9}0 & \frac{2}{9}0 & \frac{2}{9}0 & \frac{2}{9}0 \\ A_3B_5 & A_3B_6 & A_3B_7 & A_3B_8 & A_3B_9 \\ \frac{2}{9}0 & \frac{2}{9}0 & \frac{2}{9}0 & \frac{2}{9}\frac{1}{2} & \frac{2}{9}\frac{1}{2} \end{pmatrix}.$$

If we substitute those probability assignments in Equations (A11), using the condition 2, we get exactly Equation (A2), since

$$H(p_1q_{11}, p_1q_{12}, \ldots, p_nq_{nm}) \to h(9)$$

and

$$H(q_{i1}, \ldots, q_{im}) \to h(n_i).$$

A4 Frequency connection

Consider the classic problem of a random experiment on the variable X, which can take the values x_k, for $k = 1, \ldots, K$. Thus in N trials we have K^N possible outcomes yielding a set of observed frequencies for each possible value of X given by

$$f_i = \frac{n_i}{N}. \tag{A12}$$

This experiment can be repeated many times. Thus, we may wonder: what set of frequencies $\{f_i\}$ can be realized in the greatest number of ways? A possible way to find this is by maximizing the multiplicity (subjected to any set of linear constraints), which is given by

$$W(f_1, \ldots, f_K) = \frac{N!}{(Nf_1)! \cdots (Nf_K)!}. \tag{A13}$$

Or alternatively by maximizing any other monotonic function of $W(f_1, \ldots, f_K)$. Jaynes (1982) shows that when $N \to \infty$, $n_i \to \infty$ so that $f_i \to$ constant, using Stirling's approximation for the factorial

$$N^{-1} \log W(f_1, \ldots, f_K) \to -\sum_{i=1}^{K} f_i \log f_i.$$

This means that for N large, by maximizing the entropy we also find the set of frequencies that can be realized in greatest number of ways.

A natural question that may arise concerns how far, in an entropy sense, other possible distributions are from the one of maximum entropy. That is, a certain fraction F of all runs of the experiment will yield distributions of frequencies with entropies in the range

$$H_{\max} - \Delta H \leq H(f_1, \ldots, f_K) \leq H_{\max}.$$

This is summarized by the Jaynes' Entropy Concentration theorem (Jaynes, 1982), which says that $2N\Delta H$ is asymptotically χ^2 distributed with $\nu = K - M - 1$ degrees of freedom, where K is the number of probabilities (frequencies) M is the number of linear constraints in the maximization problem (same as the number of multipliers (excluding l_0)) and the 1 results from the normalization constraint $\sum_i f_i = 1$. Thus, in terms of the upper tail area $(1 - F)$, ΔH is given by

$$2N\Delta H = \chi^2_\nu (1 - F)\,. \tag{A14}$$

APPENDIX B: Examples

The principle of maximum entropy provides the most conservative probability distribution (i.e., the least informative) $p(\mathbf{x})$ that is consistent with moments of the underlying process. In this section, we illustrate via examples the computation of such distributions when moments up to the second order are available. We will work mainly in the multidimensional continuous case. The unidimensional case can be derived from those results in a straightforward way. It is our intention to be specific about which form of the reference state of information given by $q(\mathbf{x})$ is used in the entropy computations.

As mentioned before, the general expression for the maximum entropy probability density is

$$\begin{aligned} p(\mathbf{x}) &= -\kappa q(\mathbf{x}) \exp\left(-\lambda_0 - \lambda_i x_i - \right. \\ &\quad \left. \lambda_{ij} x_i x_j - \lambda_{ijk} x_i x_j x_k - \ldots\right) \end{aligned} \tag{B1}$$

where the λ are Lagrange multipliers and κ is a normalization constant that from now on will be incorporated into λ_0. Here, we use Einstein's summation convention such that $\lambda_{ij} x_i x_j = \sum_i \sum_j \lambda_{ij} x_i x_j$. Direct substitution of Equation (B1) into the definition of H provides the expression of the entropy for this probability distribution function:

$$\begin{aligned} H\left[p(\mathbf{x}); q(\mathbf{x})\right] &= \lambda_0 + \lambda_i \langle x_i \rangle + \\ &\quad \lambda_{ij} \langle x_i x_j \rangle + \lambda_{ijk} \langle x_i x_j x_k \rangle \ldots \end{aligned} \tag{B2}$$

Here, $\langle\rangle$ means the expectation with respect to $p(x)$. Equation (B2) is helpful for the calculation of the relative entropy of the probability distributions studied in this section.

B1 Zero-th Order Moment

Consider the case in which the new state of information described by $p(\mathbf{x})$ consists of only the zero-th order moment of $p(\mathbf{x})$. That is, the only "new" information available to the problem is:

$$\int_V p(\mathbf{x}) d\mathbf{x} = 1 \tag{B3}$$

In this case, the maximum entropy distribution is given by:

$$p(\mathbf{x}) = q(\mathbf{x}) \exp\left(-\lambda_0\right). \tag{B4}$$

Of course because $q(\mathbf{x})$ is normalized the constant λ_0 is equal to zero. The relative entropy entropy as given by Equation (B2) is equal to zero as expected since the probability function $p(\mathbf{x})$ is equal to the reference state of information $q(\mathbf{x})$.

B2 First-order moments

Now assume that the zero-th and the first order moments are available for the maximum entropy computation:

$$\int_V p(\mathbf{x}) d\mathbf{x} = 1. \tag{B5}$$

$$\int_V x_i p(\mathbf{x}) d\mathbf{x} = \langle x_i \rangle, \quad i = 1, ..., n. \tag{B6}$$

Here, n is the dimension of the process. We will assume that the prior probability is uniform within the interval $[0, \mathbf{X}]$, as given by the following expression:

$$q(\mathbf{x}) = \kappa \mathcal{R}\left[0, \mathbf{X}\right] = \begin{cases} \kappa & \text{if } 0 \leq x_i \leq X_i \\ 0 & \text{otherwise} \end{cases} \tag{B7}$$

where κ is a normalization constant. The maximum entropy distribution associated with Equations (B6) and (B5) is given by:

$$p(\mathbf{x}) = \mathcal{R}[0, \mathbf{x}] \exp\left[-\lambda_0 - \lambda_i x_i\right], \tag{B8}$$

where κ was incorporated into λ_0. So in this case, the maximum entropy distribution is a *truncated* exponential distribution. This distribution arises in the situation when n independent exponential random variables are restricted to be in a given region of R^n, specified by $\mathcal{R}[0, \mathbf{X}]$. The normalization constant λ_0 is given by:

$$\lambda_0 = \sum_{i=1}^{n} \log\left\{\frac{1}{\lambda_i}\left[1 - \exp\left(-\lambda_i X_i\right)\right]\right\}. \tag{B9}$$

And the mean of this distribution is given by:

$$\langle x_i \rangle = \frac{1}{\lambda_i} \frac{\left[1 - (1 + \lambda_i X_i)\exp\left(-\lambda_i X_i\right)\right]}{1 - \exp\left(-\lambda_i X_i\right)}. \tag{B10}$$

Notice that the mean and the unknown Lagrange multipliers are related in a nonlinear way. Iterative numerical procedures are likely to be the method of choice for the determination of those parameters. Once these parameters are computed, the relative entropy of the distribution given in Equation (B8), can be computed from Equation (B2), yielding:

$$\begin{aligned} H[p(\mathbf{x}); q(\mathbf{x})] &= \sum_{i=1}^{n} \log\left\{\frac{1}{\lambda_i}\left[1 - \exp\left(-\lambda_i X_i\right)\right]\right\} \\ &+ \frac{\left[1 - (1 + \lambda_i X_i)\exp\left(-\lambda_i X_i\right)\right]}{1 - \exp\left(-\lambda_i X_i\right)}. \end{aligned} \tag{B11}$$

In the situation of a less informative prior $q(\mathbf{x})$, i.e. for large values of the components of $\mathbf{X}$, Equations (B9), (B10) and (B11) can be approximated by

$$\lambda_0 = \sum_{i=1}^{n} \log\left(\frac{1}{\lambda_i}\right), \tag{B12}$$

$$\langle x_i \rangle = \frac{1}{\lambda_i} \tag{B13}$$

$$H[p(\mathbf{x}); q(\mathbf{x})] = \sum_{i=1}^{n} \log\left(\langle x_i \rangle\right) + n. \tag{B14}$$

Clearly in this case, the maximum entropy distribution (Equation (B8)) can be approximated by a exponential distribution. This example illustrates that, if only the first moments $\langle x_i \rangle$ are used *and* when the prior distribution $q(\mathbf{x})$ assumes the form of Equation (B7) the probability density function derived from the maximization of the entropy is a truncated exponential. Results will change if a different reference state of information $q(\mathbf{x})$ is used.

B3 Second-order moments

Consider next the situation where moments up to the second order are available for the maximum entropy calculation:

$$\int_V p(\mathbf{x}) d\mathbf{x} = 1. \tag{B15}$$

$$\int_V x_i p(\mathbf{x}) d\mathbf{x} = \langle x_i \rangle, \quad i = 1, \ldots, n. \tag{B16}$$

$$\int_V x_i x_j p(\mathbf{x}) d\mathbf{x} = \langle x_i x_j \rangle, \quad i, j = 1, \ldots, n. \tag{B17}$$

The probability distribution that maximizes the relative entropy functional is given by:

$$p(\mathbf{x}) = q(\mathbf{x}) \exp\left(-\lambda_0 - \lambda_i x_i - \lambda_{ij} x_i x_j\right). \tag{B18}$$

Assuming that the reference state of information is given by an uniform probability distribution function such as

$$q(\mathbf{x}) = \kappa \mathcal{R}\left[-\mathbf{X}, \mathbf{X}\right] = \begin{cases} 1 & \text{if } -X_i \le x_i \le X_i \\ 0 & \text{otherwise.} \end{cases} \tag{B19}$$

where κ is a normalization constant. Therefore, we have for the maximum entropy distribution

$$p(\mathbf{x}) = \mathcal{R}[-\mathbf{X}, \mathbf{X}] \exp\left(-\lambda_0 - \lambda_i x_i - \lambda_{ij} x_i x_j\right). \tag{B20}$$

As before κ was absorbed into λ_0. This equation defines a *truncated* Gaussian probability distribution. In this case, it is complicated to derive explicit expressions relating the Lagrange multipliers and the parameters (mean and covariances) of this distribution, as was done in the previous section, due to the complexity of the integrations. A diagonalization of the matrix λ_{ij} facilitates this matter but still the algebra involved is rather cumbersome in the multidimensional case. In the one-dimensional case, the problem is more tractable, and for the probability distribution given by

$$p(x_1) = \mathcal{R}[-X_1, X_1] \exp\left(-\lambda_0 - \lambda_1 x_1 - \lambda_{11} x_1^2\right), \tag{B21}$$

the following expressions can be derived:

$$\lambda_0 = \log \Lambda, \tag{B22}$$

$$\begin{aligned} \langle x_1 \rangle &= \frac{1}{2\lambda_{11}} \left\{\frac{1}{\Lambda}\left[\exp\left(-\lambda_1 X_1 - \lambda_{11} X_1^2\right)\right.\right. \\ &- \left.\left.\exp\left(-\lambda_1 X_1 - \lambda_{11} X_1^2\right)\right] - \lambda_1\right\}, \end{aligned} \tag{B23}$$

and

$$\begin{aligned} \langle x_1^2 \rangle = \frac{1}{4\lambda_{11}^2} \left\{\left[\frac{(\lambda_1 - 2\lambda_{11} X_1)}{\Lambda} \exp\left(-\lambda_1 X_1 - \lambda_{11} X_1^2\right) - \right.\right. \\ \left.\left.(\lambda_1 + 2\lambda_{11} X_1)\, \exp\left(-\lambda_1 X_1 - \lambda_{11} X_1^2\right)\right] + \lambda_1^2 + 2\lambda_{11}\right\} \end{aligned} \tag{B24}$$

where Λ is given by

$$\Lambda = \frac{\sqrt{\pi}}{2\sqrt{\lambda_{11}}} \exp\left(\frac{\lambda_1^2}{4\lambda_{11}}\right) \left\{\mathrm{Erf}\left[\sqrt{\lambda_{11}}\left(\frac{\lambda_1}{2\lambda_{11}} + X_1\right)\right]\right.$$

$$- \text{Erf}\left[\sqrt{\lambda_{11}}\left(\frac{\lambda_1}{2\lambda_{11}} - X_1\right)\right]\Big\}, \quad \text{(B25)}$$

where Erf[.] is the error function. As in the exponential case the Lagrange multipliers and the moments of the distribution are related in a nonlinear fashion. Again iterative procedure should probably be used to solve for the Lagrange multipliers in Equations (B23) and (B24). Once they are computed, the entropy of the distribution expressed in Equation (B21) is derived from Equation (B2). The result follows

$$\begin{aligned} & H\left[p(x_1); \mathcal{R}[-X_1, X_1]\right] = \\ & -\log\frac{1}{\Lambda} + \frac{\lambda_1}{2\lambda_{11}}\left\{\frac{1}{\Lambda}\left[\exp\left(-\lambda_1 X_1 - \lambda_{11}X_1^2\right)\right.\right. \\ & \left.\left. - \exp\left(-\lambda_1 X_1 - \lambda_{11}X_1^2\right)\right] - \lambda_1\right\} \\ & + \frac{1}{4\lambda_{11}}\left\{\frac{1}{\Lambda}\left[(\lambda_1 - 2\lambda_{11}X_1)\exp\left(-\lambda_1 X_1 - \lambda_{11}X_1^2\right)\right.\right. \\ & \left.\left. - \ (\lambda_1 + 2\lambda_{11}X_1)\exp\left(-\lambda_1 X_1 - \lambda_{11}X_1^2\right)\right] \right. \\ & \left. + \lambda_1^2 + 2\lambda_{11}\right\}. \end{aligned} \quad \text{(B26)}$$

As the components of $\mathbf{X}$ increase, Equation (B20) approaches a multidimensional Gaussian probability distribution given by

$$p(\mathbf{x}) = \sqrt{\frac{\pi^{-n}}{2\det C}}\exp\left[-\frac{1}{2}(x_i - x_{0i})C_{ij}^{-1}(x_j - x_{0j})\right], \quad \text{(B27)}$$

where n is the dimension of the variable $\mathbf{x}$, x_{0i} is the $i-th$ component of the mean vector, C_{ij}^{-1} is the $ij-th$ element of the inverse covariance matrix and $\det C$ its determinant. In this case, a direct comparison between Equations (B20) and (B27) yields the following relationships between the Lagrange multipliers and the mean and covariances of the Gaussian distribution

$$\begin{aligned} & \lambda_0 = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det C + \frac{1}{2}x_{0i}C_{ij}^{-1}x_{0j}, \\ & \lambda_i = -x_{0j}C_{ij}^{-1}, \ \ \textit{and} \\ & \lambda_{ij} = \frac{1}{2}C_{ij}^{-1}. \end{aligned} \quad \text{(B28)}$$

Use in Equation (B2) of the Lagrange multipliers defined in Equation (B28) yields

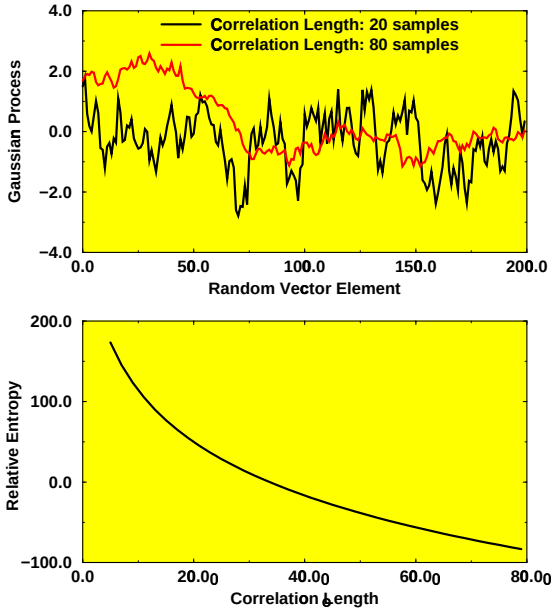
$$\begin{aligned} H\left[p(\mathbf{x}); q(\mathbf{x})\right] &= \lambda_0 + \lambda_i\langle x_i\rangle + \lambda_{ij}\langle x_i x_j\rangle \\ &= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det C + \frac{1}{2}x_{0i}C_{ij}^{-1}x_{0j} - \\ & \quad x_{0i}C_{ij}^{-1}x_{0j} + \frac{1}{2}C_{ij}^{-1}\langle x_i x_j\rangle \\ &= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det C + \frac{1}{2}C_{ij}^{-1}\left(\langle x_i x_j\rangle - \right. \\ & \quad \left.\langle x_i\rangle\langle x_j\rangle\right). \end{aligned} \quad \text{(B29)}$$

Considering in the above equation that:

$$C_{ij} = \langle(x_i - x_{0i})(x_j - x_{0j})\rangle = \langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle, \quad \text{(B30)}$$

we finally obtain:

Figure B1. (a) Gaussian correlated sequences. (b) Entropy as a function of the correlation length.



$$\begin{aligned} H\left[p(\mathbf{x}); q(\mathbf{x})\right] &= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det C + \frac{1}{2}C_{ij}^{-1}C_{ij} \\ &= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det C + \frac{n}{2}. \end{aligned} \quad \text{(B31)}$$

This equation is an approximate expression for the relative entropy of the maximum entropy distribution computed in the situation where up to the second order moments are known *and* the prior reference state of information is given by Equation (B19), when the components of $\mathbf{X}$ are large. It is left to the interested reader to check that Equation (B26) reduces to Equation (B31) in the one-dimensional case when X_1 gets large enough.

Figure (B1) illustrates the behavior of Equation (B31) for a multidimensional Gaussian process of dimension 200 as a function of its correlation length. As expected for smaller correlation lengths the relative entropy should be larger due to the more erratic pattern of the random sequence. As the correlation length increases the information content of the sequence increases, and so the relative entropy decreases. Notice that it is possible to have a *negative* relative entropy as Equation (B31) indicates. Tarantola (1987) worked out an example similar to the one just discussed, however with the reference state of information described by a multidimensional Gaussian distribution. In that case, we have for for the distributions $p(\mathbf{x})$ and $q(\mathbf{x})$ the following expressions:

$$p(\mathbf{x}) = \sqrt{\frac{(2\pi)^n}{\det C^p}}\exp\left[-\frac{1}{2}(x_i - x_i^p)C_{ij}^{p\,-1}(x_j - x_j^p)\right], \quad \text{(B32)}$$

and

$$q(\mathbf{x}) = \sqrt{\frac{(2\pi)^n}{\det C^q}} \exp\left[-\frac{1}{2}(x_i - x_i^q)C_{ij}^{q\,-1}(x_j - x_j^q)\right]. \quad \text{(B33)}$$

Where C^p and C^q are the covariance matrices, and x_i^p and x_i^q the $i-th$ component of the mean vector of the probability distributions $p(\mathbf{x})$ and $q(\mathbf{x})$ respectively. Computation of the relative entropy given by Equation (B2) is straightforward but cumbersome. The main steps of this computation are shown in Equation (B34)

$$\begin{aligned}
H\left[p(\mathbf{x}); q(\mathbf{x})\right] = & -\int_V p(\mathbf{x}) \log\left\{\sqrt{\frac{\det C^q}{\det C^p}}\right. \\
& \exp\left[-\frac{1}{2}(x_i - x_i^p)C_{ij}^{p\,-1}(x_j - x_j^p)\right. \\
& \left.\left. + \frac{1}{2}(x_i - x_i^q)C_{ij}^{q\,-1}(x_j - x_j^q)\right]\right\} d\mathbf{x} \\
= & -\int_V p(\mathbf{x}) \left[\kappa - \frac{1}{2}\left(x_i C_{ij}^{p\,-1} x_j - 2x_i^p C_{ij}^{p\,-1} x_j + \right.\right. \\
& \left. x_i^p C_{ij}^{p\,-1} x_j^p\right) + \frac{1}{2}\left(x_i C_{ij}^{q\,-1} x_j - 2x_i^q C_{ij}^{q\,-1} x_j + \right. \\
& \left.\left. x_i^q C_{ij}^{q\,-1} x_j^q d\right)\right] d\mathbf{x} \\
= & -\kappa + \frac{1}{2}\langle x_i C_{ij}^{p\,-1} x_j\rangle - \frac{1}{2}\langle x_i C^{q\,-1} x_j\rangle - \\
& \frac{1}{2}x_i^p C_{ij}^{p\,-1} x_j^p - \frac{1}{2}x_i^q C_{ij}^{q\,-1} x_j^q + x_i^q C_{ij}^{q\,-1} x_j^p \\
= & -\kappa + \frac{1}{2}C_{ij}^{p\,-1}\left(\langle x_i x_j\rangle - x_i^p x_j^p\right) - \frac{1}{2}C_{ij}^{q\,-1}\langle x_i x_j\rangle \\
& -\frac{1}{2}x_i^q C_{ij}^{q\,-1} x_j^q + x_i^q C_{ij}^{q\,-1} x_j^p \\
= & -\kappa + \frac{1}{2}C_{ij}^{p\,-1} C_{ij}^p - \frac{1}{2}C_{ij}^{q\,-1}\langle x_i x_j\rangle - \\
& \frac{1}{2}x_i^q C_{ij}^{q\,-1} x_j^q + x_i^q C_{ij}^{q\,-1} x_j^p \\
= & -\kappa + \frac{1}{2}\text{trace}\,(I) - \frac{1}{2}C_{ij}^{q\,-1}\left[C_{ij}^p + x_i^p x_j^p\right] - \\
& \frac{1}{2}x_i^q C_{ij}^{q\,-1} x_j^q + x_i^q C_{ij}^{q\,-1} x_j^p \\
= & -\kappa - \frac{1}{2}\text{trace}\left(C^{q\,-1} C^p - I\right) - \\
& \frac{1}{2}(x_i^p + x_i^q)C_{ij}^{q\,-1}(x_j^p + x_j^q) + x_i^p C_{ij}^{q\,-1} x_j^q, \qquad \text{(B34)}
\end{aligned}$$

where

$$\kappa = \log\sqrt{\frac{\det C^q}{\det C^p}}.$$

In the above derivation the following relationships were used:

$$\begin{aligned}
& C_{ij} = \langle x_i x_j\rangle - \langle x_i\rangle\langle x_i\rangle, \text{ and} \\
& A_{ij}B_{ij} = \text{trace}\,(AB). \qquad \text{(B35)}
\end{aligned}$$

The last expression holds if A and B are symmetric matrices. Equation (B34) corrects a minor error in the derivation shown in Tarantola (1987). Figure (B2) illustrates, for a random process of dimension 100, the relative entropy derived in Equation (B34) as a function of the ratio of the correlation lengths of the probability functions $p(\mathbf{x})$ and $q(\mathbf{x})$. Recall that a larger correlation length implies a more informative distribution.

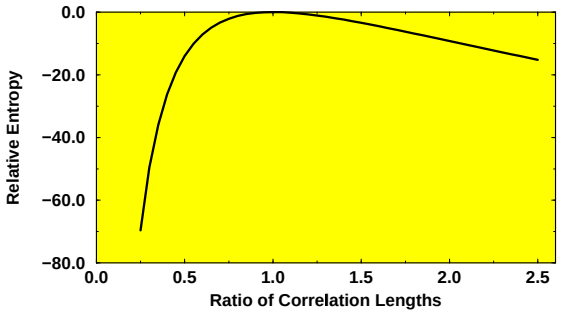


Figure B2. Relative entropy as a function of the ratio of the correlation lengths of $p(\mathbf{x})$ and $q(\mathbf{x})$.

Note the similarity between Figures (2) and (B2). This is expected since both Figures display the relative entropy $H[p(\mathbf{x}); q(\mathbf{x})]$ in the situation where $p(\mathbf{x})$ ranges from *less* informative (ratio of correlation lengths less than 1) to *more* informative (ratio of correlation lengths greater than 1) than $q(\mathbf{x})$. Again one would expect that the relative entropy would decrease smoothly as $p(\mathbf{x})$ gets progressively more informative and not the behavior shown in Figure (B2), for the case in which the ratio of correlation lengths is smaller than 1.

As a final remark notice that the distribution that maximizes the entropy when just the first and second moments are specified is not necessarily Gaussian as often said, although it was in the above example. As should be clear from these examples, such computation depends on the prior probability $q(\mathbf{x})$.