About two ways to combine information; maximum entropy and conjunction, and the role of the null-distribution in these.

Rob Devilee.

November 2, 1998

Abstract

I give an overview of two probabilistic (bayesian) ways of combining two pieces of information, which are defined in terms of (probability-) density functions. Tarantola (1982) uses the conjunction of two states. Rietsch (1977) uses the entropy formalism. In both methods, the null-distribution plays an important role to make the two measures invariant. Differences and properties of the two methods are explained.

1 Introduction.

There is an ambiguity in designing measures of uncertainty and of the information content of the resulting state of information when two pieces of information are combined. I discuss two methods: the maximum entropy method by Rietsch (1977) and the conjunction of two states by Tarantola and Valette (1982). The first method is intuitively appealing, the second is more stable in practice. Both measures are made invariant with the help of the so-called null-distribution, also referred to as 'weighting function'.This distribution eliminates 'trivial' information that is present in all d.f.'s before they are joined, namely the information due to a particular parameterization. In this view, the null-distribution standardizes the solution, in terms of a reference physical framework; the choice of such a framework is not always ambiguous, but since the best frameworks are invariant. At the end of the overview, I discuss some examples of null-distributions which can be applied in inversions and conclude by performing a simple synthetic experiment which highlights all the issues this overview adresses.

The motivation for this overview comes from the fact that I've had enormous trouble to understand these measures in the way they are presented by different authors. For example, it is not clear what kind of information the measures present. In particular the concept of the so-called null-distribution is vague. This text attempts to clarify the first issue; the issue of the null-distribution is discussed in another text.

2 Deterministic versus probabilistic

First let us see what happens if we have deterministic information, in the form of 2 functions. Then we have two conditions, for example $y_1(x) = 1 + x$ and $y_2(x) = 3 - x$, and only one point exists where both conditions are satisfied, namely the one defined by

$$y_1(x) = y_2(x)$$

and the point is at:

$$x = 1 \quad ; \quad y = y_1 = y_2 = 2.$$

However, when we have two uncertain constraints, usually there exists a whole distribution of points (x, y) that satisfy both constraints $f_1(x, y)$ and $f_2(x, y)$, though each with different accuracy. Such a set of solutions can be described by a general function $\sigma(x, y)$, where the value of σ defines the likelihood with which a point (x, y) satisfies both conditions.

$$\sigma(x, y) = g(f_1(x), f_2(x))$$

For example, the deterministic case can be rewritten into a form using probability density functions:

$$\begin{aligned} f_1(x, y) = &\quad \delta(y - y_1(x)) \\ f_2(x, y) = &\quad \delta(y - y_2(x)) \end{aligned}$$

These density functions are infinitely narrow distribution of points centered around where the lines would be in the deterministic case. If we require that both constraints are satisfied, we can for example multiply the two (i.e. we take the conjunction and use a uniform null-distribution; see section 2.3):

$$\begin{aligned} \sigma(x, y) &= f_1(x, y) f_2(x, y) \\ &= \delta(y - y_1(x)) \delta(y - y_2(x)) \end{aligned}$$

which has zero probability everywhere, except at the point (x, y) defined by:

$$y = y_1(x) = y_2(x)$$

which is the point of inter-section of the two lines of the deterministic example.

3 Methods of measuring information contents.

When information content is to be measured, first firtly we define the concept of information content, then we define a way of measuring it. The latter step is ambiguous, and depends utterly on the requirement imposed on the problem by the user.

The information content is defined as the uncertainty of reproducing a set of values obtained from an experiment. When the problem is deterministic, the outcome of an experiment is perfectly predictable: the information content is

1. When each value in the set has equal probability, it is impossible to put any constraints on a (future) outcome of an experiment: the information content is 0. Of particular interest is, to have a measure of the way information changes when two pieces of information are combined.

The information content which results after combination of information is measured by the entropy, or log-likelihood function. In a different way it is also measured by the conjunction? It depends on the requirements formulated by the person who designes the experiment to choose the best method.

3.1 Entropy

The entropy gives a measure of the information content of a probability function P and can also be used to measure the fit between two distributions P and Q.

3.1.1 Reasoning

We follow [Rietsch, 1977] here. Suppose we have a set of values $\{x_i : i = 1..N\}$, with probabilities of occurence given by $p(x_i) = p_i$. How likely is it, that such a particular set of values is measured? What does this say about the values $\{x'_i : i = 1..N\}$ we would get if the experiment was repeated? To predict the outcome of an experiment, it is desirable to have a measure H for the amount of uncertainty in a particular scheme. We need 3 constraints to arrive at a proper definition:

1) The uncertainty takes its largest value if all values x_i are equally likely, i.e. $p_i = p_j$. This implies that

$$H(p_1, ..., p_N) \leq H(\frac{1}{N}, ..., \frac{1}{N})$$

2) The addition of an impossible value x_{N+1} (with $p_{N+1} = 0$) to the set of measured values does not change the amount of uncertainty.

$$H(p_1, ..., p_N) = H(p_1, ..., p_N, 0)$$

It is a limiting case, which can never happen. Intuitively it can be understood as follows: if a value is (almost) never measured, it forms an (almost) neglectable part of the set of measurements. Note that the value of the function tends to approach zero when the probability approaches zero (see the plot).

$$\lim_{p \to 0} (-p \log p) \to 0$$

because we have

$$\lim_{p \to 0} p^p \to 1$$

3) The uncertainty of measuring the product $x_i y_j$ of the values of two different random variables X and Y, is equal to the uncertainty in x_i increased by the uncertainty remaining in y_j after x_i was realized.

$$H(XY) = H(X) + H(Y|X)$$

These three conditions are uniquely constrained only by the following function:

$$H(p_1 ... p_N) = -c \sum_{j=1}^{N} p_j \log p_j$$

which is called the entropy, in analogy with the entropy in thermodynamics. The entropy achieves its zero-crossing ($H = 0$) if the outcome of an experiment is certain, or deterministic. In such a case only one value is possible, i.e. all p_i are 0 except for one. The entropy achieves its maximum when all values are equally likely: $p_i = \frac{1}{N}$: this is a case of total ignorance (if we scale by a null-distribution, which we introduce later, the total ignorance is not a uniform distribution)!

3.1.2 Alternative derivation

Rietsch gave a fairly abstract derivation to define the entropy function, based on three conditions. The following derivation may not be better mathematically, but may be better to understand. We start with the discrete case, where we measure samples x_j with probabilities P_j. After they are measured, the total probability of measuring all those samples, is given by the likelihood function

$$G(p, p) = \prod_j P(x_j) \tag{1}$$

which is the standard way of requiring that each condition $P(x_j)$ is satisfied. Note that some x_j have identical values (i.e. duplicate samples); the relative number of identical samples with a value ξ is given by $NP(\xi)$. If we take equation 1, we can define a new multiplication over non-identical samples:

$$G(p, p) = \prod_k P(k\Delta x)^{NP(k\Delta x)} \tag{2}$$

This measure takes the form of a product, which is nasty to handle analytically. We don't lose any information if we recast this in a new form, by taking the negative log (because it is a single-valued function; so we can always transform back, using equation 4):

$$H(p, p) = -c \log(G) = -c \sum_k P(k\Delta x) \log P(k\Delta x) \tag{3}$$

$$G(p, p) = e^{-H} \tag{4}$$

Measure H is called the entropy, or the log-likelihood. The minus sign compensates for the fact that the logarithm produces always a negative value, since $0 < P < 1$. Also note that this measure is determined up to some constant c, which reflects the (arbitrary) dependency of this measure upon the number of data.

If we want to measure the fit of another distribution $Q(x)$ to a set of data which take values x with probabilities $P(x)$, then we use:

$$G(p,q) = \prod_j Q(x_j) \tag{5}$$

which leads to expressions analogous to the equations 2 and 3 derived before:

$$G(p,q) = \prod_k Q(k\Delta x)^{NP(k\Delta x)} \tag{6}$$

$$H(p,q) = -c \sum_k P(k\Delta x) \log Q(k\Delta x) \tag{7}$$

Summarizing we can say that the first case (equation 3) gives a measure of the uncertainty in a particular set of realizations, i.e. how likely they are reproduced. The second case (equation 7) gives a measure of how certain the samples of P could be reproduced by another distribution Q.

By defining an entropy we have gained a measure for the information content of distributions $P(x)$ and $Q(x)$, in the form of a number.

3.1.3 Normalization

The maximum entropy method can be extended to the case of continuous functions. I repeat the discussion by Rietsch (1977) here, slightly modified to treat the more general case of $H(p,q)$. Let x be the value of some random variable X, and let $p(x)dx$ be the probability that a value of X be in the interval $[x, x+dx]$.

We assume a general subdivision of the axis of X into segments $[x_{j-1}, x_j]$ of width Δx_j. We define p_j such that $p_j \Delta x_j$ gives the probability that x is such a segment, i.e.

$$p_j = \frac{1}{\Delta x_j} \int_{x_{j-1}}^{x_j} p(x)dx$$

and a similar equation also holds for q_j. The entropy of this discretized probability density function is:

$$H(p,q) = -\sum_j p_j \Delta x_j \log(q_j \Delta x_j)$$

which is split into:

$$H(p,q) = -\sum_j p_j \Delta x_j \log(q_j) - \sum_j p_j \Delta x_j \log(\Delta x_j)$$

Rietsch defines weights w_j such that

$$\Delta x_j = \frac{\delta}{w_j}$$

Then the log in the last term becomes:

$$-\sum_j \log(\Delta x_j) = \sum_j \log(w_j) + \log N$$

Note that the Δx_j outside the log is kept. It is needed, not δ, when the limit to 0 is taken because the parameter values are not equidistant. When we take the limit such that $\Delta x_j \to 0, j = 1, .., N$ then the continuous case becomes:

$$H(p,q) = -\int p(x) \log \frac{q(x)}{w(x)} dx + \log \infty$$

The divergent term is in fact a constant, and does not depend on the parameterization of x. Hence it can be ignored. Thus, the entropy of a continuous pdf depends not only on $p(x)$, but also on $w(x)$, an invariant measure function proportional to the varying density of the x_j in the limiting case [Rietsch, 1977].

Actually, in a true inverse problem, where we try to match a density function q to a density function p, the weights are not needed; they cancel out. This is shown in a different note of mine.

3.2 Conjunction

Now there is a different approach to describing the content of information after two pieces are combined.

3.2.1 Tarantola:

We start with an overview of T&V 1982 article. For clarity we leave some parts out. The starting point is the requirement that we need a measure of A, namely

$$p(A) = \int_A f(x) dx$$

where $f(x)$ is a density function, p is the probability of A. We want to the measure $p(A)$ to be invariant.

T&V state the following problem: if we receive two pieces of information on our system, represented by the density functions $f_1(x)$ and $f_1(x)$, how do we combine f_1 and f_2 to obtain a d.f. $f(x)$ rerpesenting the final state of information? For that they use the conjunction of two propositions $p_1 \wedge p_2$ (p_1 and p_2). The conjunction of two states satisfies the following two conditions (and some others which are not important here):

a) The first thing is that, under the trivial assumption that f_1 and f_2 are always ¿ 0:

$$\int_A f_1(x) dx = 0 \quad or \quad \int_A f_1(x) dx = 0 \iff \int_A f_1(x) \wedge f_2(x) dx = 0$$

This condition must hold for any A. Then necessarily we have:

$$\sigma(x) = f_1(x) \wedge f_2(x) = f_1(x)\ f_2(x)\ \Phi(x)$$

where $\Phi(x)$ is any function.

b) To define $\Phi(x)$ we use the requirement that two probabilities are equal, after a coordinate transformation with respect to some reference space (or physical framework) y:

$$\int_{A(x)} f_1(x) f_2(x) \Phi(x) |\frac{\partial y}{\partial x}| dx = \int_{A(x')} f_1(x') f_2(x') \Phi(x') \left| \frac{\partial y}{\partial x'} \right| dx' \qquad (8)$$

where $A(x)$ and $A(x')$ refer to related areas in the different coordinate spaces before and after the transformation (see prev. pict). Up to now, the function $\Phi(x)$ can be anything. Suppose we require the functional shape σ of the joint distribution to be identical for all coordinate systems:

$$\int_{A(x)} \sigma(x) dx = \int_{A(x')} \sigma(x') dx \qquad (9)$$

A function that satisfies $\sigma(x) = \sigma(x')$ and hence also equation 9 is called invariant. This requirement is satisfied for the following definition of $\Phi(x)$:

$$\Phi(x) = \Phi(y(x)) \left| \frac{\partial x}{\partial y(x)} \right|$$

because then 8 becomes:

$$\int_{A(x)} f_1(x) f_2(x) \Phi(y(x)) dx = \int_{A(x')} f_1(x') f_2(x') \Phi(y(x')) dx'$$

which is an invariant form.

Actually, we are kidding ourselves a bit, because all the dependency on the coordinate system is still present in the choice of y. The null-information $\mu(x)$ is defined as follows:

$$\mu(x) \equiv \frac{1}{\Phi(x)}$$

and therefore

$$\mu(x) = \mu(y) |\frac{\partial y}{\partial x}|$$

The null-information $\mu(x)$ (and also $\Phi(x)$ of course) represent the density function, imposed on the solution, due to a particular choice of coordinate system, with respect to a reference system. In a later section I give some examples of null-distributions in certain physical frame-works. We must choose the correct reference system y, where the null-distribution $\mu(y)$ is known. For example in an unbounded space we have $\mu(x) = 1$ and in a half-space we have $\mu(x) = x^{-1}$, as is shown in a different note of mine (and it was also mentioned by Tarantola, who read it in Jeffreys).

3.2.2 alternative derivation

Now I follow the same reasoning as Rietsch to derive the conjunction measure of two states. We define the measure in the form of the summation of the probabilities of $p(x)$ and $q(x)$.

$$H(p,q) = \sum_j p_j \Delta x_j \; q_j \Delta x_j$$

Following the notation used above, we write:

$$\Delta x_j = \frac{\delta}{w_j}$$

and hence

$$H(p,q) = \sum_j p_j \Delta x_j \; q_j \frac{\delta}{w_j}$$

$$H(p,q) = \delta \sum_j \frac{p_j \; q_j}{w_j} \Delta x_j$$

We keep the terms Δx_j which define the infinitesimal distances between the non-uniformly spaced function samples. Take the limit $\Delta x_j \to 0, j = 1, .., N$, so that:

$$H(p,q) = \delta \int \frac{p(x)q(x)}{w(x)} dx$$

where $\delta \to 0$ is a constant and can be ignored.

The density function inside the integral sign defines the conjunction of p and q:

$$\sigma(x) = p(x) \wedge q(x) \equiv \frac{p(x)q(x)}{w(x)}$$

where the $w(x)$ can be identified as the null-distribution. Note that $\sigma(x)$ can only be used within an integral.

4 Meaning

We give a short discussion of the meaning of the entropy and the conjunction in a probabilistic sense. I will show that the entropy is a measure of the AND operation; the conjunction is a measure of the OR operation.

To fix ideas, suppose we throw a dice. The probability of throwing one number between 1 and 6 is $\frac{1}{6}$. The chances of throwing first $x_1 = 1$, then $x_2 = 2$ and then $x_3 = 3$ is $\frac{1}{6}\frac{1}{6}\frac{1}{6} = \frac{1}{216}$. This is the way that the entropy (without the log) is measured: without the log, it gives the chance of obtaining the values of a set A, in a specific order.

Suppose we do another experiment with the dice. Now we only want to predict that the value x that we will throw is either a 1, a 2 or a 3. The probability is $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$. Note that the chance of obaining any number

between 1 and 6 in a throw is 1. This is the kind of information obtained from the conjunction: it gives the chance of obtaining any value from a specified set A.

When written down in terms of probabilities $p(x_i)$ the entropy method gives:

$$R = \prod_{i=1}^{N} p(x_i)$$

the chance that a set of values $\{x_i\}$ is drawn from distribution $p(x)$ in a specific order. The conjunction gives:

$$R = \sum_{i=1}^{N} p(x_i)$$

the chance that any one x_i (you don't care which one) is drawn from the set of values $\{x_i\}$.

This reasoning shows that the entropy gives a much stronger measure of the uncertainty than the conjunction. But when it comes to combining information, the entropy is very unstable. Suppose we draw a set of samples $\{x_i\}$ from distribution $p(x)$. The measure of information of $q(x)$ given this set of observations is:

$$R = \prod_{i=1}^{N} q(x_i)$$

Since samples x_i are chosen from a different distribution, it is possible that if $p \neq q$, for some samples the distribution q gives $q(x_i) = 0$. In that case, the product also becomes zero. There is no such problem is $p = q$, since then such samples do not exist. This shows that the entropy measure is very sensitive to outliers (with respect to q) in the data (obtained from p). On the other hand, the conjunction is a summation. If for one sample x_i we have a term that is zero, the sum is not affected. The conjunction is very stable and when it comes to combining different pieces of information, it is preferred over the entropy (at least in the geophysics community: in the mathematics community the entropy measure is more widely used, as far as I can tell; the term 'conjunction' doesn't even exist in the standard math handbooks, neither does the term null-distribution).

5 Example.

I will illustrate the use of the entropy method with an example, where we have a data set of N travel times of a sound wave t_i measured at different distances from the source r_i . We have a theory which relates the distances to the travel times using some parameter m:

$$t = G(r, m)$$

We assume that the measurements of travel times contain noise, which is distributed in the form of a Gaussian:

$$\theta(t, r, m, s) = \frac{1}{s} \exp\left(\frac{-(t - G(r, m))^2}{2s^2}\right)$$

with s some unknown positive constant describing the spread of the data. Note that it is *very important* that the scaling by $\frac{1}{s}$ is present! Now we measure the information content of θ given this data set of N measurements:

$$H(m, s) = -\frac{1}{N} \sum_{i=1}^{N} \log \theta(d_i, r_i, m) \tag{10}$$

When we substitute the assumed theory we can express the entropy as a function of the model parameter:

$$H(m, s) = \frac{1}{N} \sum_{i=1}^{N} \log s + \frac{1}{N} \sum_{i=1}^{N} (d_i - G(r_i, m))^2 / 2s^2 \tag{11}$$

Assume for the moment that G is a linear relation and that the model parameter is given by the slowness of air p:

$$G(p) = pr \tag{12}$$

Then this reduces to a linear regression problem. If we substitute 12 in 11 we get:

$$H(m, s) = \frac{1}{N} \sum_{i=1}^{N} \log s + \frac{1}{N} \sum_{i=1}^{N} (t_i - pr_i)^2 / 2s^2$$

When we maximize this measure, we obtain the maximum likelihood model m. The optimal model is found by:

$$\frac{d}{dp} H(p, s) = \sum_{i=1}^{N} r_i (t_i - pr_i) / s^2 = 0$$

$$p = \frac{\sum r_i t_i}{\sum r_i} \tag{13}$$

We can consider s to be some model parameter which describes the level of Gaussian noise in the observed travel times. The optimal width of the distribution to fit the data is:

$$\frac{d}{ds} H(p, s) = \frac{N}{s} - \frac{1}{s^3} \sum_{i=1}^{N} (t_i - pr_i)^2 = 0$$

$$s^2 = \frac{1}{N} \sum_{i=1}^{N} (t_i - pr_i)^2 \tag{14}$$

where p is given by equation 13. Equation 14 which is the well-known definition for the standard deviation.

What happens if we try to fit a linear relation to the travel time measurements, using the conjunction method? The conjunction of the data and the theoretical distribution is:

$$\sigma(m,s) = \frac{1}{N}\sum_{i=1}^{N}\frac{\theta(t_i,r_i,m,s)}{\mu_t(t_i)\mu_r(r_i)}$$

What happens if we use bayesian inversion of the observed data distribution? This is doing a real inverse problem, by applying the inverse relation directly to the data. In that case, the solution is given by:

$$S(m,s) = \frac{1}{N}\sum_{i=1}^{N}\theta^{-1}(m,s|t_i,r_i)$$

6 Appendix: invariance

If a function $p(x)$ is invariant then after a coordinate transformation there exists a function $g(x')$ related to $p(x)$ by:

$$g(x') = p(x)\left|\frac{\partial x}{\partial x'}\right|$$

and in the case of a multivariate function (which depends on a number of parameters) this becomes

$$g(x') = p(x)\left|\frac{\partial x}{\partial x'}\right| = p(x)\begin{vmatrix} \frac{\partial x_1}{\partial x'_1} & \frac{\partial x_2}{\partial x'_1} & .. \\ \frac{\partial x_2}{\partial x'_1} & \frac{\partial x_2}{\partial x'_2} & .. \\ .. & .. & .. \end{vmatrix}$$

where $J = \left|\frac{\partial x}{\partial x'}\right|$ is the Jacobian: this scaling function is used in calculus to represent a change in unit surface when it is mapped from one parameter to a transformed parameter. For example we could describe a point either as (x,y) or as $(r\cos\theta, r\sin\theta)$.

The Jacobian becomes:

$$J = r\cos\theta$$

Note that in both coordinate systems we describe the same physical point in space. We don't lose any information, but we do describe it in different ways. This will also affect probabilistic description. Due to such a coordinate transformation, we get:

$$p(A) = \int_{A(x,y)} f(x,y)dxdy = \int_{A(r,\theta)} f(r,\theta)r\cos\theta drd\theta$$

7 References:

References

[Tarantola and Valette, 1982] Tarantola A. and Valette B., 1982, Inverse problems = Quest for information, J.Geophys., 50, 159-170.

[Tarantola, 1988] Tarantola A., 1988, Probabilistic foundations of inverse theory, Les Houches, Session L.

[Devilee, 1998] Devilee R.J.R., 1998, The null-distribution of a half-space, personal note.

[Papoulis, 1991] Papoulis A., Probability, random variables and stochastic processes, McGraw Hill, Singapore, third edition.

[Pearlmutter and Parra,] Pearlmutter and Parra, ... , ...

[Rietsch, 1977] Rietsch E., 1977, The maximum entropy approach to inverse problems, J.Geophys., 42, 489-506.