Société de Calcul Mathématique, S. A.
Algorithmes et Optimisation

The Experimental Probabilistic Hypersurface

a new concept in order to represent and interpret the results of the measures in a physical experiment or in a computer simulation

by Bernard Beauzamy

Siège social et bureaux : 111, Faubourg Saint Honoré, 75008 Paris. Tel. : 01 42 89 10 89. Fax : 01 42 89 10 69
Société Anonyme au capital de 46 200 Euros. RCS : Paris B 399 991 041. SIRET : 399 991 041 00035. APE : 731Z

Abstract

We introduce a new concept, called « Experimental Probabilistic Hypersurface » (in short EPH). It allows us to represent the information obtained from any number of measures, in a physical experiment or in a computational code. This information is stored as a density of probability, « above » each point in the configuration space. If the experiment depends upon K parameters $(x_1,\ldots,x_K)$, the EPH consists in the collection of the density functions $f(t\,;x_1,...,x_K)$.

These densities are built from the existing information (the measures that have already been made). The existing information « propagates » all over the space, with the following rule : the entropy should always be maximal. The principle of maximal entropy thus governs the whole construction, which allows us a construction with no artificial rules or probability laws. If you are close to a place where the experiment has been performed, the density will be more concentrated ; if you are far away, the density will be less concentrated, because you know less.

The applications are multiple. The EPH is a « storage » of information, which grows and becomes more precise when more and more experiments are performed. It allows you to get immediately « local » results : which regions or points are dangerous, which are safe, and so on. The EPH is intended to replace both the deterministic methods (for instance interpolation between existing values), which are artificial, and the statistical methods, which are only global. The EPH gives local results, but still keeps the global characteristics.

The EPH was initially constructed in order to meet a request from Framatome-ANP (*Commande Framatome-ANP no 12A/1003006066 / 17.12.2003 / PSC),* concerning the results given by a computational code, named « CATHARE ». This code computes a temperature in a nuclear reactor, in case of a breach. What conclusions can be drawn from a limited number of computations (the number is limited because the computation takes several hours), when the number of parameters is quite high (around 50) ? We present here the theory in its full extend.

First Part

Introduction and description of the need

Any physical experiment usually brings many measures, which may concern various quantities, at a given time or over an interval of time. The results themselves depend upon various parameters, and the choice of these parameters is usually unclear : should we take into account the atmospheric temperature ? the speed of wind ? personal characteristics linked with the experimenter ? If we think a little bit, we see that the list may grow indefinitely, and we usually do not know which parameters are mostly influential.

But conversely, the number of data that are collected is usually extremely small, no matter what topic we are in. This comes from the fact that the experiment is always delicate : the preparation takes time, it requires human and technical resources. In some cases, the experiment may be dangerous, or the favourable circumstances are rare. In all cases, it is costly. For instance,

- To fly a missile, in order to study the trajectography, is both dangerous and quite costly ;

- In ecotoxicology, the toxicity experiments are usually made upon a small number of species, over a small number of days : this comes from the fact that only few species are available for laboratory work. One wishes to « extrapolate » the results, to a larger number of species and to longer periods (see our work [Ecotox]).

But we immediately come to a fundamental difficulty, which is of mathematical nature, linked to the size of the configuration space. Let's imagine, for the sake of simplicity, that we take into account $K = 50$ parameters, and that each of them may take 10 values (this is obviously oversimplified : in general, the parameters vary continuously and may take infinitely many values). So we get 10^{50} possible values, for all possible configurations. Even if the experiment is purely on a computer (for instance a computational code), even if the computer performs a billion operations per second, we will need more than 3.10^{33} years to perform it, and we see that a few more GHz change nothing !

So we see that one can never explore completely the space of all configurations, and even not explore it in a significant manner. The amount of information we gather can be only infinitesimal, compared to the global exploration. So we are led to this fundamental question : how can we exploit the information we got, and, since it is so rare, how can we exploit it at best ?

I. Two methods : interpolation or statistics

We may distinguish globally between two methods, or, more exactly, between two « ways of thought ». The first one is totally deterministic : one tries to alleviate the missing data, using the existing one, for instance using linear interpolation. If we succeed, then we have data are every point of the configuration space.

The problem is that these data are artificial : they depend completely upon the method of interpolation or extrapolation that has been chosend. Another difficulty, as we see later, is that the methods used for such extensions are quite inefficient when the dimension is high (as it is always the case for real-life problems).

The second approach consists in a global, statistical, information. If we perform $N = 300$ measures, we may take the average, the standard deviation, and so on. All this is quite useful, but does not answer the question : if I perform a new experience, with new values $x_1, \ldots, x_K$ of the parameters (a configuration which has never been met), what may I expect as a result of the experiment ?

II. The Experimental Probabilistic Hypersurface

The concept we introduce here, « Experimental Probabilistic Hypersurface », precisely answers this question. If N experiments have already been performed, upon N configurations, the Hypersurface gives a result, as a density of probability : for any new configuration $x_1, \ldots, x_K$, here is the density of probability of the expected result.

If the configuration has already been tested, the result is a certainty, and the density is a Dirac measure. Othewise, it is a true density, and this density is less and less concentrated if the points where the experiment has already been performed are further. In other words, if you are interested in a set of parameters close to another where the experiment has already been performed, the Hypersurface gives you a very concentrated density, and if you consider a configuration far from the known ones, the density is quite « flat ».

So, what we get is not a deterministic result, as was the case with the interpolation methods, but a probabilistic result. This result is local, in the sense that this density is given at every point of the configuration space : this differs from global statistics.

The Experimental Probabilistic Hypersurface is constructed using, in a very strict manner, the principle of minimal information, that is, the principle of maximum entropy. Using this principle, we show how each measure « propagates » an information over the whole configuration space. Indeed, the concept of entropy will be the key-concept for our construction. It allows us to ensure that we do not introduce (as one sees quite often in probabilistic approaches) unjustified or arbitrary laws.

III. Applications

The applications of the hypersurface are numerous, since it is intended to replace both deterministic and statistical methods. It keeps, obviously, the global statistical information (means, standard deviation, and so on), but it allows a local treatment : we may observe the probability density above each point, see how it changes from one point to the next, find the regions in which it is concentrated near high values, or near low values, in which the variance is high, or is small, and so on.

We may decide to have an exploitation which is different in each region. For instance, we may find what are the dangerous regions, in the space of configurations, in order to concentrate the future experiments on these zones.

The Hypersurface meets a need connected with the representation of the information that has been obtained, no matter how it has been obtained. The measure points may have been chosen at random, or in a deterministic way.

IV. The dangers of a random exploration

Indeed, it may be tempting, in order to explore the space of configurations, to rely solely upon chance. Each parameter will be affected precise values, randomly chosen (experimental design), and the set of results is supposed to represent a good exploration of the configuration space.

Indeed, if we divide the configuration space into 20 « boxes », each of them of probability 0.05, and if we perform 300 random samples, independently, the probability that one of the boxes should be ignored is only $0.95^{300} = 0.2 \cdot 10^{-6}$; in other words, we are quite certain that all boxes have been penetrated. If, for instance, the result of the experiment is a temperature, we are quite sure that the highest observed temperature, in the 300 experiments, is among the 5 % highest of all possible temperatures.

This information is interesting in itself ; however, it depends in an essential manner on the probability laws which have been introduced on each of the parameters. But these laws are necessarily artificial, since, by definition ! very little is known about the whole experiment. If we assume these laws to be quite concentrated, or conversely quite flat, we modify the final probability. This is explained quite in detail in our book [BB2].

The Experimental Probabilistic Hypersurface does not answer this difficulty, but it turns around. It will allow us to represent the information, no matter how it has been obtained. Perhaps, our sampling was made using absurd laws : no problem, still some results have been obtained, and these result bring some information, which will be incorporated into the EPH, which can be built. If, later, we want to change the laws upon the parameters, we can easily do so. To represent the information and to treat it are two different things.

Second Part

The concept of

Experimental Probabilistic Hypersurface

The result of an experiment may be described, mathematically speaking, as a function $CT(x_1,...,x_K)$; the variables $x_1,\ldots,x_K$ are the « input parameters ». Here, K is the number of parameters (around 50). We assume that the result is a real number (scalar result). Of course, for a given experiment, we may measure several physical data and obtain a result which is a vector. Then, we have as many scalar functions. In other words, for us, the function CT is real-valued. The notation CT was chosen as a reference to the question posed by Framatome-ANP, where the experiment was in fact a computational code, called « Cathare », and the output is a temperature.

I. Global Hypercube

Each parameter, for physical reasons, is set to vary in a closed bounded interval : $x_k \in [a_k, b_k]$, $k = 1,...,K$. The cartesian product :

$$GH = \prod_{k=1}^{K} [a_k, b_k] \tag{1.1}$$

is therefore an hypercube in the space R^{K}. We call it GH (« Global Hypercube »).

II. Reduced Hypercube

In order to compare the influence of the various parameters, we need to normalize the variation intervals, and bring all of them to be $[0,1]$. This is obtained in a very simple manner : if a parameter x varies in an interval $[a,b]$, the parameter $x-a$ varies in $[0,b-a]$ and the parameter $\dfrac{x-a}{b-a}$ in $[0,1]$.

We denote by $GRH = [0,1]^K$ the global reduced hypercube.

Let us observe that, in the present situation, any point of GH is admissible in the space of parameters. This is the case for the code « Cathare », but is not necessarily true in general. Indeed, we may imagine that, for some computational codes, for some physical experiments, some parameters have restrictions upon their values, for instance some sub-interval may be excluded.

In short, in the present situation, the space of parameters is a closed bounded hypercube in R^{K}. In order to simplify our notation, we set $X = (x_1,...,x_K)$. This is a point in R^{K}.

III. Sampling

We choose (in an arbitrary manner) some values for the K parameters $x_1,...,x_K$ (it may be random sampling, or deterministic choices). We denote by $X_n = (x_1^{(n)},...,x_K^{(n)})$ the result of the le résultat n-th choice, for $n = 1,...,N$ (N is therefore the number of choices). So we get N points in the space R^{K}. Of course, N is large compared to K :

the number of measures is larger than the number of parameters. In our work with Framatome-ANP, N was of the order of 300.

IV. Constructing the deterministic hypersurface associated with a function of K variables

If we knew the value of $T = CT(x_1,...,x_K)$ for all values of the parameters $x_1,...,x_K$, we would have an hypersurface in the space $\mathbb{R}^{K+1}$. For instance, for $K=1$, the equation $T = CT(x_1)$ determines a curve in the space $\mathbb{R}^2$; for $K=2$, the equation $T = CT(x_1, x_2)$ determines a surface in the space $\mathbb{R}^3$. For $K>2$, we have an hypersurface, which divides the space into two regions, situated on each side of the hypersurface : the region where $T < CT(x_1,...,x_K)$ and the where $T > CT(x_1,...,x_K)$, in the space $\mathbb{R}^{K+1}$. This notion of (deterministic) hypersurface is quite customary ; for instance, it is used in order to represent the « graph » of a function of K variables in a space of dimension $K+1$; as we explained earlier, it generalizes the notion of a curve (dimension 1) and surface (dimension 2).

V. Defining the Probabilistic Hypersurface associated with a sampling

We study K parameters (around 50) and we perform N measures (around 300), for which the complete computation of $CT(x_1,...,x_K)$ is done, or the physical experiment, if we deal with a real physical experiment. We use the word « sampling » to refer to the successive choices of parameters, but, quite clearly, these choices may not be at random.

In order to build the Probabilistic Hypersurface, the idea is as follows : we consider that the result of the computation of $CT(x_1,...,x_K)$ is unknown, except if it has been perform precisely for these values of $x_1,...,x_K$. When it is known, it is perfectly deterministic : a computational code does not make mistakes, and, if we deal with a physical experiment, we neglect the errors (the errors linked with the experiments might be taken into account, replacing the Dirac measures by other densities ; see our book [BB2]).

When the result is unknown, we will consider that it is given by a probability law ; this law will be more precise if the computations have been made for a point close to the considered point.

In other words, for any value of $x_1,...,x_K$, when N measures have been made, we have a density of probability :

$$f(t\,;x_1,...,x_K \mid X_1,...,X_N) \tag{5.1}$$

which depends on the place where we are $(x_1,...,x_K)$ and on the previous measures : $(X_1,...,X_N)$. Recall that each X_j stands for $(x_1^{(j)},...,x_N^{(j)})$.

The vertical bar | in the definition of f reminds us that this is a conditional probability: we are looking for a density, knowing that the measures at points $X_1,...,X_N$ have been performed.

This density will be more concentrated, if the point $x_1,...,x_K$ where we are is closer to a point where the measure has been made. If we are exactly at a point where a previous computation has been made, the density is a Dirac mass : it is completely deterministic.

The collection of all densities $f(t\,;x_1,...,x_K\,;X_1,...,X_N)$ will be the « Experimental Probabilistic Hypersurface » (in short EPH) Probabiliste » describing the experiment (or the computational code). We may view it, not as a thin surface, but as a thick surface.

Let's see an example, in the case of one parameter x_1. Assume that two computations have been made, at the values x_1' and x_1''. We might have the following picture :

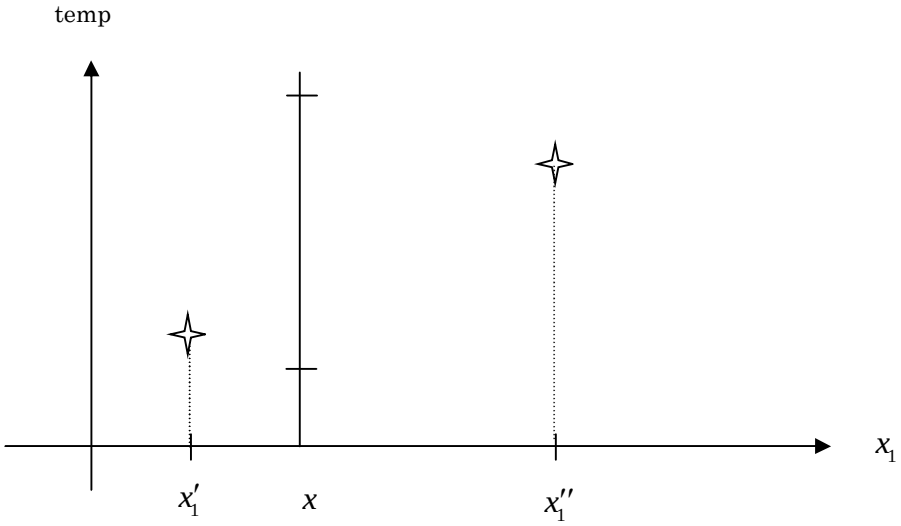

Figure 1 : an example of conditional density

Above x_1' and x_1'', we have precise values, since the computation has been made. Above any point x, we do not have a precise value, but a range of possible values, and a probability density, inside that range. This density takes into account the proximity of x_1' and x_1'' with respect to x.

If $T_1' = CT(x_1')$ and $T_1'' = CT(x_1'')$ are the results of the two computations that were made, the density of probability above x might look like this :

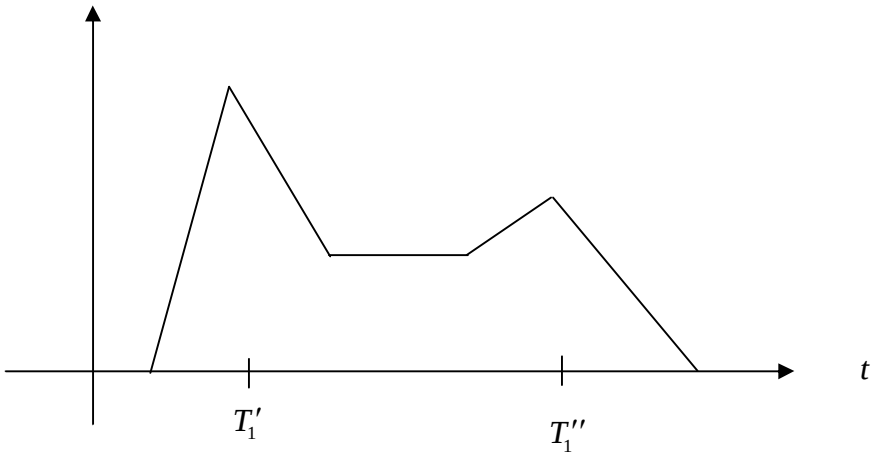

Figure 2 : shape of the conditional density above the point x

In our example, the point x is closer to x_1' than to x_1''. So we may assume that the triangle above x_1' is « sharper » than the triangle above x_1'' : all this will be explained later on.

VI. Why not a deterministic hypersurface ?

As we explained in the introduction, the most natural idea, a priori, is not that of a probabilistic hypersurface, but rather a deterministic one, the following way : we interpolate between all the measure points (« Krigeage » methods). If for example the space of parameters had dimension 1, we would interpolate linearly between measure points. If it had dimension 2, we would use small triangles and we would construct a plane surface above each triangle. This method can be generalized to arbitrary dimension : we need « hyper-triangles », made of $K+1$ points in a space of dimension K (we still speak of « triangles »).

This idea, though it is the most natural one, must still be discarded, for three reasons :

- We would have to assume that our computational code, above each triangle, acts linearly. But this is wrong. Our code does not need to be linear, even on small regions.

- We cannot accept the idea to attribute to the computational code a precise, unknown, value : the idea to have a probability, reflecting the uncertainty, is much more satisfactory. We can measure the standard deviation, which characterizes the uncertainty.

- Building the triangles can be done in many ways, none is imposed. Depending on the method which is chosen, the result will be different. There is a particular triangulation, named « Delaunay Triangulation », for which the triangles are not too long and narrow, and this triangulation can be performed in any dimension. But why should we choose these triangles ? There is no physical reason to such a choice.

Computer implementation of Delaunay Triangulation is extremely costly, as far as computation time is concerned. We launched an international consultation on the net « NA-Net » (numerical analysis net), in order to obtain information about the most recent performances in terms of triangulation algorithms. One can perform very quickly the triangulation of thousands of points in dimension 2, one hundred in dimension up to 10, but 300 points in a space of dimension 50 is totally out of reach. The theoretical complexity is in $0(n^{d/2})$, where n is the number of points (here 300) and d is the dimension (here 50) : we find something like 300^{25} operations ! If the computer performs a billion operations per second, we need 2.10^{45} years.

One may consult the paper « *Delaunay Triangulation in Higher Dimensional Spaces* », by Kurt Melhorn and Michaël Seel, Max Planck Institut für Informatik, 1998, from which we extract the following table. It concerns the triangulation of 100 points, as a function of the dimension :

dimension	time (seconds)
2	0.53
3	2.65
4	15.7
5	76
6	387
7	out of memory

Third Part

Constructing the Experimental Probabilistic Hypersurface

I. Before the first measure

Before we start any measure, we may justifiably consider that all results will be in a closed bounded interval $[T_{\min}, T_{\max}]$. Indeed, no physical measure goes to $-\infty$ nor to $+\infty$. But this interval should be viewed as a security interval : if it is replaced by a larger interval, the security will of course be better.

Since we know nothing about the result of the first measure, we may consider that this result follows a uniform distribution upon this interval : nothing, indeed, allows us to privilege any subset of this interval. But of course, if some information is available about what the first measure will be, it can be used at this early stage.

If we want to study a critical interval, of the type $[T_{seuil}, +\infty[$ (temperatures that are larger than a certain threshold), we need obviously $T_{\max} > T_{seuil}$ (otherwise the problem does not exist!). If we increase $T_{\max}$, we increase the probability of the interval $[T_{Threshold}, T_{\max}[$: in other words, we penalize ourselves. So, conceptually speaking, there is no objection in chosing $T_{\max}$ quite large.

But conversely, if we diminish $T_{\min}$ in an artificial manner, this implies that we increase the probability of the interval $[T_{\min}, T_{Threshold}]$, and so we decrease that of the critical interval $[T_{Threshold}, T_{\max}[$. This is forbidden.

So the rule is as follows : we should choose as global interval for our study $[T_{\min}, T_{\max}]$ the smallest interval such that we are sure that it will contain all measures that will come. This information comes from the physics of the problem, and also from all previous similar situations, if any.

So we observe this somewhat surprising fact : if we are interested only in large temperatures, the choice of $T_{\max}$ is not critical, but the choice of $T_{\min}$ is critical !

Before the first measure, the hypersurface already exists : above each point X, we have a uniform density upon the interval $[T_{\min}, T_{\max}]$. This first hypersurface will progressively get modified, in order to incorporate the information coming from each measure.

II. After the first measure

The first measure is made at a point X_1 and gives a value $T_1 = CT(X_1)$. We now want to evaluate the resulting information.

Quite clearly, the total available information has increase, since a measure has been made (except, of course, if it contradicts the initial interval !). In fact, we have at our disposal two different types of information :

– The probability density above a point X is more or less similar to the density constructed above the point X_1 ; we may legitimately assume that it is more similar when X is closer to X_1.

– The probability density above a point X is entirely supported by the interval $[T_{\min}, T_{\max}]$.

Since the measure at point X_1 gave the precise value $T_1 = CT(X_1)$, we have above the point X_1 a density which is a Dirac mass : δ_{T_1}. If we took the errors of measure into account (which is not the case here), we would put here, instead of a Dirac mass, a density of probability which would be obtained from the calibration of the device used for the measure : see our book [BB2].

Now, if we want to make these remarks precise, we need to use the notion of information connected with a law of probability : this is the concept of entropy, which we now present.

III. Information connected with a law of probability : the entropy

There are two kinds of entropies : discrete and continuous.

A. Discrete Entropy

If we start with a discrete law of probability (p_i), the associated entropy is defined by the formula :

$$I = -\sum_i p_i \text{Log } p_i \qquad (3.a.1)$$

This entropy is obviously positive (since $p_i \leq 1$) ; it is equal to 0 if and only if the distribution is concentrated in a single point (all p_i's are 0, except one, equal to 1) : this is a Dirac mass.

In the case of equirepartition of N points over an interval $[T_{\min}, T_{\max}]$, all having same probability $1/N$, the entropy is :

$$I = \text{Log } N = \text{Log } \frac{T_{\max} - T_{\min}}{a}, \qquad (3.a.2)$$

where a is the distance between two consecutive points (width of the subdivision).

B. Continuous Entropy

We define the entropy of a law of probability on $\mathbb{R}$ by the formula :

$$I(f) = -\int_{-\infty}^{+\infty} f(t) \text{ Log } f(t)\, dt. \qquad (3.b.1)$$

The larger this entropy is, the less concentrated the law is. We now treat two examples, in order to illustrate this definition :

– Entropy of a uniform distribution

Let $f(t) = \dfrac{1}{T_{\max} - T_{\min}}$ if $T_{\min} \le t \le T_{\max}$, 0 otherwise. Then :

$$I(f) = \mathrm{Log}\,(T_{\max} - T_{\min}) \tag{3.b.2}$$

When this interval shrinks, this entropy decreases (the law is more and more concentrated). If the interval is reduced to a single point (case of the Dirac mass), its value is $-\infty$. The continuous entropy, unlike the discrete one, takes its values between $-\infty$ and $+\infty$.

– Entropy of a gaussian variable :

Let us consider the gaussian density :

$$h_\sigma(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-t^2 / 2\sigma^2),$$

then :

$$\begin{aligned} I(h_\sigma) &= -\int_{-\infty}^{+\infty} h_\sigma(t)\,\mathrm{Log}\,h_\sigma(t)\,dt \\ &= \frac{1}{2\sigma^2}\int_{-\infty}^{+\infty} t^2 h_\sigma(t)\,dt + \mathrm{Log}\left(\sigma\sqrt{2\pi}\right)\int_{-\infty}^{+\infty} h_\sigma(t)\,dt \\ &= \frac{1}{2} + \mathrm{Log}\left(\sigma\sqrt{2\pi}\right) \end{aligned}$$

and therefore :

$$I(h_\sigma) = \mathrm{Log}\left(\sigma\sqrt{2\pi e}\right) \tag{3.b.3}$$

Here again, of course, the entropy increases with σ ; when $\sigma \to 0$, $I(h_\sigma) \to -\infty$.

The entropy is, as we can see, a measure of the quantity of information : the larger the entropy is, the less significant is the information. We will respect this principle of maximal entropy when we construct the Hypersurface : to say that the entropy is maximal means that we do not add any information. This is the point of view we adopted already when we said that, before any measure, a uniform density was taken above all points.

IV. Constructing the Hypersurface in the case of a single measure

Let us assume now that a single measure has been performed. We will construct completely the Hypersurface in that case. We start with the discrete case, which is the only one that has a physical meaning (an observation is always of discrete nature ; see our book [BB2]). The continuous version is interesting, however, since it provides useful tools, to which we come later.

Let us start with the case of a single parameter, denoted by x. The measure $CT(x)$ was made for $x = 0$ (we may of course bring back ourselves to this case) ; it gave $CT(0) = T_0$.

The values of the parameter are discretized under the form :

$$x_i = \frac{id}{K},\ i = 0,...,K\ . \tag{4.1}$$

For $i = 0$, we have $x_0 = 0$: this is the point where the measure was made. For $i = K$, we have $x_K = d$: this is the maximal distance between any point of the domain and the point where the measure was made. So the discretization was made with a width equal to $\varepsilon = \frac{d}{K}$. We may also have points x on the negative side of the axis : we do not care about that here.

Concerning the temperatures, we also have a discretization of the observed values, between $T_{\min}$ and $T_{\max}$, under the form :

$$t_j = T_{\min} + \frac{j}{K'}(T_{\max} - T_{\min})\ , \tag{4.2}$$

where K' is the number of points in the subdivision. The width of the subdivision, for the temperatures, will be denoted $\varepsilon' = \frac{T_{\max} - T_{\min}}{K'}$.

Let j_0 be the index of the observed temperature :

$$t_{j_0} = T_0\ .$$

In order to construct the Hypersurface, we will compute the probability, denoted by $p_{i,j}$, corresponding to each t_j above each x_i. These probabilities verify :

$$\sum_{j=0}^{K'} p_{i,j} = 1\text{, for }\ i = 0,...,K. \tag{4.3}$$

For $i = 0,...,K,$ the t_{j_0} is that of largest probability (since it is the one that was observed). The probability decreases when j gets further from j_0 (both on the right and on the left). In other words, the sequence $p_{i,j}$ is increasing for $j \le j_0$, decreasing for $j \ge j_0$.

We may legitimately consider that the sequence $p_{i,j}$ is symmetric with respect to the index j_0 (same decay both on the left and on the right). If we neglect the truncation effect due to $T_{\min}$ and $T_{\max}$ (which we can do in general), this last property will translate into the fact that the expectation is equal to T_0 :

$$\sum_{j=0}^{K'} p_{i,j}\, t_j = T_0, \text{ for } \; i = 0,...,K. \tag{4.4}$$

Finally, we may legitimately consider that the variance of the law increases when we get further from the measure point. This gives :

$$\sum_{j=0}^{K'} p_{i,j}\, t_j^2 \text{ decreases for } \; i = 0,...,K. \tag{4.5}$$

So we have enumerated all properties that the probability laws must meet, taking into account the physical constraints.

The discrete entropy I_i above a point x_i is defined by :

$$I_i = \sum_{j=0}^{K'} p_{i,j} \text{ Log } p_{i,j}. \tag{4.6}$$

It is increasing with $i = 0,...,K$. Its value is 0 at $i = 0$, since we have a precise measure : $p_{0,j} = 0$ except $p_{0,j_0} = 1$. At the other end, for $i = K$, we have the information :

$$I_K = \text{Log}\frac{T_{\max} - T_{\min}}{\varepsilon'}.$$

In order to simplify the notation, we set $A = \text{Log}\frac{T_{\max} - T_{\min}}{\varepsilon'}$. This quantity is known : the width of the subdivision is known, as well as the extreme temperatures.

We do not know the precise value for each I_i ; so we consider that they follow uniform laws, satisfying the inequalities :

$$0 = I_0 \le I_1 \le \cdots \le I_i \le \cdots \le I_K = A \tag{4.7}$$

This means that the joint law of the $K-1$-uple $I_1,...,I_{K-1}$ is proportional to the function $1_{0 \le y_1 \le \cdots \le y_{K-1} \le A}$. Then we take as an estimate for each I_i the expectation of the marginal law for each variable. This gives the formulas :

$$E(I_i) = \frac{\int_0^A \int_0^{x_{K-1}} \cdots \int_0^{x_2} x_i \, dx_1 \cdots dx_{K-2} dx_{K-1}}{\int_0^A \int_0^{x_{K-1}} \cdots \int_0^{x_2} dx_1 \cdots dx_{K-2} dx_{K-1}} \tag{4.8}$$

The denominator gives $\frac{A^{K-1}}{(K-1)!}$ and the numerator $\frac{iA^K}{K!}$. So we get :

$$E(I_i) = \frac{iA}{K} \;,\; i = 0,...,K \tag{4.9}$$

and this is the value we give now to the entropy above each x_i.

It will be convenient to present this result as a Lemma, which will be useful in the sequel :

Lemma 1 : Minimal Information Lemma

If $K-1$ *variables* z_i *satisfy :*

$$0 \le z_1 \le \cdots \le z_{K-1} \le A,$$

the minimal information is obtained when the value :

$$z_i = \frac{iA}{K}, i = 1,..., K-1.$$

is attributed to each of them.

The knowledge of the entropy is obviously not enough to compute completely the probabilities $p_{i,j}$. But we assume moreover that on each layer the variance is maximum (we penalize ourselves). In this case, the $p_{i,j}$ can be computed explicitly, as we see now. In order to simplify our notation, we omit the index i in the next paragraph : the computation is done separately on each layer.

Let us first observe that, for a fixed distribution of probability (p_j), to maximize the variance when the entropy is fixed is the same as to maximize the entropy when the variance is fixed : the solutions to both problems will have the same shape. The second problem is easier to solve :

Lemma 2. – Let p_j be a distribution of probability at the points t_j, with fixed variance. The distribution with maximal entropy is a gaussian :

$$p_j = \exp(\alpha + \beta t_j + \gamma t_j^2).$$

Proof of Lemma 2

This extremality property of gaussian variables is well-known (see for instance the book [Sobolev], chapter 10, proposition 10.2.2); we present here a discrete analogue.

Let p_j be the distribution maximizing the entropy. We have obviously :

$$\sum_j p_j = 1. \tag{4.10}$$

We may assume (modifying the points t_j if necessary) that the expectation is 0, that is :

$$\sum_j p_j t_j = 0. \tag{4.11}$$

If the variance is fixed, that means :

$$\sum_j p_j t_j^2 = \sigma^2. \tag{4.12}$$

Let now q_j be a sequence such that $(p_j + q_j)$ is also a probability, with mean 0 and same variance. This means :

$$\sum_j q_j = 0\,, \tag{4.13}$$

$$\sum_j q_j t_j = 0\,, \tag{4.14}$$

and using (4.12),

$$\sum_j q_j t_j^2 = 0\,. \tag{4.15}$$

Since p_j is extremal for the entropy,

$$-\sum_j p_j \mathrm{Log}\ p_j \geq -\sum_j (p_j + q_j)\mathrm{Log}\,(p_j + q_j)\,,$$

which can be written :

$$\sum_j p_j \mathrm{Log}\,(1+\frac{q_j}{p_j}) + \sum_j q_j \mathrm{Log}\ p_j + \sum_j q_j \mathrm{Log}\,(1+\frac{q_j}{p_j}) \geq 0\,. \tag{4.16}$$

Let λ be a real number and let us consider the sequence $\lambda\, q_j$, when $\lambda \to 0$. We get from (4.16) :

$$\sum_j p_j \mathrm{Log}\,(1+\frac{\lambda q_j}{p_j}) + \lambda \sum_j q_j \mathrm{Log}\ p_j + \lambda \sum_j q_j \mathrm{Log}\,(1+\frac{\lambda q_j}{p_j}) \geq 0\,,$$

and thus :

$$\lambda \sum_j q_j + \lambda \sum_j q_j \mathrm{Log}\ p_j + \lambda^2 \sum_j \frac{q_j^2}{p_j} \geq 0\,,$$

using (4.13) :

$$\lambda \sum_j q_j \mathrm{Log}\ p_j + \lambda^2 \sum_j \frac{q_j^2}{p_j} \geq 0\,.$$

But this must be true for any λ , positive or negative. This is possible only if :

$$\sum_j q_j \mathrm{Log}\ p_j = 0\,. \tag{4.17}$$

Let Q be the vector of the q_j's. The conditions (4.13), (4.14), (4.15) imply respectively that, in the space l^2, the vector Q is orthogonal to the constant sequence, to the sequence $X = (t_j)$, and to the sequence $X^2 = (t_j^2)$. Property (4.17) shows that, if these conditions are fulfilled, Q is orthogonal to the sequence $U = \text{Log } p_j$.

So we see that if Q is orthogonal to the vector space spanned by 1, X, X^2, then U is orthogonal to Q. A classical result in Functional Analysis, the « bipolar theorem » (see for instance the book [BB1]) shows that U is itself in the vector space spanned by 1, X, X^2. By definition, this means that we can find three real numbers α, β, γ such that :

$$U = \alpha + \beta X + \gamma X^2,$$

and, coming back to our original notation :

$$\text{Log } p_j = \alpha + \beta t_j + \gamma t_j^2,$$

which proves the Lemma.

We observe that the distribution of probability maximizing the entropy is completely determined when the expectation and the variance are known. Indeed, the equations :

$$\sum_j p_j = 1,\ \sum_j p_j t_j = m,\ \sum_j p_j t_j^2 = \sigma^2 + m^2$$

can be written :

$$e^\alpha \sum_j \exp\left(\beta t_j + \gamma t_j^2\right) = 1, \tag{4.18}$$

$$e^\alpha \sum_j t_j \exp\left(\beta t_j + \gamma t_j^2\right) = T_0, \tag{4.19}$$

$$e^\alpha \sum_j t_j^2 \exp\left(\beta t_j + \gamma t_j^2\right) = \sigma^2 + T_0^2, \tag{4.20}$$

and from this we determine explicitly the parameters α, β, γ. The same holds if the expectation and the entropy are known ; the third equation becomes :

$$-\sum_j (\alpha + \beta t_j + \gamma t_j^2) \exp(\alpha + \beta t_j + \gamma t_j^2) = I. \tag{4.21}$$

V. Simplified computations

We may omit the effect of the truncation between $T_{\min}$ and $T_{\max}$ if the observed temperature T_0 is not too close either from $T_{\min}$ or from $T_{\max}$. Then the p_j are of « gaussian type ». Indeed, p_j is maximum at j_0, decreasing on both sides, the same way. Under these conditions, the Lemma gives :

$$\text{Log } p_j = -a(t_j - t_{j_0})^2 + b \text{, with } \ a > 0. \tag{5.1}$$

The value of b comes from the condition $\sum p_j = 1$. We find :

$$b = -\text{Log} \sum_j \exp\left(-a(t_j - T_0)^2\right), \tag{5.2}$$

and replacing into (5.1), we obtain an equation depending upon the sole parameter a :

$$\text{Log } p_j = -a(t_j - t_{j_0})^2 - \text{Log} \sum_{j'} \exp\left(-a(t_{j'} - T_0)^2\right). \tag{5.3}$$

If the entropy is fixed, we get the equation :

$$I = \frac{a}{S} \sum_j (t_j - T_0)^2 \exp\left(-a(t_j - T_0)^2\right) + \text{Log } S, \tag{5.4}$$

with $S = \sum_j \exp\left(-a(t_j - t_{j_0})^2\right)$, and this equation allows us to find a numerically.

VI. Evolution of the entropy at small distance to the measure point

We will determine the form of the probability distribution at a point x_1, situated immediately nearby the measure point x_0. Recall (formula (4.1)) that $x_1 = d / K$. Since the probability distribution is quite concentrated, we may limit ourselves with three terms, of indices $j_0 - 1,\ j,\ j_0 + 1$. In order to simplify our notation, we take $j_0 = 0$. The sum $S = \sum_j \exp\left(-a(t_j - t_{j_0})^2\right)$ can be written :

$$S = 1 + 2e^{-a\varepsilon'^2} \ , \tag{6.1}$$

with :

$$\varepsilon' = \frac{T_{\max} - T_{\min}}{K'} \, . \tag{6.2}$$

(recall that K' is the number of points in the subdivision in t.)

Formula (5.3) gives :

$$\text{Log } p_1 = -a\varepsilon'^2 - \text{Log}(1 + 2e^{-a\varepsilon'^2}) \, . \tag{6.3}$$

But, when $\varepsilon \to 0$, $p_1 \to 0$ (the law concentrates at 0). For a fixed ε', this implies that $a \to +\infty$.

Since the entropy above the point x_1 is $I_1 = \dfrac{A}{K}$ (formula (4.9)), we obtain the relation :

$$\frac{A}{K} = \frac{2a\varepsilon'^2 e^{-a\varepsilon'^2}}{1 + 2e^{-a\varepsilon'^2}} + \text{Log}\left(1 + 2e^{-a\varepsilon'^2}\right) \tag{6.4}$$

When $a \to +\infty$, we have $\mathrm{Log}\left(1+2e^{-a\varepsilon'^2}\right) \approx 2e^{-a\varepsilon'^2}$ and so, with $u = a\varepsilon'^2$,

$$\frac{A\varepsilon}{2d} \approx ue^{-u}. \tag{6.5}$$

Since $e^{-u} \le ue^{-u} \le e^{\frac{u}{2}}e^{-u} = e^{-u/2}$, we obtain :

$$\mathrm{Log}\frac{2d}{A\varepsilon} \le u \le 2\mathrm{Log}\frac{2d}{A\varepsilon}$$

and therefore :

$$\frac{1}{\varepsilon'^2}\mathrm{Log}\frac{2d}{A\varepsilon} \le a \le \frac{2}{\varepsilon'^2}\mathrm{Log}\frac{2d}{A\varepsilon}. \tag{6.6}$$

In short, we obtain approximately :

When the distance between the point x and the measure point x_0 tends to 0, the coefficient a defined by (5.1) for the law of minimal information is of the order of :

$$a \sim c_1\mathrm{Log}\frac{c_2}{d(x,x_0)}. \tag{6.7}$$

Remark : The formula given in (6.7) is a simplification which is not quite correct. In fact, formula (6.5) shows that the result is the solution of the equation $xe^{-x} = \frac{A\varepsilon}{2d}$, which uses the W Lambert function.

VII. Evolution of the entropy at large distance from the measure point

As we already mentioned, if we omit the truncation effect, the formulas give gaussian functions, which make sense as continuous probability distributions.

At distance $d > d_0$ from the measure point, we may as well use the discrete or the continous entropy. The continous one is much easier to compute and gives explicit formulas. We have approximately :

$$\int f(t)\mathrm{Log}\ f(t)dt \approx \varepsilon'\sum_j p_j \mathrm{Log}\ p_j,$$

where ε' is the width of the subdivision in t. If the discrete entropy increases linearly, the same holds for the continous entropy.

If we consider the distribution :

$$h_\sigma(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-(t-T_0)^2 / 2\sigma^2\right),$$

it has mean T_0 and continous entropy :

$$I_c = \mathrm{Log}\left(\sigma\sqrt{2\pi e}\right). \tag{7.1}$$

So we obtain a continous analogue the following way : the parameter x will vary continuously between x_0 and the boundary of the domain ; above the point x we put a continuous entropy of the form :

$$I_c(x) = \lambda x,$$

since the entropy increases linearly with the distance.

Under these conditions, formula (7.1) gives :

$$\sigma(x) = \frac{e^{\lambda x}}{\sqrt{2\pi e}}. \tag{7.2}$$

So finally we obtained explicit formulas :

- Within short distance to the point of measure, the minimal information may be represented by a gaussian, with variance :

$$\sigma \sim \frac{1}{\alpha_1 \mathrm{Log}\left(\frac{\beta_1}{d}\right)}$$

- At large distance from the point of measure, the minimal information may be represente by a gaussian, with variance :

$$\sigma \sim \alpha_2 \exp\left(\beta_2 d\right)$$

where d is the distance between the current point and the measure point. The parameters $\alpha_1, \beta_1, \alpha_2, \beta_2$ depend upon the physical characteristics of the model.

To generalize to several parameters is immediate, but we need first to choose a distance upon the space of parameters.

VIII. Choosing the distance for several variables

At this stage, we consider all parameters as equivalent, in terms of importance : we do not know which ones will be preponderant. So we take a distance which will be symmetric with respect to the various parameters : there are no weights. Such distances may be :

$$d_1(X, X') = \sum_{k=1}^{K} \left| X_k - X'_k \right| \tag{8.1}$$

$$d_2(X, X') = \sqrt{\sum_{k=1}^{K} \left(X_k - X'_k \right)^2} \tag{8.2}$$

$$d_\infty(X, X') = \max_k \left| X_k - X'_k \right| \tag{8.3}$$

Let us observe, however, that the third one does not meet our needs. Indeed, two points X and X' are close to each other for that distance if all their coordinates are similar. But, among 50 coordinates, all of them between 0 and 1, in most cases, two at least will differ significantly, and the distance d_∞ can never be used in practice : one may never say that two points are close to each other, for this distance.

So we request to be able to say that two points are close to each other if most of their coordinates are similar. This is satisfied for the distances d_1 or d_2. In practice, we use d_2, euclidean distance, simply denoted by d.

IX. Two measures

When only one measure has been performed, one feels an impression of intellectual comfort : the only observation gives $T_1 = CT(X_1)$. We propagate this information, making it less and less precise, but still the value T_1 remains the most probable everywhere. We are in the situation of the Englishman who came to Calais, saw a woman with red hair, and deduced that all French women have red hair.

When a second observation is performed, it questions this pleasant certainty. Indeed, in general, $T_2 = CT(X_2)$ does not give the same result as $T_1 = CT(X_1)$. We cannot believe any longer that T_1 is the most probable value everywhere. We have now two informations, apparently contradictory, and we have to conciliate both. At first sight, this is troublesome. But an obvious remark is that two measures are better than just one. The available quantity of information has increased, and the entropy has decreased.

This situation is of course quite common, but it is still troublesome, and we will have to develop a specific mathematical model in order to answer it.

At each point X of the parameter space, we now have two densities of probability. The first one, simply denoted by $f_1(t, X)$, is generated by the measure $T_1 = CT(X_1)$ considered alone, the second, $f_2(t, X)$, is generated by the measure $T_2 = CT(X_2)$ considered alone. We have to conciliate both, and turn them into a single density.

The natural idea is that if the first one is centered at T_1 and the second one at T_2, the resultant will have two « bumps », one around T_1 and the other one around T_2. Here is the aspect of a possible graph, with $T_1 = 1$, $T_2 = 1.4$.

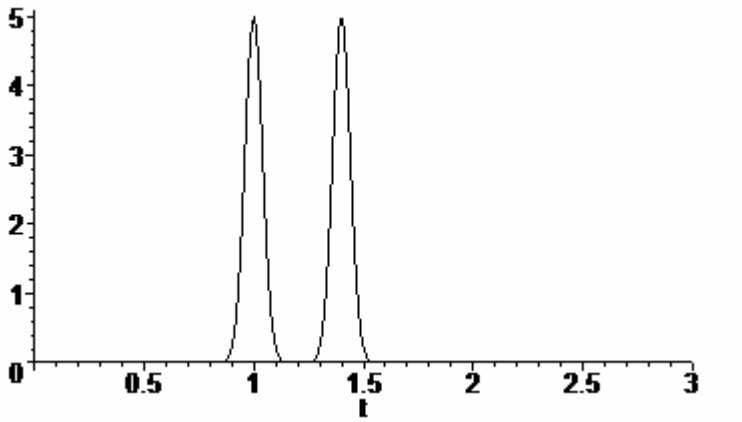

Figure 9.1 : a possible graph in the case of two measures

Why don't we decide that the resulting density should have just one maximum (a single bump), situated somewhere between T_1 and T_2 ? Because, if we did so, we would introduce points that have never been showed by any observation. The only existing measures have shown the possibility of the values T_1 and T_2 ; they never showed the possibility, for instance, of the value $\frac{T_1 + T_2}{2}$, yet that it might have a large probability !

Except if there is a specific information, as a complement, a probability density with several maxima (one for each measure point) is the only shape compatible with the existing measures, no matter if we have two measures or more. But we still have to explain how the two densities should be combined.

Quite naturally, we consider that the resulting density, $f(t, X)$, will be given by a formula of the following type :

$$f(t\,;X) = \lambda\, f_1(t\,;X) + (1-\lambda)\, f_2(t\,;X) \tag{9.1}$$

with $0 \le \lambda \le 1$. Indeed, we need to have a density of probability : we cannot just add the two densities f_1 and f_2.

A. Case of the dimension 1

To start with, let us take $X_1 = 0$ and $X_2 = 1$, in a one dimensional space (just one parameter). If the point X has x, $0 \le x \le 1$,as abscissa, we know that the densities f_1 and f_2 are given by gaussian formulas :

$$f_1(t, x) = \frac{\exp\left(-x^2 / 2\sigma_1(x)^2\right)}{\sigma_1(x)\sqrt{2\pi}}, \tag{9.a.1}$$

$$f_2(t,x) = \frac{\exp\left(-(x-1)^2 / 2\sigma_2(x)^2\right)}{\sigma_2(x)\sqrt{2\pi}}, \qquad (9.a.2)$$

where $\sigma_1(x)$ depends upon the distance between x and 0, and decreases when $x \to 0$, and $\sigma_2(x)$ depends upon the distance between x and 1, and decreases when $x \to 1$. Depending upon the physical characteristics of the model, the functions $\sigma_1(x)$ and $\sigma_2(x)$ are precisely known.

So we get, as a resulting density, an expression of the form :

$$f(t;x) = (1-\lambda(x))\, f_1(t;x) + \lambda(x)\, f_2(t;x). \qquad (9.a.3)$$

In this expression, the functions f_1 and f_2 are known, but not the function $\lambda(x)$. We call this last function « proportion of influence of the second source upon the information at the point x ». If $x = 0$, we are at the point X_1 and this influence is zero, and $\lambda = 0$. If $x = 1$, we are at the point X_2 and this influence is 1. Moreover, quite obviously, this proportion of influence of X_2 may only increase when we get closer to X_2.

The Minimal Information Lemma tells us that, under these circumstances, we may consider that λ is linear. Since $\lambda(0) = 0$ and $\lambda(1) = 1$, we get $\lambda(x) = x$.

If we are at the middle between both points, the influences of X_1 and X_2 are equal, and :

$$f(t;x) = \frac{f_1(t;x) + f_2(t;x)}{2}. \qquad (9.a.4)$$

Let us now treat the case of a space of dimension larger than 1.

B. Dimension larger than 1

Let's see first the case of the dimension 2. We discretize the whole space : let's denote by $x_{i,j}$ the points of the subdivision. The measures, as previously, are in $(0,0)$ and $(1,0)$, to take an example.

The influence of the second measure (at $(1,0)$) will be increasing, on the square $[0,1]\times[0,1]$, if we move, for fixed y, in the sense of increasing x, since we get further from $(0,0)$ and since we get closer to $(1,0)$. The same way, the influence of $(1,0)$ at $x_{1,0}$ is weaker than at $x_{0,1}$: these two points are at the same distance from $(0,0)$, but the second is closer than the first from $(1,0)$. We write this way all the inequalities that should satisfy the couples of variables. This defines a volume V in the space of all coordinates that have been introduced (as many as points $x_{i,j}$). After that, we assume, as we did in the Minimal Information Lemma, in dimension 1, that the influence at each point follows a uniform law, with restrictive inequalities (some are bigger than some others). We assign to each influence the expectation of this law, which gives, for the influence $\lambda(i,j)$:

$$\lambda(i,\, j) = \frac{\int_V x_{i,j}\, dx_{0,0} \ldots dx_{N,N}}{\int_V dx_{0,0} \ldots dx_{N,N}}\,. \tag{9.b.1}$$

Other discretisations are possible, for instance using balls centered at the measure points.

X. The case of N measures

The density at point X will be of the form :

$$f(t\,;X) = \lambda_1 f_1(t\,;X) + \cdots + \lambda_N f_N(t\,;X)\,, \tag{10.1}$$

where $f_j(t\,;X)$ represents the density of probability above the point X resulting from the sole j-th measure. The coefficient λ_j depends from the distances between X and the different points of measure $X_1,\ldots, X_N$. We call it « influence of the j-th measure ».

Set $d_j = d(X, X_j)$. The coefficient λ_j must satisfy the following properties :

1. $0 \le \lambda_j \le 1$ and $\sum_{j=1}^{N} \lambda_j = 1$;
2. If $d_j = 0$, $\lambda_j = 1$ (at the point X_j, only this measure has an influence) ;
3. If $d_i = 0$, $i \ne j$, $\lambda_j = 0$ (at another point X_i, the measure X_j has no influence) ;
4. If all d_i are fixed, except d_j, the coefficient λ_j is a decreasing function of d_j (one gets further from the j-th source of information) ;
5. If all the d_i are fixed, except d_k, $k \ne j$, the coefficient λ_j is an increasing function of d_k (we get further from the k-th source of information).

We see easily that a choice of the form :

$$\lambda_j = \frac{\dfrac{1}{d_j^{\gamma}}}{\displaystyle\sum_{i=1}^{N} \frac{1}{d_i^{\gamma}}}\,, \tag{10.2}$$

with $\gamma > 0$, satisfies all these requests. We now show how to compute the parameter γ, which depends upon the characteristics of the model, that is upon the physical experience which is performed. This parameter is related to the propagation of the information and to the way we combine several informations.

Theorem. – In a space with K parameters, the value $\gamma = K$ is the one which represents the minimal information. In other words, the combination of elementary densities will be given by the explicit formula :

$$f(t\,;X) = \frac{1}{\sum 1/\,d_i^K}\left(\frac{1}{d_1^K}\,f_1(t\,;X) + \cdots + \frac{1}{d_N^K}\,f_N(t\,;X)\right), \tag{10.3}$$

Proof of the Theorem. We may legitimately consider that the coefficient γ does not depend of the number of measures that have been realized.

Let us consider the case of two measures, in dimension K, according to the following figure :

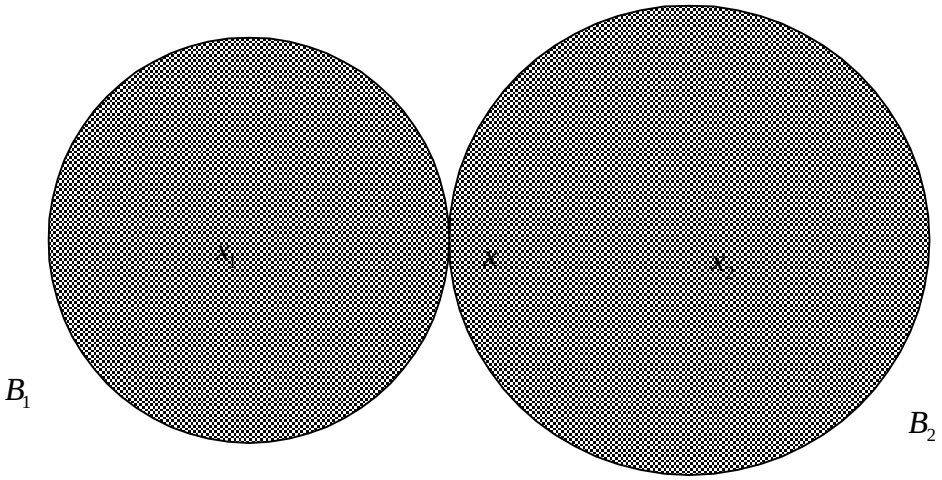

Figure 10.1 : two points of measure and their influence

Here, x_1 and x_2 are two points of measure and x is any point in the segment $[x_1, x_2]$. Let d_1 be the distance between x and x_1 (this is the radius of the first ball in the above figure) and d_2 be the distance between x and x_2 (this is the radius of the second ball). We come back to formula 9.a.3 :

$$f(t\,;x) = (1-\lambda(x))\,f_1(t\,;x) + \lambda(x)\,f_2(t\,;x),$$

where $\lambda(x)$ represents the proportion of influence of the second source at the point x.

Quite obviously, at every point of the ball B_2, this proportion of influence will be larger : these points are closer to x_2 and further from x_1.

The same way, quite obviously, at every point of the ball B_1, this proportion of influence will be weaker : these points are closer to x_1 and further from x_2.

Assume that these balls have been discretized by m_1 points z_i inside the first ball and m_2 points z_i $(i = m_1+1,...,m_1+m_2)$ inside the second one. Let Z_i be the random variable representing the proportion of influence of the second source at the point z_i and X be

the random variable representing the proportion of influence at the point x. This proportion of influence being unknown, as we already did in the Lemma of Minimal Information, we assume it follows a uniform law. The variables Z_i respect the following relations :

$$0 \le Z_1,...,Z_{m_1} \le X \le Z_{m_1+1},...,Z_{m_1+m_2} \le 1. \tag{10.4}$$

In other words, their joint law is proportional to :

$$1_V(z_1,...,z_{m_1},x,z_{m_1+1},...,z_{m_1+m_2}), \tag{10.5}$$

where V is the volume :

$$V = \left\{(z_1,...,z_{m_1},x,z_{m_1+1},...,z_{m_1+m_2}) \in \mathbb{R}^{m_1+m_2+1}, 0 \le z_1,...,z_{m_1} \le x \le z_{m_1+1},...,z_{m_1+m_2} \le 1\right\}. \tag{10.6}$$

As we already did in the Lemma of Minimal Information, the value we attribute to X is the expectation, namely :

$$E(X) = \frac{\int_V x dz_1...dz_{m_1} dx dz_{m_1+1}...dz_{m_1+m_2}}{\int_V dz_1...dz_{m_1} dx dz_{m_1+1}...dz_{m_1+m_2}}. \tag{10.7}$$

The computation is exactly the same as in the Lemma of Minimal Information, noticing that the volume of V is :

$$vol(V) = m_1!m_2!vol(V') \tag{10.8}$$

with :

$$V' = \left\{(z_1,...,z_{m_1},x,z_{m_1+1},...,z_{m_1+m_2}) \in \mathbb{R}^{m_1+m_2+1}, 0 \le z_1 \le ... \le z_{m_1} \le x \le z_{m_1+1} \le ... \le z_{m_1+m_2} \le 1\right\} \tag{10.9}$$

and the denominator of (10.7) is $\frac{m_1!m_2!}{(m_1+m_2)!}$, and the numerator is $\frac{m_1!m_2!(m_1+1)}{(m_1+m_2+1)!}$.

So we get :

$$E(X) = \frac{m_1+1}{m_1+m_2+1}. \tag{10.10}$$

But the number of points m_1 in the ball B_1 is proportional to the volume of this ball, which is itself proportional to d_1^K (recall that K is the dimension of the space). The same way, the number of points in the ball B_2 is proportional to d_2^K. So we get :

$$\lambda(x) \sim \frac{d_1^K}{d_1^K + d_2^K} \tag{10.11}$$

and, putting back into formula (9.a. 3) :

$$f(t;x) = (1 - \frac{d_1^K}{d_1^K + d_2^K}) f_1(t;x) + \frac{d_1^K}{d_1^K + d_2^K} f_2(t;x)$$
$$= \frac{d_2^K}{d_1^K + d_2^K} f_1(t;x) + \frac{d_1^K}{d_1^K + d_2^K} f_2(t;x)$$

Dividing both the numerator and the denominator by the product $d_1^K d_2^K$, we obtain the annouced expression. This proves the theorem.

We observe that the formula defining f is symmetric with respect to the set of observations : the order in which they were made does not matter. This condition is obviously necessary : the result should not depend of the order of the computations. Both for a computational code and for a real-life experiment, the results obtained at each trial are independent of the order of the trials.

So, as we see, each new piece of information (each new computation) makes the density of probability more precise : it « shrinks » it, and this effect is stronger if the measure point is closer.

The definition of these densities of probability allows us to incorporate all the available information, no matter when it was obtained. All the results, even the old ones, are worth using : they contribute to the global information.

In summary, the choice we made for the function f obeys three rules :

- It must be a density of probability ;
- The order of the computations should not have any importance : only the results count ;
- The density should tend to a Dirac mass if we approach a point already known.

XI. Global characteristics of the Hypersurface

We have seen, at paragraph VII, that, in the case of one measure, the propagation of the information, at a point at distance d from the measure, was characterizd by a gaussian density, with variance $\sigma = \lambda a^d$, with $\lambda > 0$ and $a > 1$.

The parameters λ and a need to be determined : it is reasonable to believe that they depend of the model. The propagation of the information has no reason to be the same, in the case of a meteorological measure and in the case of a thermohydraulic computational code. But, for the moment, let us assume that these parameters are fixed.

If the parameters λ and a are chosen, all local densities $f(t;X)$ are known. From these densities, by integration, we compute the local repartition functions. We denote them by $F(t;X)$.

Then, using the local repartition functions $F(t;X)$, we can compute the global repartition function connected with the Hypersurface (that is, to the computational code, or to the physical experience, depending on the situation). It answers the following question : if random values of the parameters $x_1,...,x_K$ are obtained, what value may I

expect for the final result $CT(x_1,...,x_K)$? We denote by $F_{EPH}(t)$ this global repartition function.

Several cases may occur :

1. If I know nothing about the parameters $x_1,...,x_K$, I will consider that each follows a uniform law over some interval, and I will assume that they are independent. In this case, the global repartition function of the Hypersurface will be given by the formula :

$$F_{EPH}(t) = \int_{A_K}^{B_K} \frac{1}{B_K - A_K} ... \int_{A_1}^{B_1} \frac{1}{B_1 - A_1} F(t\,; x_1,..., x_K)\, dx_1 ... dx_K \,. \tag{11.1}$$

2. If I have fixe a priori laws upon the parameters, still assuming they are independent, we have the formula :

$$F_{EPH}(t) = \int ... \int F(t\,; x_1,..., x_K)\, h_1(x_1) \cdots h_K(x_K)\, dx_1 ... dx_K \,, \tag{11.2}$$

where $h_1,\ldots,h_K$ are the densities for each parameter $x_1,...,x_K$.

3. Finally, if I do not assume the parameters to be independent, but if I have the joint law of the K - uple $x_1,...,x_K$, under the form of a density $h(x_1,...,x_K)$, the formula will be :

$$F_{EPH}(t) = \int ... \int F(t\,; x_1,..., x_K)\, h(x_1,..., x_K)\, dx_1 ... dx_K \,. \tag{11.3}$$

We should insist upon the fact that the densities $h_1,\ldots,h_K$, or the joint density $h(x_1,...,x_K)$, come necessarily from the outside : they are given by the experimenter, for physical reasons. They are by no means an internal characteristic of the Hypersurface. But if these densities are modified, one can perform again the computations (11 .2) or (11.3) (which is very simple) and see the impact of the modifications upon the result.

XII. Empirical repartition function

Recall (this is an elementary tool in Probability Theory) that we have at our disposal an empirical repartition function, denoted by $F_{emp}(t)$, related with the experience. Indeed, N measures have been performed (no matter how they were made), and, if the results are put in increasing order :

$$T_1 \le T_2 \le \cdots \le T_N \,, \tag{12.1}$$

the empirical repartition function F_{emp} is defined by :

$F_{emp}(t) = 0$ if $t < T_1$

$F_{emp}(t) = \frac{1}{N}$ if $T_1 \le t < T_2$

$F_{emp}(t) = \dfrac{k}{N}$ if $T_k \le t < T_{k+1}$ $k = 1,..., N-1$

$F_{emp}(t) = 1$ if $t \ge T_N$.

One may consult the book [BB2] for the main properties of this function. Obviously, it is piecewise constant.

If some probability laws have been introduced upon the input parameters, either indepedently, that is $h_1,\ldots,h_K$, or as a joint law, that is $h(x_1,...,x_K)$, and if the values of the parameters are obtained according to these laws, the empirical repartition function F_{emp} will converge, almost everywhere and in L^1 towards a limit, denoted by $F(t)$, when the number of measures increases : see [BB2]. In other terms, we have a deterministic experience, $CT(x_1,...,x_K)$, and we consider a random variable :

$$Y = CT(X_1,...,X_K), \tag{12.2}$$

where the $K-$uple $(X_1,...,X_K)$ has a known law, with density $h(x_1,...,x_K)$. The function F is simply the repartition function for the variable Y. The more measures we have, the better this repartition function is known.

Quite obviously, the result depends essentially from the density $h(x_1,...,x_K)$ that has been chosen. If we decide, for instance, to put the emphasis upon some subset of the whole configuration space (space of all possible values for the parameters), or to restrict the variation intervals for some parameters, the empirical repartition function will be different in each situation. There will be more measures in some subsets of the configuration space.

The choice of the measure points may be « biased » : one chose, intently or not, to privilege some zones which have no interest, and therefore the interesting zones are « diluted » (see the book [BB2] for a description of this phenomenon). In this case, the global statistical characteristics of the model, deduced from the empirical repartition function F_{emp}, will reflect these poor choices, and will be of little use, since of low credibility. Any observer, familiar with the physics of the problem, will notice that the measures were not made where they should have been.

However, even if the laws that have been chosen are poor, the function F_{emp} can still be used in order to determine the propagation parameters inside the Hypersurface, as we now explain.

XIII. Estimating the parameters for the propagation of the information

In the case of a single measure , the propagation of the infomation, at a point situated at distance d from the measure, is characterized by a gaussian density, with variance $\sigma = \lambda a^d$, with $\lambda > 0$ and $a > 1$ (§ VII) ; in the case of N measures, these gaussian are combined using the formula given at § X.

When N measures have been realized, with a joint law of density $h(x_1,...,x_K)$, we have at our disposal :

- an empirical repartition function $F_{emp}(t)$;
- a global repartition function for the hypersurface, $F_{EPH}(t)$.

We may legitimately consider that the two must be close, because, when the number of measures increases, both must be close to the « real » repartition function $F(t)$, which is unknown.

As we will see in the Appendix, one has to work with functions of repartition, and not with densities. The densities, indeed, do not provide a convenient tool to study proximity questions.

So we will estimate the parameters $\lambda > 0$ and $a > 1$ in order to fulfill this proximity. In order to make this clear, let us first imagine that we take λ and a quite large : the variances will be high. The propagated laws will be quite close to a uniform law, and this will hold almost at every place inside the Hypersurface (in fact, every where, except in a small neighborhood, near the measure points). under these conditions, the mean, computed using (11.1) or (11.2) or (11.3), will be very close to the repartition function of a uniform law. Il will not coincide with $F_{emp}(t)$. In other words, if we do not propagate the information, we obtain everywhere a uniform law, and the Hypersurface does not reflect the global statistical properties.

But conversely, if the propagation of the information is too strong, chosing λ and a too small, we find ourselves with, essentially, a density above every point which is a combination of Dirac masses. For instance, if two measures only have been performed, and if they gave respectively T_1 and T_2, the « absolute » propagation of infomation will consist in the density $\frac{1}{2}\left(\delta_{T_1} + \delta_{T_2}\right)$ above each point. In other words, this would say that the only possible values are those which have already been observed, in an equiprobable manner. In this case, the repartition function $F_{EPH}(t)$ linked with the Hypersurface is exactly the same as the empirical repartition function $F_{emp}(t)$, but the construction does not meet our request : we want that other values, than those already observed, should be possible, and we require that the information should decrease with the distance to the existing measures. So we see that it is not legitimate either to propagate the information in too strong a manner.

Our choice will be fixed as follows :

Let us fix a small real number $\varepsilon > 0$, corresponding to the « confidence threshold » we give to F_{emp}. In other words, it is a bound for the distance between the true repartition function (unknown) and the empirical repartition function : $\left\| F_{emp} - F_{réelle} \right\| \leq \varepsilon$ (the norm is here the L^1 norm). It is legitimate that the number ε should depend upon the number of measures : the larger the number of measures is, the closer the empirical repartition function is to reality.

When this is done, we choose λ and a that maximize the global entropy of the Hypersurface, under the constraint :

$$\left\| F_{emp} - F_{EPH} \right\| \leq \varepsilon \ . \tag{13.1}$$

The global entropy of the Hypersurface is computed as follows : we start with the repartition function $F_{EPH}(t)$; we compute the density $f_{EPH}(t)$, and the global entropy is by definition :

$$Ent(EPH) = -\int f_{EPH}(t) \text{ Log } f_{EPH}(t)\, dt \, . \tag{13.2}$$

This global entropy depends of course from λ and a, since $F_{EPH}(t)$ depends from them.

This principle to estimate the two parameters means the following : let us choose the parameters which insure a sufficient proximity with the empirical repartition function, still keeping minimal information. Indeed, we saw that if we propagate the information too much, we get closer and closer to the empirical law.

Remark

We chose here to estimate the parameters, using a reference to the global proximity of the two repartition functions ; this proximity is computed using the L^1-norm. This choice is quite robust, since the repartition function is the one which best describes a probabilistic experience. We might have thought of simpler characteristics, such as the variance. The choice of the expectation is not suitable, since for the Hypersurface, the global expectation is, roughly speaking, equal to the empirical expectation : it is almost independent from λ and a. The only differences come from the truncations (recall that the laws are truncated between $T_{\min}$ and $T_{\max}$). The same way, for the same reason, the choice of any quantile, that is ($P\left(T > T_0\right)$), is not suitable.

Let us also observe that, since the entropy is not linear, the global entropy is not the integral of all entropies above each point.

XIV. When should we stop the construction ?

The concept of Experimental Probabilistic Hypersurface is self-adaptating : it can handle all the information coming from any number of measures and it becomes reacher and more precise each time. Of course, each new measure modifies the Hypersurface locally, since a density (generally diffuse) is replaced by a Dirac mass. But, if the number of

measures is high, this does not modify the Hypersurface globally. Let's take this remark as a criterium for the stopping time in the process, allowing us to conclude that the number of measures which have already been performed is sufficient.

Let $F_{EPH,N}(t)$ be the global repartition function of the Hypersurface, computed using N measures. Let us $\varepsilon > 0$. The number of measures already performed will be considered as sufficient if, when $N' > N$,

$$\left\| F_{EPH,N}(t) - F_{EPH,N'}(t) \right\| < \varepsilon$$

In other words, to perform more measures does not modify significantly the repartition function, measured using the L^1- norm.

In practice, it is of course best the use as many measures as possible in order to constitute the Hypersurface. So one should proceed as follows : for instance with 300 measures : we build the Hypersurface using the 300 measures, we eliminate a small number at random (say 10) and we check that the Hypersurface built using the 290 remaining points is not significantly different from the Hypersurface built using the 300 points.

Fourth Part

The Applications of the Experimental Probabilistic Hypersurface

I. Application of the EPH to the estimate of the probability to go over a certain threshold

Assume that N computations have been made and that the densities $f_N(t\,;X \mid X_1,...,X_N)$ have been built. We do not care about the way they have been constructed, we just have densities for each value of $X=(x_1,...,x_K)$. We denote them simply by $f(t\,;X)$.

Let us see how to use them in order to compute the probability to go over a certain threshold : $P(T>T_0)$. This is a very simple probability problem ; the simplest version uses boxes : assume we have a certain number of boxes, containing white and red balls, in variable proportions (these proportions are known). What is the probability to take a red ball ? Such problems are solved using Bayes formula.

A. Case of a single parameter , with discrete values

In order to illustrate the method, let us start with the case where the function CT depends on a single parameter z ; assume moreover that this parameter takes only discrete values z_i. We write :

$$P(T>T_0)=\sum_{i=1}^{I} P(T>T_0 \text{ et } Z=z_i)$$
$$=\sum_{i=1}^{I} P(T>T_0 \mid Z=z_i)P(Z=z_i)$$

and if the z_i's all have same probability :

$$=\frac{1}{I}\sum_{i=1}^{I} P(T>T_0 \mid Z=z_i)$$

$$=\frac{1}{I}\sum_{i=1}^{I} \int_{T_0}^{+\infty} f(t\,;z_i)\,dt\,.$$

B. Cas of several parameters, with discrete values

In the case of several parameters $(x_1,...,x_K)$, we get the same way :

$$P(T>T_0)=\sum_{x_1,..,x_K} P\left(T>T_0 \text{ et } X=(x_1,...,x_K)\right)$$
$$=\sum_{x_1,..,x_K} P\left(T>T_0 \mid X=(x_1,...,x_K)\right)P\left(X=(x_1,...,x_K)\right)$$

$$= \sum_{x_1,...,x_K} \int_{T_0}^{+\infty} f(t\,;x_1,...,x_K)dt\; P\big(X=(x_1,...,x_K)\big).$$

C. General case : several parameters, laws with densities

We simply denote by $f(t)$ the density of probability of the t (temperature, in our case) : this is what we are looking for.

Let $h(x_1,...,x_K)$ be the density of probability of the K - uple $(x_1,...,x_K)$. As we already saw, if the input parameters are considered as independent, it can be written as a product $h(x_1,...,x_K)=h_1(x_1)\cdots h_K(x_K)$. If they are not independent, we use the joint law of the K - uple under the global form $h(x_1,...,x_K)$. We get :

$$f(t)=\int\!\dots\int_{x_1,...,x_K} f(t\,;x_1,...,x_K)\,h(x_1,...,x_K)\,dx_1\dots dx_K\,. \qquad (1.c.1)$$

If the input parameters are independent, this can be written :

$$f(t)=\int_{x_1}\!\dots\int_{x_K} f(t\,;x_1,...,x_K)\,h_1(x_1)\cdots h_K(x_K)\,dx_1\dots dx_K\,. \qquad (1.c.2)$$

When we know the density, we can obviously compute $P(T>T_0)=\int_{T_0}^{+\infty} f(t)\,dt$:

$$P(T>T_0)=\int_{x_1}\!\dots\int_{x_K}\int_{T_0}^{+\infty} f(t\,;x_1,...,x_K)\,dt\;h_1(x_1)\cdots h_K(x_K)\,dx_1\dots dx_K\,. \qquad (1.c.3)$$

This last formula is quite interesting, because it shows that, in order to compute the probability to pass a certain threshold, we do not need to know completely the functions $f(t\,;x_1,...,x_K)$: all we need to know is, for each of them, the probability that it passes this threshold, that is $\int_{T_0}^{+\infty} f(t\,;x_1,...,x_K)\,dt$.

II. Application to the determination of secondary parameters

A primary parameter is a parameter which influences mostly the result of the computation. Conversely, a parameter will be secondary if the result does not depend much upon the value it takes. We will make this distinction clear, because we want to make a precise difference between primary and secondary parameters.

First of all, a parameter would be useless if the result of the computation was totally independent from its value. For instance, x_1 would be useless if $CT(x_1,x_2,...,x_K)$ was independent of x_1, for all values of $(x_2,...,x_K)$. This will guide us in our definition of a secondary parameter.

A. Definition

Let $\varepsilon > 0$. The parameter x_1 will be secondary of order ε if, for all x_1, x_1', for all $(x_2,...,x_K)$,

$$|CT(x_1,x_2,...,x_K) - CT(x_1',x_2,...,x_K)| \leq \varepsilon \,. \qquad (2.a.1)$$

In other words, the amplitude of all possible variations of F, with respect to the first parameter, is at most ε, no matter what are the values of the other parameters.

The characterization of the fact that a parameter is secondary can be read on the Hypersurface, as we now explain.

B. Experimental Probabilistic Hypersurface and secondary parameters

Let S be the Hypersurface obtained above. We do not have anymore a precise value for $CT(x_1,x_2,...,x_K)$, but a density of probability , $f(t\,;x_1,x_2,...,x_K)$. Let $F(t\,;x_1,x_2,...,x_K)$ be the associated repartition function.

We say that the parameter x_1 is secondary of order ε if, for all x_1, x_1', for all $(x_2,...,x_K)$,

$$\int |F(t\,;x_1,x_2,...,x_K) - F(t\,;x_1',x_2,...,x_K)|\,dt \leq \varepsilon \,. \qquad (2.b.1)$$

This means that the repartition functions connected with the points x_1 and x_1' are close to each other. Again, the reason why we work with repartition functions is explained in the Appendix.

C. Practical Method

- Using simple statistical techniques, one checks that some parameters look secondary ;
- One confronts these choices with arguments coming from physics : if a parameter is secondary, there must be good reasons for that ;
- One computes the densities $f(t\,;x_1,x_2,...,x_K)$ and the repartition functions and one checks formula (2.b.1) above.
- In order to validate a posteriori the fact that the parameter x_1 is secondary, one finally proceeds as follows :

One performs a certain number of random sampling of $(x_2,...,x_K)$: we denote them by $(x_2^{(i)},...,x_K^{(i)})$, $i=1,...,M$. Then x_1 is given several values, in a deterministic way (for instance, dividing its interval of variation into 10 equal pieces) ; we denote them by $x_1^{(j)}$, $j=1,...,10$. We perform the computation $F(x_1^{(j)},x_2^{(i)},...,x_K^{(i)})$ for each value of i and j. For each value of i, the result must be almost independent from j.

III. Finding the critical zones

Quite possibly, for some parameters, some domains of variation may be identified : they do not lead to a critical temperature. If, for instance, the parameter x_1 varies between two bounds a and b, it may happen that, for physical reasons, some interval $[a,c]$ ($a < c < b$) will be safe, in the sense that it will never lead to high temperatures, or, more generally, to dangerous situations.

This must be observed immediately upon the densities $f(t\,;x_1,...,x_K)$ built above each point, since then they will be supported (or mostly supported) inside an interval $[T_{\min},T_1]$, with $T_1 < T_0$ (T_0 is the threshold that we are considering).

IV. How to proceed

1. First, one computes the Experimental Probabilistic Hypersurface, using all the available data ;

2. Then one uses it to eliminate all secondary parameters and all non critical zones. Simple statistical reasoning and observations coming from physics are first used to identify such parameters and zones.

3. When this is done, one completes the data with new computations, concentrated upon the remaining parameters and the critical zones. One builds a new Experimental Probabilistic Hypersurface.

Appendix : Size of random perturbations

In practice, the following question comes quite often : how can we measure the size of a perturbation ? Or, if the same phenomenon has been observed by two different people, how can we measure the size of the difference between the two observations ? Or, if two people have different records, how can we conclude that they observed the same phenomenon, taking into account the errors in the measures ?

A real-life phenomenon, whatever it is, is always described by a table, which represents a law of probability : we put in the table the list of observed values, together with the number of times each of them was observed. If $x_1 < x_2 < \cdots < x_n$ are the observed values, in increasing order, if N is the total number of observations, if n_i is the number of times where the value x_i was observed, we set $p_i = n_i / N$; this is the probability of the value x_i. As we explained in the book [BB2], any observation is always discrete and bounded (finite number of possible values). We may of course build the repartition function F, defined by the general formula :

$$F(x) = P\{X \leq x\} \tag{A.1}$$

where X is the random variable associated to the phenomenon : variable which takes the values x_i with the probabilies p_i.

In some cases, we can also (cf. [BB2]) build artificially a density of probability f and assume it to be continuous and even with a derivative : this depends upon the underlying physical process.

In short, in all cases, our phenomenon will be given either by a table (x_i, p_i), or by a repartition function F, or by a density f ; we will have to work on these three forms. Let us start with the simplest case.

I. Case of a perturbation of known law

In some cases, we have a perturbation with known law. For instance, we may measure directly some phenomenon X and, at the aame time, the same phenomenon through a glass : one records at each time a variable Y. The difference $E = X - Y$ is characteristic of the perturbation due to the glass, and the law of E is known, since at each time we recorded both the value of X and that of Y. Let us observe that what we recorded is in fact the joint law of the couple (X, Y) (see [BB2]).

In such a situation, we know the values $e_1 < e_2 < \cdots < e_n$ taken by E with their probabilities, denoted by p_i as before. How can we characterize the fact that E is small ?

There are many definitions of « tends to zero » for a sequence of random variables :

- The sequence of random variables E_n tends to 0 in probability if, for all $\alpha > 0$, $P\{|E_n| > \alpha\} \to 0$ when $n \to +\infty$.

– It tends to 0 for the L^1-norm if $\mathrm{E}\left(\left|E_n\right|\right) \to 0$ when $n \to +\infty$ (E is the expectation).

– It tends to 0 for the L^2-norm if $\mathrm{E}\left(\left|E_n\right|^2\right) \to 0$ when $n \to +\infty$.

Many other definitions are possible. Which one should we choose, and why ?

We have explained in the book [BB2] that the good tool in order to work on real life phenomena was the repartition function F. According to this, we will try to measure the difference between the repartition function when the error is zero and the repartition function for the recorded error.

The repartition function for the zero error is the Heaviside function $H(x)$ defined by :

$$H(x) = 0 \text{ if } x < 0,\ H(x) = 1 \text{ if } x \geq 0. \tag{A.2}$$

Indeed, if the error is zero, $P\{E < 0\} = 0$ and $P\{E \leq x\} = 1$ if $x \geq 0$.

If now F is the repartition function for the error that has been recorded, we can legitimately take as size of this error :

$$\begin{aligned} size\_error &= \int_{-\infty}^{+\infty} \left|H(x) - F(x)\right| dx \\ &= \int_{-\infty}^{0} \left|F(x)\right| dx + \int_{0}^{+\infty} \left|1 - F(x)\right| dx \end{aligned} \tag{A.3}$$

If the error is given by its repartition function, the above computation is explicit. If the error is given by a table (e_i, p_i), we have :

$$size\_error = \sum_i p_i \left|e_i\right| = \mathrm{E}\left(\left|E\right|\right). \tag{A.4}$$

Indeed, if the error is given by the table (e_i, p_i), the repartition function is :

$F(x) = 0$ if $x < e_1$
$F(x) = p_1$ if $e_1 \leq x < e_2$
$F(x) = p_1 + p_2$ if $e_2 \leq x < e_3$

$F(x) = p_1 + p_2 + \cdots + p_k$ if $e_k \leq x < e_{k+1}$ $(k = 1, \ldots, n-1)$

$F(x) = 1$ if $x \geq e_n$.

Among the values e_i taken by the error, some may be positive, other negative. Let's write :

$$e_1 < e_2 < \cdots < e_m < 0 < e_{m+1} < \cdots < e_n$$

(of course, in some cases, all may be positive, or all negative, in which case the index m does not exist).

Then we have :

$$\int_{-\infty}^{0} |F(x)|\,dx = \sum_{k=1}^{m-1} (p_1+\cdots+p_k)(e_{k+1}-e_k) + (p_1+\cdots+p_m)(0-e_m)$$

$$= -p_1e_1 - \cdots - p_me_m = \sum_{i=1}^{m} p_i|e_i|$$

and the same way :

$$\int_{0}^{+\infty} |1-F(x)|\,dx = e_{m+1}\big(1-(p_1+\cdots+p_m)\big) + \sum_{k=m+1}^{n-1} \big(1-(p_1+\cdots+p_k)\big)(e_{k+1}-e_k) = \sum_{k=m+1}^{n} p_ke_k$$

which shows the formula (A.4).

Of course, if the error is given by a density f, the above formula becomes :

$$size\_error = \int_{-\infty}^{+\infty} |x|\, f(x)dx \tag{A.5}$$

since $\mathrm{E}(|E|) = \int_{-\infty}^{+\infty} |x|\, f(x)dx.$

Since, for any random variable X,

$$\mathrm{E}(|X|) \le \big(\mathrm{E}(X^2)\big)^{1/2}, \tag{A.6}$$

the choice we made to control $\mathrm{E}(|E|)$ brings less penalty than the choice to control $\big(\mathrm{E}(E^2)\big)^{1/2}$.

The convergence in probability does not lead to a control which is satisfactory in practice. Indeed, we control only the probability to pass a threshold, but not the value which has been attained. In other words, if E takes the value 1 with probability 0.01 and if E takes the value 10 with probability 0.01, it is not the same thing in practice. In both cases, the probability to pass the threshold is small (0.01), but the value which was attained is not the same.

Let us observe that, in the formula we chose, we control not only the values and their probabilities, but in fact the sum of all quantities : what we take into account is $\sum p_i|e_i|$, and not only each $p_i|e_i|$ taken separately.

II. Distance between two phenomena

Let us now turn to a situation that is much more frequent than the previous one : the law of the perturbation is unknown. Two different people, or the same person at two different times, have observed similar phenomena, and we wonder whether or not it is the same phenomenon.

In practice, each person filled his own table : (x_i, p_i) for the first, (y_j, q_j) for the second. The number of observations do not need to be the same ; the observed values neither, nor of course the probabilities. We have to compare the two phenomena and decide if they are similar.

Here again, the most pertinent indicator is the L^1-norm of the difference between the two repartition functions :

$$dist\_phenomena = \int_{-\infty}^{+\infty} |F_1(x) - F_2(x)|\, dx \tag{A.7}$$

where F_1 and F_2 are, respectively, the repartition functions associated with (x_i, p_i) and (y_j, q_j).

One should observe that it is impossible, in the present case, to speak about the law of the difference, as we did in the previous paragraph. If X represents the first random variable and Y the second, we know the laws of X and Y, but we do not know the law of the couple. This law, quite often, does not even make sense, because the two variables have been recorded at different places or different times.

Let us take a very simple example in order to illustrate this. The laws of X and Y may have both the following form :

value	-1	0	1
proba	1/3	1/3	1/3

and they look identical. But, quite possibly, X indicated -1 when Y indicated 1, and conversely ! This is the case, for instance, if $Y = -X$. In the case of a symmetric phenomenon, ($P\{X = x\} = P\{X = -x\}$), if $Y = -X$, X and Y have the same law, but they are not close to each other, in the sense that $X - Y = 2X$ may be very large.

If the two variables are given by their repartition functions, the computation of (A.7) is immediate. If they are given by tables (x_i, p_i), (y_j, q_j), we proceed as follows :

First, we build a common list, denoted z_l, from the points x_i and the points y_j. Then, for the variable X, we define its law of probability, on the list z_l, by : if z_l is one of the x_i, the probability is the corresponding p_i. If z_l is one of the y_j, the probability is 0.

For the variable Y, it is the converse ; the probability on the list z_l is : if z_l is one of the x_i, the probability is 0, if z_l is one of the y_j, the probability is q_j.

Let p'_l be the probabilities obtained for X on the list z_l and q'_l the probabilities obtained for Y. Then :

$$dist\_phenomena = \sum_{k=1}^{N} |(p'_1 + \cdots + p'_k) - (q'_1 + \cdots + q'_k)|\,(z_k - z_{k-1}) \tag{A.8}$$

where N is the total number of points in the list z_l. Indeed, formula (A.8) is the traduction of the definition (A.7), since the repartition functions are piecewise constant.

We observe that the distance between two phenomena cannot be correctly measured using the densities. Indeed, suppose that the variables X and Y have densities, denoted by f and g. We might want to consider :

$$d(X,Y) = \int_{-\infty}^{+\infty} |f(x) - g(x)| dx \tag{A.9}$$

which is obviously a distance. But it cannot be used in practice. Suppose for example that both X and Y are certain : their density of probability is a Dirac mass. Assume moreover that their values are quite close : for instance, $X = 1$ constantly and $Y = 1.001$ constantly. Still, we would have $d(X,Y) = 2$, which is not satisfactory.

References

[BB1] Bernard Beauzamy : Introduction to Banach Spaces and their Geometry. *North Holland, Collection « Notas de Matematica »,* vol. 68. First edition : 1982, second edition : 1985.

[BB2] Bernard Beauzamy : Méthodes probabilistes pour l'étude des phénomènes réels. *Ouvrage édité par la Société de Calcul Mathématique,* mars 2004.

[Ecotox] Analyse statistique et modélisation mathématique des effets toxiques sur les espèces vivantes. *Etude réalisée par la SCM dans le cadre d'un contrat avec l'INERIS,* 1999-2002.

[Sobolev] "Sur les inégalités de Sobolev logarithmiques", C. Ané, S. Blachère, D. Chafaï, P. Fougères, I. Gentil, F. Malrieu, C. Roberto, G. Scheffer, *coll. Panoramas et Synthèses 10, Société Mathématique de France,* 2000.