

July 13, 2007

SOME NOTES AND QUESTIONS ON PROBABILISTIC HYPERSURFACE

MANUSCRIPT FROM S. DESTERCKE

ABSTRACT. Some more comments and questions added to the one made during the presentation.

answers by SCM in blue

1. INTRODUCTION

Here is some comments, questions, propositions about the probabilistic hypersurface presented at the INSTN.

We start with some general considerations and questions of practical importance (section 3), to pursue by some (simple) suggestions of interesting studies, mainly for the parts C to F of the presentation (section 4). We finish with more conceptual questions, that are perhaps less important for down-to-earth and practical applications.

2. NOTATIONS

We use the following notations along the comments:

_ $X; X_{\max}; X_{\min}$: vectors of input (lower bound and upper bound of) parameters

_ $x_i; x_{i;\max}; x_{i;\min}$: one particular parameter (here, the i th)

_ $Y; Y_{\max}; Y_{\min}$: vectors of (lower bound and upper bound of) output parameters

_ v : Number of discretized values

_ f : function corresponding to the model

_ $p; P; F$: respectively probability density, measure and cumulative distribution

_ $d_{\max}; I_{\max}$: maximal distance defined in the EPH and maximal entropy.

_ w_i : weights of probability densities defined in the EPH

For convenience, we assume the metric to be the same for all parameters, unless otherwise stated.

3. GENERAL COMMENTS AND QUESTIONS

We tried to make clearer some discussion made during the presentation, particularly the one concerning expert knowledge.

Main interests of the method.

First, thanks again for the presentation. From what we've understood, it seems that the main interesting points of the method are:

(1) its ability to handle high-dimensional problems and how efficient can be made subsequent calculations, all this with simple concepts.

(2) the idea of propagating the information given by an experimental point and make it less and less certain as we go away from it.

(3) adding as few information as possible while making reasonable assumptions about the underlying model (indeed, if you don't make such assumptions, adding no information except from the one you already have comes down to say that you don't have any way of learning (or deducing) new information from the one you have)

So, the method offers a simple and coherent way to build a probabilistic response surface, and thus deserve some attention.

In our opinion, the main interest of EPH is that it gives a well-defined way of "storing" information, without unnecessary assumptions. If you made 321 measures, this is what you can expect

Robustness with respect to parameter bounds and discretization.

It is clear that the result is both dependent on the discretization of parameter space and of the discretization of this space. So the questions are:

- _ If, instead of considering $X = [X_{\min}; X_{\max}]$, we consider $X_{\varepsilon} = [X_{\min-\varepsilon}; X_{\max+\varepsilon}]$, with ε a small value. Are the built surface robust to such small changes?
- _ In the same way, what happens if we multiply or divide ν by, for example, two?

Since the model assume regularity of the response surface and the entropy decreases linearly, we assume the method is robust to such small changes (i.e. a small change in the input will only provoke a small change in the result). This is nevertheless an important matter, since in practice, given bounds are often precise approximations of imprecisely known bounds.

Yes, this is correct, that a small change in the parameter provokes only a small change in the result. In practice the bounds of the parameters are governed by the physical characteristics. The modifications on the EPH induced by changes either on the max or on the min are studied in the paper.

If we increase ν , we also increase the maximal entropy I_{\max} , hence we can expect the response of a given point to be less influenced by the closest local models in the final EPH, and thus the final result should be smoother. At the same time, increasing ν imply a higher computational cost. If this is true, how do we define a good value for ν ? (e.g. we could define a value of ν s.t. the maximal entropy I_{\max} is close enough to the entropy corresponding to the continuous case)

The same as for the boundaries of the parameters, the discretization ν is characterized by physical conditions and expert knowledge. In fact, the choice of ν is simply given by the precision which is sought (for instance one degree Celsius). We regard it as something which is imposed to us, not as a parameter at our choice.

Adding information, expert knowledge and physical constraints to the method.

These comments are related to the questions asked during the presentation concerning prior information. we use a simple and purely illustrative example to give a better idea of the problem.

Example 1. Let us assume we have a model, predicting the bacteria population (expressed in concentration) ($B \in [0;1]$) of a particular bio process, and that this model depends both on pressure ($P \in [1;3]$ bars), temperature ($T \in [300;500]$ K) and the input flow rate of some substance on which bacteria feed ($Q_{in} \in [0;0.9]$). Now, before any use of this model and of the EPH, we assume to have one or more of the following pieces of information:

(1) An expert (that works in the field and on the same physical phenomenon, but can use a different model) was asked the following question: *Assume the temperature is close to 330 K, the pressure is of 2 bar and we do not know the input flow rate, because of some sensor deficiency, what do you think the state of the population will be?* the expert then express his answer about the bacteria population in term of three percentiles 5%;50%;95% for which he gives the following values $F^{-1}(5\%) \in [0;0.02]$, $F^{-1}(50\%) \in [0.08;0.1]$, $F^{-1}(95\%) \in [0.12;0.15]$. In short, the expert thinks that in such a situation, the bacteria population concentration is likely to be low.

(2) It is commonly admitted in the scientific community that, provided the temperature is high enough and the bacteria have enough food, the concentration will be high, whatever the pressure value. This can be traduce by the probability $P(B > 0,8 | T \in [450;500]; Qin > 0,5) > 0,9$

(3) It is known for sure that bacteria cannot survive without food, thus we have the constraint $P(B > 0 | Qin = 0) = 0$

Now assume we perform a batch of code runs and thus get a set of points. Of course, we can expect the model to integrate such information as the one given above, but it would be unrealistic to think that the code integrate all the available information about the given problem. So our questions are:

_ Can you integrate the above information (constraints) to the EPH so that, used in conjunction with code runs, they improve the final result ? If no, then do you see an easy way of doing it ?

_ Now, suppose that the model or the EPH are conflicting with some of these information (for example, when the temperature is close to 330 K and the pressure is around 2 bars, both the model and the EPH based on this model find a probability of concentration centred around 0.3, and not 0.1 as the expert thought): is it possible to detect such conflict in the information so that you can tell the expert or the person that have built the model ? how would you treat such conflicting information in the building of the EPH ?

Of course, the basic setting of the EPH assumes no prior information before any code run. Nevertheless, we think that taking account of prior information or information from other sources (experts, new physical experiments) would be very useful, since more often than not, we have some pieces of partial information that are not coming from the particular model we're working on.

See BB's file "Taking into account preliminary information in the EPH"

In the same line of thought, would it be easy to integrate (imprecise) knowledge about input parameters dependencies (for instance, if you that, generally, x_i is high when x_j also is, how would you integrate such information?).

First of all, the integration such information is possible only when the matter concerns computing the global probability.

Let us remind some notation :

$$R(X) = \sum_{t_j=T_0}^{t_{\min}} \sum_{n=1}^N \gamma_n(X) \cdot \frac{c_n \tau}{\sqrt{2\pi\sigma_n(X)}} \exp \left\{ -\frac{(t_j - \theta_n)^2}{2\sigma_n^2(X)} \right\}$$

Here is a local probability to be above a certain threshold T_0 at some point X .

So, when we compute the global probability, then three cases may occur :

- 1) If we know nothing about the parameters x_1, \dots, x_K , we will consider that each follows a uniform law over some interval, and we will assume that they are independent. In this case, the global probability to be above the threshold T_0 is :

$$P_{global} \{t \geq T_0\} \approx \int_0^1 \dots \int_0^1 R(X) dx_1 \dots dx_K$$

- 2) If we have a fixed law upon each parameter but still we are assuming they are independent, then the probability will be given by the formula :

$$P_{global} \{t \geq T_0\} \approx \int_0^1 \dots \int_0^1 R(X) h_1(x_1) \dots h_K(x_K) dx_1 \dots dx_K$$

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

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

$$P_{global} \{t \geq T_0\} \approx \int_0^1 \dots \int_0^1 R(X) h(x_1, \dots, x_K) dx_1 \dots dx_K$$

In this case, the joint law h may be anything you want, and this may include all types of dependences between parameters.

Another point is the treatment of a multi-dimensional output with eventual dependencies (and the construction of the associated joined probability density).

See the answer below.

Are local probabilities conditional probabilities ?

It seems to us that what you call local probabilities are in fact special cases (i.e. cases where the conditional values are precise) of general conditional probabilities. If it is the case, it would probably mean that computing any conditional probability becomes quite easy, which is a very interesting feature of the EPH.

Yes, what we call “a local probability” it is a special (I would say “single”) case of a global probability : because a local probability is computed for a certain single point, whereas a global probability is computed for all possible points.

Let's be more specific : all probabilities in the EPH are indeed conditional probabilities, since they are conditioned by the events : you measured this at that place(s). A local probability refers to the probability that, for a given (new) value of the parameters, you obtain some result. A global probability integrates over the whole space, or part of it.

Inverse problem.

Assume you have a model f with $f(X) = Y$, and we have some probabilistic model of Y . The inverse problem then consist of building the uncertainty on X (the joined probability density), knowing the uncertainty on Y . Otherwise stated, what is the joined probability on X s.t. $f(X)$ will give back the initial model on Y ? A lot of people are interested in solving this quite difficult problem, and we just wondered if the EPH couldn't help to solve it in some way?

There is one possible way to solve this problem, namely :

For example we are interested about the case then the temperature belongs to some interval $[t_1; t_2]$ and we would like to know what value can be taken by each parameter in order to obtain this case.

In order to find such values, we compute the local probability at each point of the configuration space and we choose those ones which satisfy our conditions : for example we accept all points for which the probability to be inside the interval $[t_1; t_2]$ is more than 95% (or more than 99%).

In practice this can be done by the method Monte Carlo.

Of course in this case we will not have the probability law for each parameter, but just a set of points which answer our needs. Then we can study each parameter using the obtained values.

Second proposition, in our opinion, will be more appropriate than the first one, but it is not studied completely by us for the time being, here is :

Case of one dimension and one measure :

We can consider the temperature as the parameter (that is input) ; and our parameter is the result of the computation code (that is output) : $x = CT(t)$

We made one experiment and we found that : $A_1 = CT(\theta_1)$

Then, we can build the local probability at any point of the temperature, such is that :

$$p_j(t) = \frac{c\tau}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(x_j - A_1)^2}{2\sigma^2}\right),$$

where t is the temperature, x_j is our parameter (which was discretized), A_1 is the value of the measure point.

So, in some sense, we look directly at the inverse problem from the computational code.

The same idea appears when we have many measure points.

Case of many-dimensional space

Now we assume that we have one parameter t (i.e. a temperature), but the result of the computational code (i.e. our parameters) lies in many dimensional space :

$$X = (x_1, \dots, x_K) = CT(t)$$

If all parameters are independent, then the problem is solved : we take each parameter separately and compute the distribution of probability as is the case with one dimensional space.

Otherwise we have to combine all density of probabilities in a way which we did not think yet.

4. INTERESTING STUDIES AND SUGGESTIONS

These are some suggestions and studies that could be of interest in the various information treatments that you've presented to us.

4.1. Using other operators to aggregate densities.

Have you tried to use other aggregation operators in the combination of densities? Is the EPH sensitive to small changes? We agree that, among possible combination operators, the arithmetic weighted mean is one of the more convenient (particularly because of the marginalization property that avoids any need of renormalization) and probably the operator having the most interesting properties. In particular, we are curious about the sensitivity of the result with respect to the use of a particular operator. For instance, given two sources (here, probability densities) p_1, p_2 , any functional f s.t.

$$p_{12} = f^{-1}(f(p_1) + f(p_2))/k$$

with k the normalization factor, can be considered as a mean between p_1 and p_2 . For example, what would give the geometric mean ? (i.e. the product of densities p_i)

There are indeed many ways of combining information ; the one we use seemed to us the most appropriate, because it does not rely upon particular assumptions. Other ones might be used, but they would need specific justification.

4.2. Improving the search of high probabilities.

We strongly think you could improve your search of high probabilities (part C of the talk) by using some features of typical heuristic methods (as simulated annealing). For instance, two ideas are :

_ To start with a ball of big radius (and drawing more stochastic points) and then to decrease (and drawing less stochastic point) it at each step of the search. The idea is that, at the begin, we want to make big jumps to probabilities that are higher and higher (thus avoiding to get stuck in a local maxima), while at the end we just want to get nearer the maximum we have found. Moreover, by this way, we think that reaching the maximum will take less steps in the iterative process.

_ Another solution (used in simulated annealing), is to sometime accept to "worsen" the solution (i.e. at one step, when there is no point with higher probability than the original one, try to take the second highest point and make a new iteration). The idea is again not to get stuck in a local extremum. and there are surely many similar ideas used in heuristics that you could use to improve the search.

Of course there are many other possibilities to search the point with the highest probability, very likely we chose the simplest one, but it was just to present the one of the application of the EPH. We agree that this question needs more improvements which we are going to do in the future.

We think that it would be interesting to try several variations of the algorithm upon a specific example. We certainly do agree that our algorithm is not best possible.

4.3. Using more finer clustering methods to detect dangerous zones.

Basically, the method you are using comes down to group parameters by average values, and then to cut a zone by taking the same hyperplane parallel to each dimensions in a given group of parameters. Although it allows to have very quick results, we think it could lead to clusterization having a poor quality, due to the following reasons:

_ It can be dangerous to subdivide parameters by average values. Just consider that we have ten points having similar outputs. Now, let us assume that parameter x_i takes the value 0.5 for these ten points, and that another parameter x_j takes value 1.0 for five points, and value 0.0 for the five other points. Both parameters have an average of 0.5, and will thus be in the same subgroups, while it is clear that their behaviors are quite different.

_ Hyperplanes parallel to each axis of the space $[0;1]^{51}$ are very simple functions and thus very easy to use, but this simplicity also imply a very low descriptive power.

we think you could advantageously use classifications or clustering algorithms that will allow for a finer analysis of your data. Working in a 51 dimensional space remains tractable for some clustering algorithm especially designed for high dimensional problems (for example, the EM algorithm).

Here again, our clustering method is indeed very preliminary. Our choice was made in a simple manner, in order to show that we could bring a solution to IRSN's problem (detecting dangerous zones). We certainly agree that improvements upon the division are possible. Also, the physical behaviour of the parameters can be very useful and interesting.

4.4. Evaluation of the parameters : finding the important ones.

As said during the presentation, you have a representation of the joined density with which it is easy to achieve computations. This is why we think you could without too much problems compute usual sensitivity indices such as Sobol indices, Partial Correlation Coef_cient,

If it is the case, it would be interesting to compare their result to your proposition.

We never tried this, at this point. We will be glad to do it on a specific example. We agree that is has to be done.

4.5. Reconstruction or prediction of data.

It would be interesting to compare, on benchmark or any real data set, how the EPH performs compared to some other usual statistical techniques (e.g. ARMA or ARIMA models, use of Kalman filters, . . .).

Same as above : we will be glad to do it. However, we wish to emphasize the fact that all the techniques you mention rely upon specific assumptions (linear, gaussian, and so on), and this is precisely what we try to avoid in the EPH.

5. CONCEPTUAL COMMENTS, QUESTIONS

Changes in the entropy slope (parameter λ).

we think basing the whole method on a single parameter is seducing, mathematically elegant and, more important, convenient. Nevertheless, we have a little remark concerning λ . Assume we have n experimental points. If we obtain one more experimental points, two things can occur:

_ either the maximal distance between the $n+1$ th point and the bounds is greater than for the n points, d_{max} increases and λ decreases, making the entropy increase less important for all other points. Thus, in this case, the importance of all previous experimental points increase (as they propagate distributions having lower entropies)

_ or the maximal distance is not increased by the interval of the new point, and this time, the importance of all previous experimental points is not changed. So, in some cases, the arrival of a new point will influence the behavior of all the others, while in other cases it won't. Another effect is that central points will get more and more importance as new experimental points arrive.

Yes, this is correct : when a new measure point arrives, sometimes lambda changes, sometimes it does not. But there is a limiting value, when the number of measures increases.

Of course, changing the slopes allows to modulate the importance of a given point with respect to its distance from the point for which we want to compute the output. But we could argue that this distance is also taken into account in the weights of the arithmetic mean used to combine probability densities, and thus that the distance of a point with respect to another point is counted twice in the building of the EPH.

So what is unclear is how parameter λ and weights w_i interact with each others, and in which measure could we consider other options (e.g. one λ_i per point)? Indeed, as more experimental points are found, the decreasing slope for λ increase the contribution of the closest local model, while the use of a weighted average lower this same contribution.

First of all λ and w_i are really independent from each other : we compute λ as soon as we obtain all measure points we need, and we recompute λ when a new measure point is coming.

This change of entropy due to increasing d_{max} concerns all measure points to the same extent, none of them gets more important in this way, just all “bumps” become narrower at the same time; and it is completely logical : when we perform a new measure then all global information can remain the same or become more precise.

Why λ is not changed every time when a new measure is arriving : it is explained by a principal of the minimal information, we should not add any unnecessary information so, we keep the rule that at the furthest point we have a uniform law.

About w_i (which consist of the distances from the unknown point to all measure points) : first, they take part in the forming of the density itself (they govern “a size” of the bump) and second, they are needed to compute w_i in order to obtain the resulting density of probability.

considering imprecise probabilistic models.

Some of our research being mainly focused on the theory of imprecise probabilities (Sebastien's thesis is about the treatment of information in this framework), we were wondering if, instead of decreasing the entropy, we could not consider wider and wider probability families as we get farther and farther from any experimental point ?

So, given an experimental point, we could propagate it so that the propagated families of probabilities always contain this Dirac measure, while containing more and more other distributions as we get farther from the point.

Any point outside an experimental point would then become a weighted arithmetic mean (if we choose this combination operator) of the propagated probability families at this point.

Yes, certainly, other approaches may be considered, with the same result, namely that the information becomes less precise when you get further away from measure points. But precisely, information is linked with entropy (this is why entropy was chosen to govern the whole construction), and information is not directly linked with variance. So variance appears as secondary to entropy. In some sense, EPH is an attempt to use information theory (and probabilities come only as a tool).