Société de Calcul Mathématique SA

Outils d'aide à la décision
depuis 1995

The confidence intervals of an extrapolation method: a surprising behavior.

Gottfried Berton
Société de Calcul Mathématique SA
111 Faubourg Saint Honoré
75008 Paris France

November 2017

I. Abstract

The EPH (Experimental Probabilistic Hypersurface) is a method, introduced by SCM (see the book by Zeydina and Beauzamy "Probabilistic Information Transfer"), which is used in order to propagate the existing information toward unknown locations. It can be used to reconstruct missing data or to make predictions.

The main advantage of this method is that it does not introduce arbitrary information or assumptions about the data. The propagation of information relies on a general principle of maximal entropy (or minimal information) which is itself an increasing function of the distance to the measurement point.

Usually, for any extrapolation method, the size of confidence intervals increases with the distance: predicting further leads to poorer predictions than short term ones. But we see here that this is not always the case, which is quite surprising in itself.

The present work originates in a contract with the French "Institut de Radioprotection et de Sûreté Nucléaire" (contract EX10/12022486, 16/04/2015), which dealt with the comparison between EPH and the usual Kriging methods. Our special thanks to Dr. Yann Richet, from IRSN, and to Dr. Giovanni Bruna, scientific director of IRSN, for bringing such questions to our attention.

II. Basic information about EPH

The result provided by EPH is given under the form of a collection of discrete probability densities having maximal variance for the fixed entropy. Such a density takes the form of a Dirac function at the measurement point location (the value is known precisely), and becomes less and less concentrated when moving away from it.

The EPH model requires two input parameters:

- bounds on each dimension. For example, if the historical data are temporal, one has to set a time min and max on which to perform the reconstruction;

- bounds on the outcome range and discretisation path; the resulting estimates has the form of a discrete probability law on the defined range.

The bounds may come from expert knowledge, physical limits or be defined by a user.

We have N observation points, denoted by A_n, $n=1,\dots,N$. Let C_n be the outcome value of the nth measure. Let X be the point where we want an estimate. Let $d_n = d\left(A_n, X\right)$ be the distance between the point to reconstruct and the nth point of measure.

Let j be the discretisation of the result range with step τ, and λ be a parameter related to the entropy which is calculated so as to maintain the information minimal at every point (see the book [PIT]).

Each of the measurements gives its own contribution to the final result, written under the following form:

$$p_{n,j}(X) = \frac{\tau}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(j-C_j)^2}{2\sigma^2} \right),$$

where $\sigma = \dfrac{\tau e^{\lambda d_n}}{\sqrt{2\pi e}}$.

At the end of the process the individual laws are recombined in order to get a single one depending on the distance of the target-point from each measurement:

$$p_j(X) = \gamma_1 p_{1,j}(X) + \dots + \gamma_N p_{N,j}(X),$$

where $\gamma_n = \dfrac{d_n^{-1}}{\sum_{i=1}^{N} d_i^{-1}}$, $n=1,\dots,N$

From this probability law, one can extract a confidence interval. The lower bound of this confidence interval is the quantile 5% of the probability law. The upper bound is the quantile 95%.

III. Confidence intervals

We will see that, under certain circumstances, the width of the confidence interval decreases when predicting farther and farther away from the observations. This is not the case in general: when reconstructing far from the observations, the confidence interval is usually larger because there is more uncertainty about the estimate, as we see on the example below. The confidence interval is represented in blue:

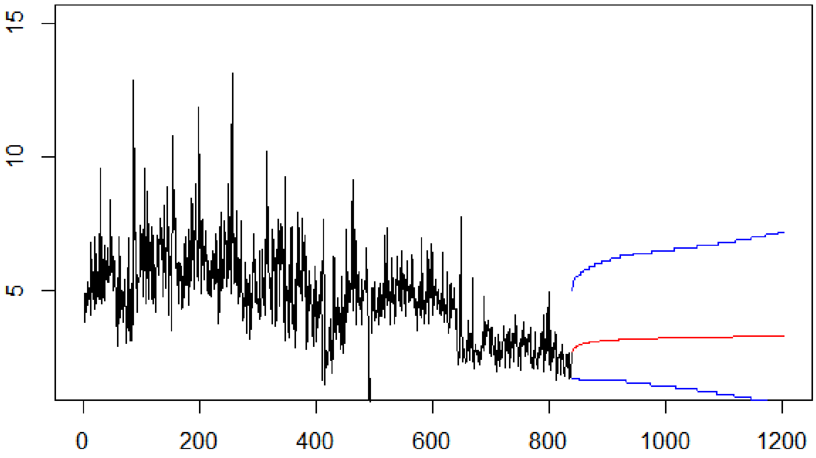

Figure 1: confidence interval of EPH in general

Let us treat a simple example.

Example of strange behavior

We wish to reconstruct the temperature function $T(t) = \sin(200t) \times t^6$ over the time range [2;15], using 150 measures located in the time interval $[0;2]$. The function to reconstruct is the following:

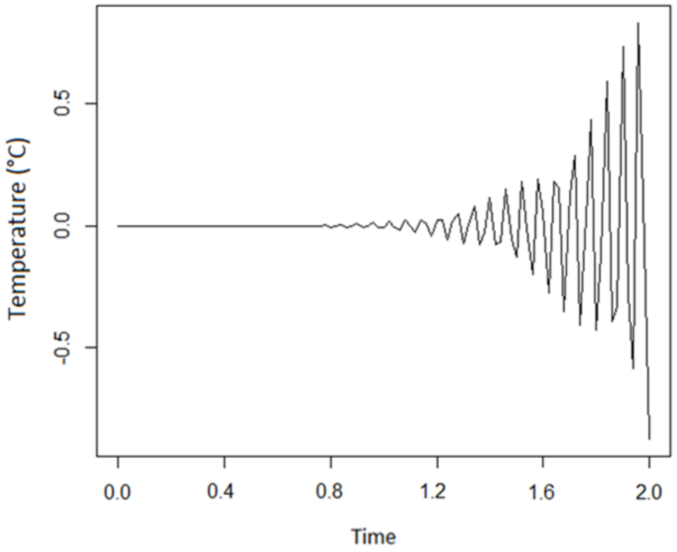

Figure 2: function to reconstruct

The prediction obtained with EPH (red) and the confidence interval (blue) are displayed below:

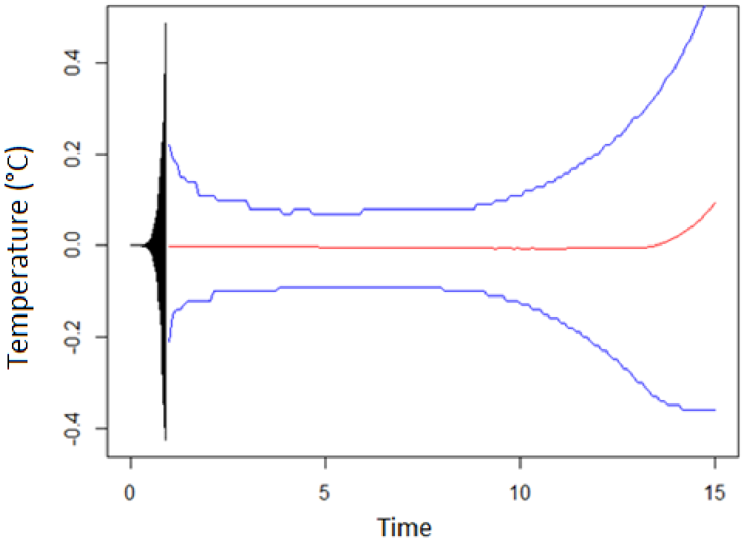

Figure 3: prediction and confidence interval

The bounds of the confidence interval are the two quantiles 5% and 95% of the probability law returned by the EPH.

We see that this interval tightens when moving away from the observations. However, when making the prediction in a sufficiently far future, the width of the confidence interval increases again. This is surprising because usually the confidence interval enlarges: the less we know the larger is the uncertainty.

The tightening of the confidence interval is due to the decreasing weights of the recent observations. The recent data have less and less importance. The following graphs shows the weights of a recent, and an ancient observation with respect to the time.

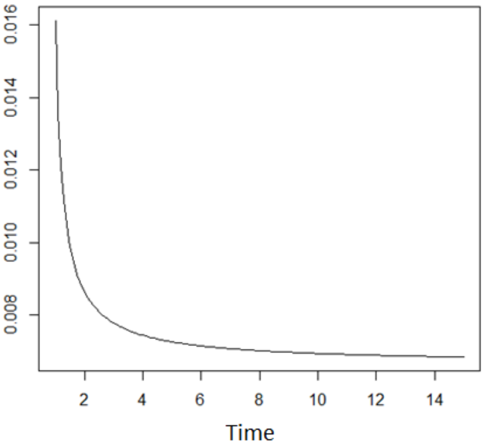

Figure 4: weight of a recent observation

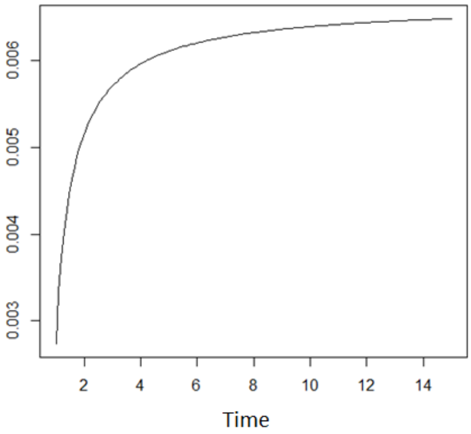

Figure 5: weight of an ancient observation

When moving away from the observations, the weights associated to the recent measures are decreasing, and conversely the weights associated to the ancient ones are increasing: they converge to the same value $\frac{1}{150}$. In this example, the recent observations are extreme. Hence, the weights associated to the extreme values decrease, and the probability law changes accordingly. The probability laws returned by the EPH at $t=2$ and $t=5$ are given below:

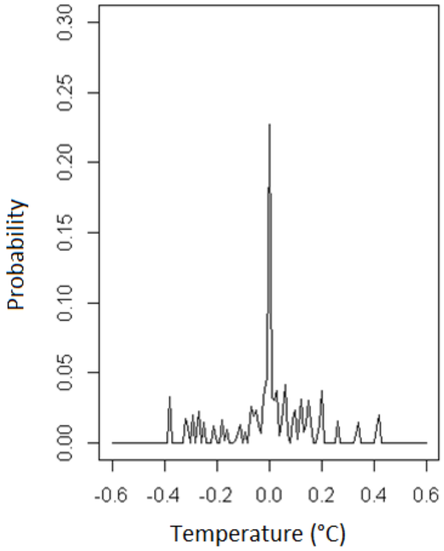

Figure 6: probability law at $t = 2$

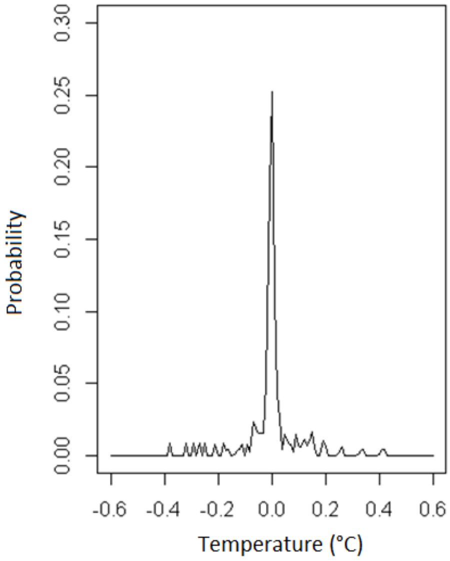

Figure 7: probability law at $t = 5$

Since the weights of the extreme observations are decreasing, the spikes associated to these observations are lower at $t = 5$ than at $t = 2$. Therefore, the quantiles of the probability distribution get closer to zero, and the confidence interval is smaller.

So the explanation to the strange behavior is this: the EPH takes into account the past, with a weight which decreases: ancient past is less important than modern past. If the ancient past is quite stable and modern past is quite oscillating, as in the example above, moving away from the recent past will put less and less weight on it, and so the oscillations will not be considered so important.

In some sense, a rough description is this: we have a phenomenon which proved to be very unstable in the recent past, but was more stable before. Then moving ahead will put less weight on the recent past and will lead to narrower confidence intervals.

However, when predicting far enough in the future, the confidence interval enlarges again, roughly from $t = 8$ in the example above. There are two reasons:

- the variance of the probability laws associated to each observation increases exponentially with the distance to the point to reconstruct;
- the weights of the recent observations are not decreasing any more.

The probability distributions associated to the observations are becoming broader and broader and the confidence interval is enlarging again.

IV. References

[PIT] Olga Zeydina et Bernard Beauzamy: Probabilistic Information Transfer. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA. ISBN: 978-2-9521458-6-2, ISSN : 1767-1175. Relié, 208 pages, mai 2013.