Multi-fidelity regression using a non-parametric relationship

Federico Zertuche, Celine Helbert, Anestis Antoniadis

MascotNum2014 conference - Multi-fidelity regression using a non-parametric relationship.

F. Zertuche, C. Helbert (ICJ) and A. Antoniadis
Université Joseph Fourier, Grenoble

Affiliation: *Laboratoire Jean Kuntzmann*, UJF Grenoble, 51 rue des Mathmatiques, BP 53 38041 Saint Martin d'Hres, France

Email: zertuche.federico@imag.fr – **URL**: https://team.inria.fr/moise/en/

Master: Université Jean Monnet, Saint-Etienne

Ph.D. (2012-2014): Université Joseph Fourier, Grenoble

Supervisor(s): Prof. A. Antoniadis (UJF) and Dr. C. Helbert (ICJ)

Abstract:

We study the synthesis of data from different experiments. These experiments are very complex computer simulations that take several hours to produce a response for a given input. Understanding the phenomenon modeled by the simulation requires a large number of responses and in practice having all of them is unfeasible due to time constraints. This is why the computer simulation is often replaced by a simpler probabilistic model, also known as metamodel, that is faster to run.

The studied metamodel is based on the hypothesis that the computer simulation is in fact the realization of a gaussian process indexed by the inputs and defined by a parametric mean function and a parametric covariance function. A small number of responses produced by the computer code are used to determine the values of the parameters of the mean and covariance functions. Given a new input, the predicted value is the expectation of the stochastic process at that input conditioned by the responses available. Since the stochastic process is gaussian, there is a formula for this expectation and the error of prediction.

When the precision of the output produced by the computer code can be tuned it is possible to incorporate responses with different levels of fidelity to enhance the prediction of the most accurate simulation at a new input while respecting the time constraints. This is usually done by adding several imprecise responses instead of a few precise ones. The main example for this type of computer experiments are the numerical solutions of differential equations. The precision can depend on the size of the mesh of the domain of resolution used to produce the response; on the space where the solution is projected or even weather a part of the physical model involved is left aside. The problem is how to take into account all the information available. This problem has been studied by many authors, most notably by LeGratiet in [1] and by Kennedy and O'Hagan in [3].

In the present work, we propose a new approch that is different from the existing ones.

For ease of notation only two precision or fidelity levels are considered: 1 for the least accurate and 2 for the most precise. First we will assume that the most precise level is a function of the least accurate. The difference between the two will be modeled by the gaussian process $Z_{(2,x)}$. If we suppose that $Y_{(1,x)}$ is the gaussian process related to 1, then $Y_{(2,x)}$ defined by equation (1) is also a gaussian process. It will model the outcomes of 2.

$$Y_{(2,x)} = \varphi(Y_{(1,x)}) + Z_{(2,x)} \tag{1}$$

Generalizing the results in [1], we propose a non-parametric approach where we compute a locally linear approximation of the function φ. We estimate the relationship and build a predictor by using all the responses to compute the conditional expected values for $Y_{(1,x)}$ and $Z_{(2,x)}$. The prediction error is built using the predicting errors of $Y_{(1,x)}$ and $Z_{(2,x)}$.

Then, we study an analog model based in [2] where the difference between the two levels is no longer a gaussian process. This time the difference between the two computer simulations will be modeled by

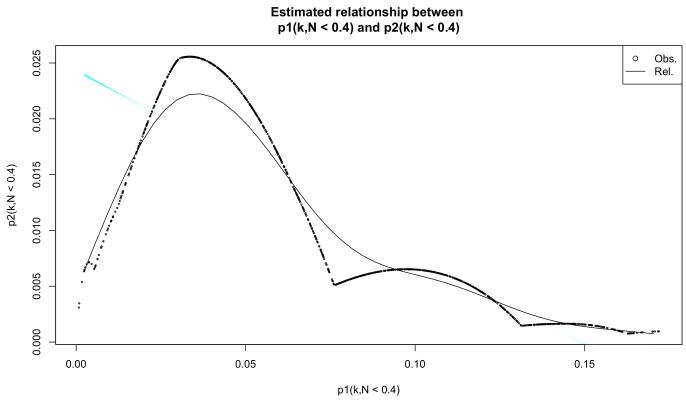

Figure 1: Estimated relationship between two successive levels of a computer code that simulates the pressure transient in a porous media.

the correlated errors ϵ_y. The correlation structure of the errors will depend on the distance between the outputs of 1. The new probabilistic model for the second simulator is given by equation (2) where $Y_{(1,x)}$ is still the gaussian process related to 1.

$$Y_{(2,x)} = \varphi(Y_{(1,x)}) + \epsilon_y \tag{2}$$

Once again we will estimate φ by using locally linear polynomials. Since we considered a particular correlation structure for the errors, we use the algorithm described by Fernandez in [2] to correct the bias in the estimation of the smoothing parameter of the non-parametric regression.

Finally, the two models are tested to illustrate their advantages and shortcomings. First by simulating the computer codes as gaussian processes we find that assuming that φ is linear when it is not can affect the results of the predictions. By using physical models we notice that the relationship between two fidelity levels of a computer code can be non-linear - as shown in Figure 1 - and in some cases not even function-like. Then, we develop briefly a case study related to a diphasic air-water flow in a rectangular domain.

References

[1] Le Gratiet, L. Recursive co-kriging model for Design of Computer experiments with multiple levels of fidelity with an application to hydrodynamic, arXiv:1210.0686; 2012.

[2] Fernandez, F and Opsomer, J. Smoothing Parameter Selection Methods for Nonparametric Regression with Spatially Correlated Errors. Canad. J. Statist., 33(2): p.279-295; 2004.

[3] Kennedy, M and O'Hagan, A. Predicting the output from a complex computer code when fast approximations are available. Biometrika. 87(1): p.1-13; 2000.

Short biography – Federico Zertuche is a third year PHD student at the Laboratoire Jean Kuntzmann (Université Joseph Fourier, Grenoble) under the supervision of Céline Helbert and Anestis Antoniadis. He is part of the INRIA team MOISE whose main research theme is the development of mathematical methods for modeling environmental phenomena.