A new approach to sensitivity analysis based on PLS regression

Jean-Pierre Gauchi

A NEW APPROACH TO SENSITIVITY ANALYSIS BASED ON PLS REGRESSION

J.-P. Gauchi
National Institute for Agronomical Research (INRA), France
Jean-Pierre.Gauchi@jouy.inra.fr

In this presentation, we propose the use of the Partial Least Squares (PLS) regression in order to more effectively carry out the Sensitivity Analysis (SA). This regression method is very well-known for data analysis in many different scientific fields (chemometrics, biometrics, etc). It was proposed by [1] and is explained in detail in [2] and [3]. Two advantages are particularly relevant for conducting a SA of a model ouput. These two advantages are: *i*) the very efficient way to manage the stochastic and structural dependences – because the partial covariances are taken into account – between the inputs; *ii*) the possibility of having a smaller number of simulations (because no matrix inversion is needed for estimating the SA indexes) than the number of inputs, which is extremely useful if simulations are very time-consuming. More information on the principle and properties of PLS regression will be given in the lecture and in a subsequent paper. We only provide the main steps of our methodology below.

The general methodology we propose is composed of four steps:

(a) N Monte Carlo simulations of the output are generated via a computer model, which leads to a simulation matrix S of N rows and p columns (the p inputs). It should be observed that the correlation structure between the p inputs is obtained by application of the method given in [4].

(b) A full quadratic polynomial model (p linear effects, p quadratic effects, $p(p\text{-}1)/2$ first-order interactions between inputs, and one intercept) is built from the p inputs, leading to a matrix M of N rows and k columns.

(c) A particular method of stepwise PLS regression – the BQ method described in [5] – is used for selecting the significant and significant expanded inputs. Even if the value of k is very large (2000, for example), the procedure works well and quite rapidly.

(d) A final PLS regression model is estimated (by means of SIMCA software Version 9.0, Umetrics AB, Sweden) with the inputs selected in step (c). If its R^2 is large enough (typically > 80%), we can consider that this final model is valid and hence provides estimated centred and scaled PLS coefficients, which can be seen as SA indexes (see Fig. 1 of the following example). Eventually, the adequate normalisation can be applied to these indexes for obtaining percentages.

Following are some results about a successful application to a real SA problem. This application is concerned with the exposure to the mycotoxin Ochratoxin-A (OTA) in food, for the population of French children. An elementary exposure to OTA is defined by the product of a food consumption (normalised by the individual weight) by the contamination level of this food. A global exposure is the sum of several (eight here) elementary exposures. The exposure distribution was estimated in [6], as well as its 95th quantile for estimating risk assessment exposure to OTA in food. A first SA was reported in [7] and at the SAMO 2001 Conference. The second SA we propose here is easier to achieve thanks to the PLS regression, and especially, to the fact that the whole variation domain of the 32 inputs can be taken into account, to the contrary of the study in [7] where Fig. 6 clearly shows the ellipsoidal domain of an input couple. The output we are interested in is thus the 95th quantile relative to the parameters of the probability density functions (*pdf*, the inputs of the SA) of the consumption and contamination distributions of the eight types of food. Indeed, these parameters are not certain and their potential ranges were estimated in [7] from real consumption and contamination data. Therefore, it is crucial to quantify the sensitivity of this high quantile to the variation of these inputs. In this case, we have $p = 32$ and $k = 561$. One trial was achieved with $N = 318$ (note that $N < k$), and the second with $N = 12{,}698$. The SA indexes are very similar for these values of N. However, we only show the significant SA indexes for $N = 12{,}698$ ($100\text{x}R^2 = 96\%$) in Fig.1.

Some brief comments can be made here. First of all, only six SA indexes are significantly different from zero among the 560 indexes; the word “significantly” has a particular meaning in the PLS that will be explained in the lecture and the subsequent paper. Secondly, we observe that the type of food “CEREALS” (see a detailed definition of this word in [6]) is the only type of food that is involved in the SA and, moreover, the SA indexes relative to the parameters of the contamination distributions are preponderant. Thus, it is of particular importance to have accurate values for these parameters and, consequently, we need to improve the collecting process of contamination data for “CEREALS”.

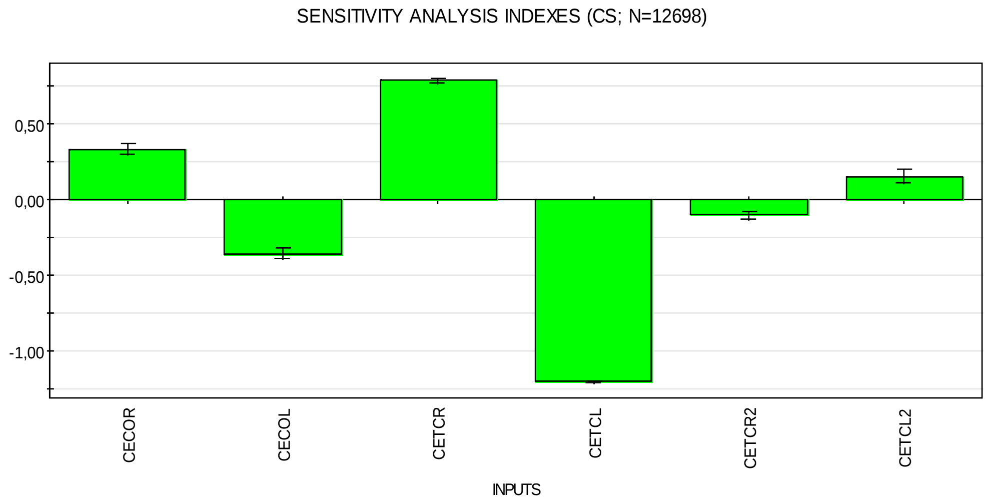

Fig. 1: SA indexes for the 95th quantile with their bootstrap type confidence intervals; inputs (all relative to "CEREALS") are: cecor = the shape parameter of the consumption Gamma *pdf*, cecol = the scale parameter of the consumption Gamma *pdf*, cetcr = the shape parameter of the contamination Gamma *pdf*, cetcl = the scale parameter of the contamination Gamma *pdf*, and quadratic CETCR and CETCL terms.

References

[1] Wold S, Martens H, Wold H: The multivariate calibration problem in chemistry solved by the PLS method. *In*: Proc. Conf. Matrix Pencils, Ruhe A, Kastrom B. (Eds.), Lecture Notes in Mathematics, Springer, Heidelberg, 1983, pp. 286-293.

[2] Garthwaite PH: An interpretation of partial least squares. *J. Amer. Statist. Assoc.* **89**, 425, 122-127 (1994)

[3] Tenenhaus M: La régression PLS: théorie et pratique. Technip, Paris (1998)

[4] Iman RL, Conover WJ: A distribution-free approach to inducing rank correlation among input variables. *Commun. Statist.-Simula. Computa.* **11**, 3, 311-334 (1982)

[5] Gauchi JP, Chagnon P: Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. *Chemometrics & Intelligent Laboratory Systems* **58**, 2, 171-193 (2001)

[6] Gauchi JP, Leblanc JC: Quantitative Assessment of Exposure to the Mycotoxin Ochratoxin-A in Food. *Risk Analysis* **22**, 2, 219-234 (2002)

[7] Albert I, Gauchi JP: Sensitivity analysis for high quantiles of Ochratoxin-A exposure distribution. *Int. J. of Food Microbiology* **75**, 2, 143-155 (2002)