5th SAMO Conference

Budapest, 18 - 22 june 2007

A NEW APPROACH TO SENSITIVITY ANALYSIS BASED ON PLS REGRESSION

Jean-Pierre Gauchi

INRA / Unité MIA (UR341), Jouy-en-Josas, France

PLAN

1. **Introduction**

2. **PLS regression background**

3. **Outlines of the new SA approach**

4. **An application in risk assessment in food**

5. **Conclusion**

1. Introduction

- The proposition: the use of the Partial Least Squares Regression (PLSR) (Wold et al., 1983, applications in chemometrics, biometrics, etc) in order to carry out the (Global) Sensitivity Analysis in case of highly correlated inputs and moderate nonlinear behaviour.

  $\longrightarrow$ Classical definition of a Sensitivity Index:

$$SI_j = V\left(E\left(Y|X_j\right)\right)/V\left(Y\right) \underset{if\ linear\ relation}{=} cor^2\left(Y, X_j\right)$$

- Advantages of the PLSR that lead to new SI_j

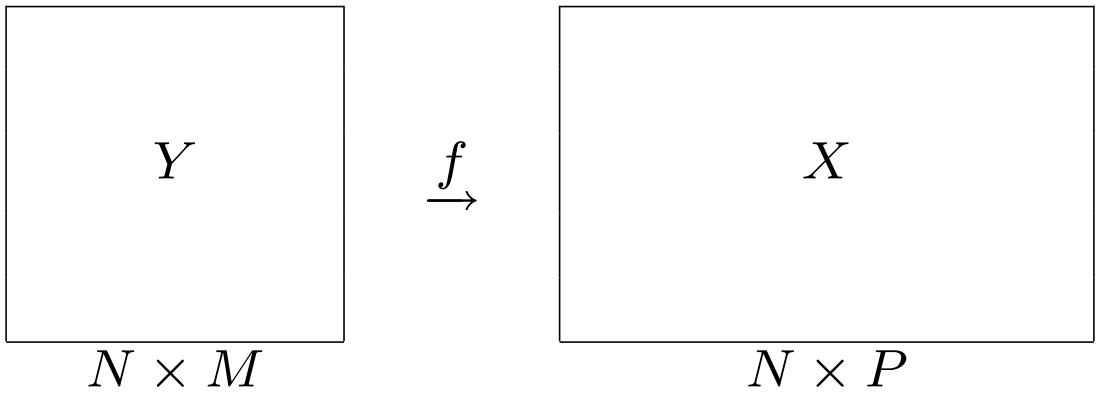
  – **the partial covariances between inputs and output are taken into account**,

  – **the number of simulations can be less than the number of inputs.**

2. PLS regression background

2.1 The formal PLS model

Let us consider the multivariate regression model: $Y = f(X)$



where X is a $N \times P$ matrix composed of P controlled inputs X_j fixed at N levels and Y is a $N \times M$ matrix composed of the M outputs Y_k observed at these N levels; f is a set of M polynomial functions of x_j to be estimated via H **crossvalidated** latent components t_h.

2.2 Algorithm of PLS1 $(M = 1)$ regression

Step 1

Step1.1

$$\underset{(N\times 1)}{t_1} = E_0 w_1 = Arg\left\{\underset{t=E_0 w\ ;\ \|w\|=1}{\max}\left[cov\left(t, y_0\right)\right]\right\}$$

where E_0 and y_0 be the centered and scaled X and y, respectively.

$\Longrightarrow$As $cov\left(t_1, y_0\right) = \sqrt{var\left(t_1\right)} \times cor\left(t_1, y_0\right)$

$\Longleftrightarrow$

to maximize simultaneously the explained variance by the latent component t_1 and the correlation between the ouptut y and t_1:

$\Longrightarrow$ **it is a compromise between the ordinary multiple regression of y on $X_1, \ldots, X_P$ and the principal component analysis of the X matrix**

$\hookrightarrow$ **Solution :**

By the Lagrange multipliers method we find that w_1 is the eigen vector of the $(P \times P)$ matrix $E_0^T y_0 y_0^T E_0$ (associated with the largest eigen value):

$$\begin{aligned} w_1 &= \left(E_0^T y_0\right) / \left\|E_0^T y_0\right\| \\ &= \left(\sum_{j=1}^{P} cor^2\left(X_j, y\right)\right)^{-1/2} \times \begin{bmatrix} cor\left(X_1, y\right) \\ \vdots \\ cor\left(X_P, y\right) \end{bmatrix} \end{aligned}$$

Step1.2

Regressions of the E_{0j} and y_0 on t_1 :

$$\begin{aligned} E_{0j} &= p_{1j}t_1 + E_{1j} \\ y_0 &= r_1 t_1 + y_1 \end{aligned}$$

with $p_{1j} = E_{0j}^T t_1 / \|t_1\|^2$ and r_1 the scalar regression coefficients,

and $E_1 = (E_{11}, \ldots, E_{1P})^T$, y_1 the "residual" matrice and vector.

$\Longrightarrow$

$$\begin{aligned} E_0 &= t_1 p_1^T + E_1 \\ y_0 &= r_1 t_1 + y_1 \end{aligned}$$

where $p_1 = (p_{11}, \ldots, p_{1P})^T$.

Step 2

Step2.1

$E_0 \longrightarrow E_1$; $y_0 \longrightarrow y_1$

$\Longrightarrow$

$$\begin{aligned} w_2 &= \left(E_1^T y_1\right) / \left\|E_1^T y_1\right\| \text{ under the constraint } w_2 \perp w_1 \\ &= \left(\sum_{j=1}^{P} cov^2\left(E_{1j}, y_1\right)\right)^{-1/2} \times \begin{bmatrix} cov\left(E_{11}, y_1\right) \\ \vdots \\ cov\left(E_{1P}, y_1\right) \end{bmatrix} \end{aligned}$$

that leads to the second PLS component:

$$t_2 = E_1 w_2 = E_0 \left(I - w_1 p_1^T\right) w_2 = E_0 w_2^*$$

Step2.2

We achieve the new regressions that lead to the decompositions:

$$\begin{aligned} E_1 &= t_2 p_2^T + E_2 \\ y_1 &= r_2 t_2 + y_2 \end{aligned}$$

where p_2 and r_2 are the corresponding regression coefficients (vector and scalar).

- Etc. for the following steps

$\hookrightarrow$ **Final results** (at step H with H crossvalidated **orthogonal** t_h)

$$
\begin{aligned}
E_0 &= t_1 p_1^T + \ldots + t_H p_H^T + E_H \\
\|E_0\|^2 &= \|t_1\|^2 \|p_1\|^2 + \ldots + \|t_H\|^2 \|p_H\|^2 + \|E_H\|^2
\end{aligned}
$$

$$
\begin{aligned}
y_0 &= r_1 t_1 + \ldots + r_H t_H + y_H = \sum_{h=1}^{H} \frac{sign(r_h) SI_h^{1/2}}{\sigma_h} t_h + y_H \\
y_0 &= \sum_{h=1}^{H} r_h \left(\sum_{j=1}^{P} w_{hj}^* E_{0j} \right) + y_H = \sum_{j=1}^{P} \left(\sum_{h=1}^{H} r_h w_{hj}^* \right) E_{0j} + y_H
\end{aligned}
$$

$$
y_0 = \hat{\beta}_1^{PLS} E_{01} + \ldots + \hat{\beta}_P^{PLS} E_{0P} + y_H
$$

$\hookrightarrow$The following normalized PLS coefficients $\hat{\beta}_j^{PLS}$ can be seen as **new (signed)** SI_j:

$$SI_j = 100 \times \frac{\hat{\beta}_j^{PLS}}{\sum_{j=1}^{P} \left|\hat{\beta}_j^{PLS}\right|}$$

This proposition is motivated by the result:

$$det\left(Varcov\left(\hat{\beta}^{PLS}\right)\right) << det\left(Varcov\left(\hat{\beta}^{OLS}\right)\right)$$

3. Outlines of the new SA approach

- **STEP 1**: N Monte Carlo simulations of the output y are generated (the "optimal" aspect of the simulation design is not considered here) via a "computer model" based on the distributions of P independent or correlated inputs X_j.

- **STEP 2**: A full quadratic polynomial model is built from the P inputs X_j and then y, and the K inputs and expanded inputs are centered-scaled.

- **STEP 3**: A **stepwise** $PLSR$ (BQ method, Gauchi & Chagnon, 2001) $\Rightarrow$ selecting the K^* significant inputs and significant expanded inputs by observing the evolution of a PLS specifical statistics named the Q^2_{cum}.

- **STEP 4**: A final $PLSR$ model is estimated with the K^* inputs.

  If $100 \times R^2 \geq 80\%$, we consider that this final model is acceptable and the K^* new SI_j are computed.

  If $100 \times R^2$ is too low, we declare that the approach does not work (we advice not to raise the degree of the polynomial in order to keep a "physical" interpretation of the inputs associated to the SIs) and we propose a PLS extended approach (see conclusion).

4. An application in risk assessment in food

Exposure to the mycotoxin Ochratoxin-A (OTA) in food (French children population)

- An elementary **theoretical** exposure to OTA is defined by the product of a i food consumption (normalised by the individual weight) by the contamination level of this food: $E_i = C_i T_i$.

- A global **theoretical** exposure is the sum of several (eight foods were studied here) elementary exposures: $E_G = \sum_{i=1}^{8} C_i T_i$.

- An **estimation** of the $\tilde{E}_G$ random variable distribution was obtained, as well as its **95th quantile** for evaluating the risk assessment exposure to OTA in food (Gauchi & Leblanc, 2002), and a previous SA approach was reported (Albert & Gauchi, 2002)

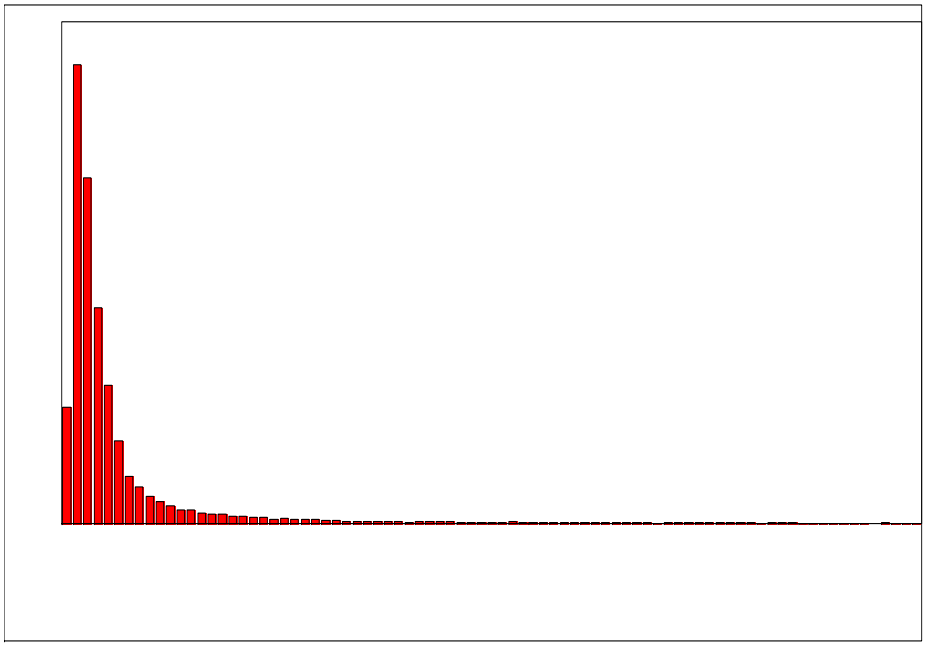


Figure 1: Exposure output relative histogram for children (unit=ng$\times$bwkg$^{-1}\times$day^{-1})

- We propose here a second SA where the whole variation domain of the 32 inputs can be taken into account, to the contrary of the previous study:

  $\Longrightarrow$**The** $32 = 8 \times (2 + 2)$ **correlated inputs are the estimated scale (L) and shape (R) parameters of the Gamma distributions fitted to the consumption and contamination histograms of the eight foods** (Indeed, the inputs depend on the collected data during the survey, and their potential ranges were estimated from real consumption and contamination data).

  $\Longrightarrow$**The output we are interested in is the 95th quantile** of the $\tilde{E}_G$ distribution for quantifying **its sensitivity to** the variation of the K $(= 560)$ inputs and expanded inputs.

Two trials were achieved with $N = 318$ (note that $N < K$), and $N = 12,698$: the significant selected inputs are the same (except CETCR2) for these values of N.

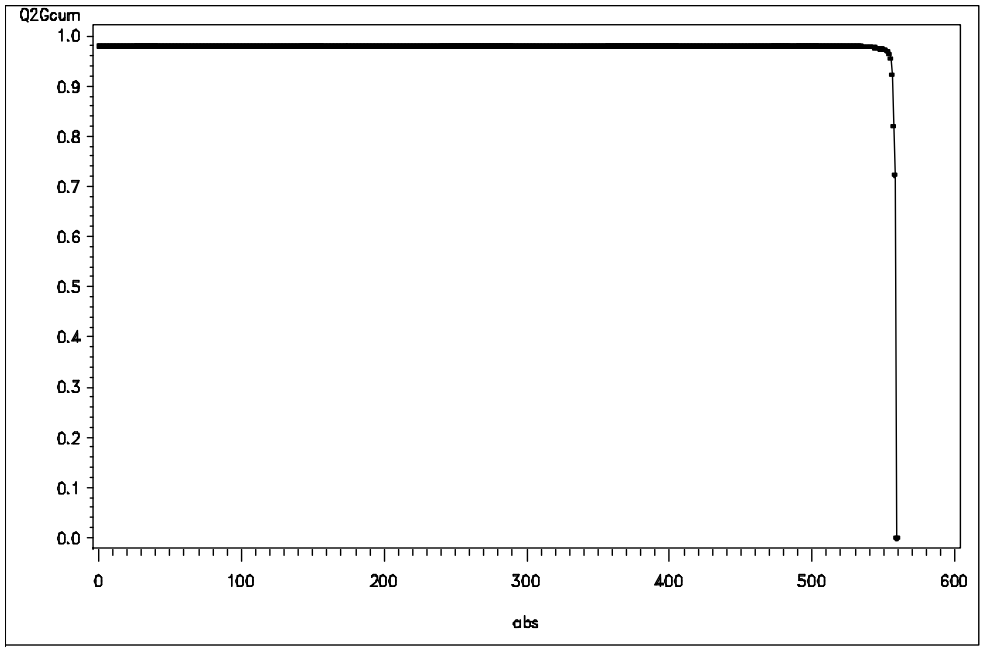


Figure 2: The Stepwise PLS regression (BQ method): Evolution of the Q2cum versus the eliminated inputs.

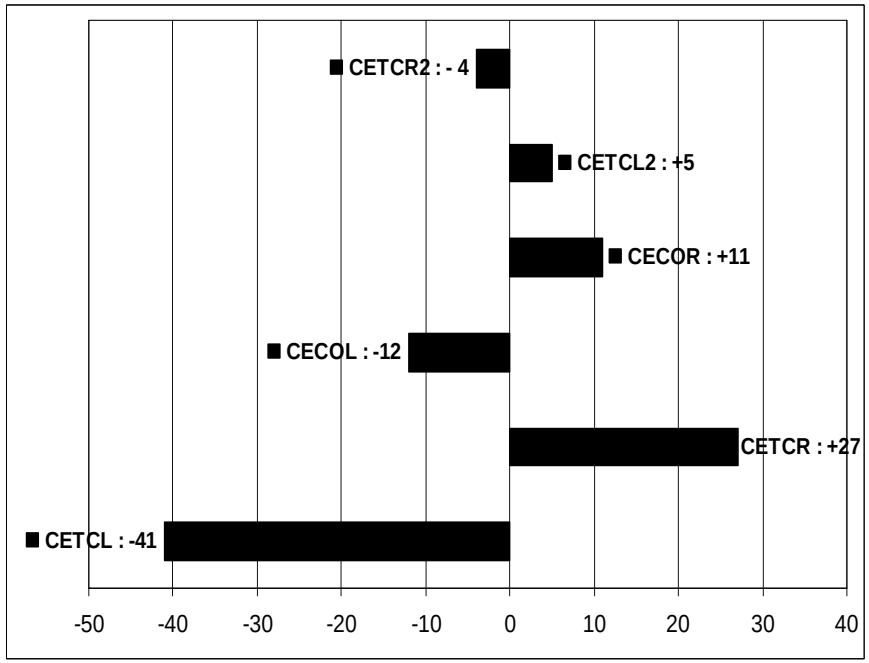


Figure 3: The $K^* = 6$ new significant SIs (for $N = 12,698$); The final PLS model has a $100 \times R^2 = 95\%$.

Comments:

- Only six SIs are significant (in the sense of stepwise PLSR-BQ and also in a bootstrap sense) among the 560 SIs.

- The type of food "CEREALS" is the only type of food that is involved in the SA and, moreover, the SIs relative to the parameters of its **contamination** distribution are preponderant.

  $\Longrightarrow$ 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".**

5. Conclusion

- PLSR (no matrix inversion) can contribute to improve the SA, especially **when the inputs are highly correlated and/or the number of simulations is less than the number of inputs.**

- Another important possibility (technical report and paper in progress) is:

  Taking into account qualitative inputs via the 0/1 coding of their levels and mixture of quantitative and qualitative inputs.

- Finally, I wish that this approach be compared to other SA methods.

- Future work: kernel (RBF) PLSR for strong nonlinear behaviour.