WAVELET-PLS REGRESSION: Application to Oil Production Data

Salwa BenAmmou, Zied Kacem, Hédi Kortas and Zouheir Dhifaoui

Computational Mathematics Laboratory

FACULTE DE DROIT ET DES SCIENCES ECONOMIQUES ET POLITIQUES DE SOUSSE

Introduction

- Statisticians are often confronted to several problems such as missing or incomplete data, the presence of a strong collinearity between the explanatory variables or the case where the number of variables exceeds the number of observations.

- The PLS method has been proposed by WOLD in the 80's to cope with these problems.

Introduction

In practical applications, however, we are confronted with the problem of noise affecting the dataset.

Actually, the noise component can strongly affect the adjustment quality and the predictive performance of the PLS model.

Objective

We propose an hybrid data analysis method based on the combination of wavelet thresholding techniques and PLS regression in order to remove or attenuate the effect of the noise.

Wavelet Theory: Multiresolution Analysis (MRA)

A MRA is a sequence $V_j, j \in \mathbb{Z}$ of closed subspaces of $L^2(\mathrm{IR})$ satisfying:

$(i)\quad \forall j \in Z : V_j \subset V_{j+1}$

$(ii)\quad \forall j \in Z : f \in V_j \Leftrightarrow f(2.) \in V_{j+1}$

$(iii)\quad \bigcap_{j \in Z} V_j = \{0\}$

$(iv)\quad \overline{\bigcup_{j \in Z} V_j} = L^2(\Re)$

(v) *There exists a function* $\varphi \in V_0$ *such that* $\{\varphi(.-k) : k \in Z\}$ *is an O.N.B of* V_0;
φ *is called scaling function.*

Wavelet Theory: Basic concepts

- The scaling function is such that:
  $\left\{ \varphi_{jk} := 2^{j/2}\varphi\left(2^{j}.-k\right) : k \in Z \right\}$ is an ONB of $V_j, j \in \mathbb{Z}$

Let W_j be the orthogonal complement of V_j in V_{j+1}

- If there exists a function $\psi \in W_0$ such that $\{\psi(.-k) : k \in Z\}$ is an ONB in W_0

ψ is called wavelet function and satisfy:

$\left\{ \psi_{jk} := 2^{j/2}\psi\left(2^{j}.-k\right), k \in Z \right\}$ is an ONB in $W_j, j \in \mathbb{Z}$

Wavelet Theory: Basic concepts

Thus a function has a unique representation in terms of a convergent series in L_2 :

$$f(x) = \sum_k \alpha_k \varphi_{0k}(x) + \sum_{j=0}^{\infty} \sum_k \beta_{jk} \psi_{jk}(x) \qquad (1)$$

where $\alpha_k = \int f(x)\overline{\varphi_{0k}(x)}dx$ and $\beta_{jk} = \int f(x)\overline{\psi_{jk}(x)}dx$

Wavelet thresholding

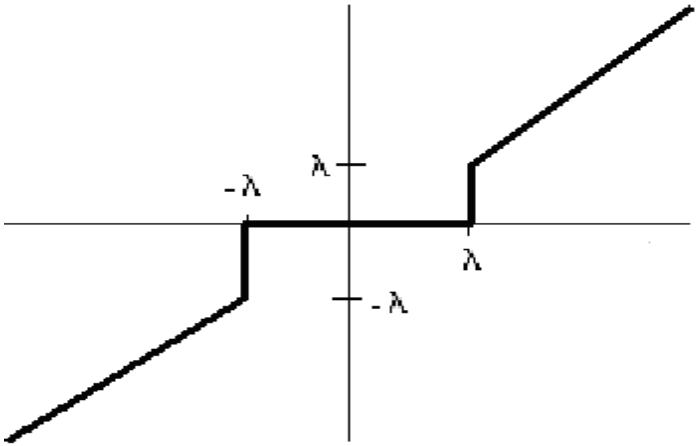
The thresholding strategy consists in three steps:

- Apply the DWT decomposition to the observed data sequence to produce a set of scale-wise approximation and detail coefficients.
- Keep the detail coefficients β_{jk} which are above a fixed threshold level and set to zero the coefficients which are below the threshold.
- Reconstruct the signal

Thresholding techniques

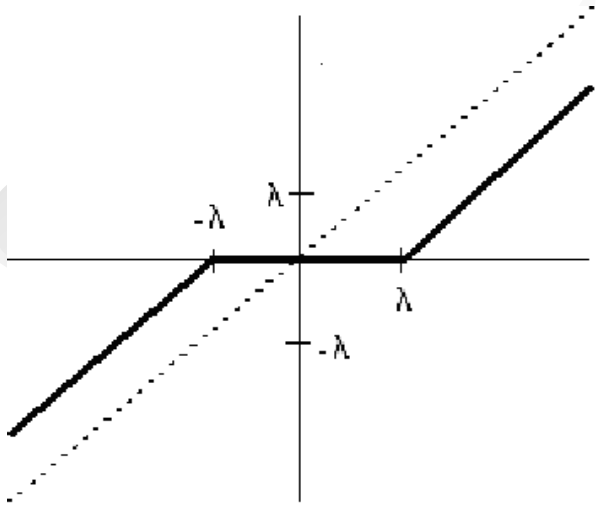
- Hard thesholding:

$$\theta(x)=\begin{cases}0 & \text{si}\,|x|<\lambda\\ x & \text{si}\,|x|\geq\lambda\end{cases}$$



- Soft thesholding:

$$\theta(x)=\begin{cases}0 & si\,|x|<\lambda\\ x-sign(x)\lambda & si\,|x|\geq\lambda\end{cases}$$



- The linear wavelet estimator $\hat{f}_{jl}$ of the function f is given by:

$$\widehat{f_{J_1}}(x) = \sum_k \widehat{c_{J_0 k}} f_{J_0 k}(x) + \sum_{J=J_0}^{J_1} \sum_k \widehat{d_{jk}} y_{jk}(x) \quad (2)$$

$$\hat{c}_{jk} = \frac{1}{n}\sum_{i=1}^{n} Y_i \varphi_{jk}(X_i) \quad \text{et} \quad \hat{d}_{jk} = \frac{1}{n}\sum_{i=1}^{n} Y_i \psi_{jk}(X_i)$$

$\hat{d}_{jk}$ are the thresholded wavelet detail coefficients

$\hat{c}_{jk}$ are the thresholded approximation coefficients

PLS regression (Partial Least Squares Regression)

- PLS regression (PLS) is a nonlinear model linking a set of dependent variables Y to a set of numerical or categorical explanatory variables X.

- It is often utilized to handle highly correlated regressors

- It is of great interest when dealing with data sets in which the number of predictors greatly exceeds the number of observations.

- It allows to deal with the problem of missing data.

PLS1 regression

PLS univariate regression (PLS1) is a nonlinear model linking a dependent variable to a set of numerical or categorical explanatory variables .

- The PLS1 regression algorithm involves several steps:
  - Construction of the first PLS component t_1

$$t_1 = w_{11}X_1 + \ldots + w_{1k}X_k$$

  - Normalisation of the coefficients w^*_{1j}

$$w^*_{1j} = \frac{w_{1j}}{\sqrt{\sum_{j=1}^{k}(w_{1j})^2}} \quad (3)$$

PLS1 regression

- Perform an OLS regression of Y on t_1

$$\hat{Y} = c_1 t_1 + Y_1$$

regression coefficient

residuals

Therefore:

$$\hat{Y} = c_1 w_{11} X_1 + \ldots + c_1 w_{1k} X_k + Y_1$$

If the model has limited explanatory power, we search for a second component which is not correlated with and is able to explain the residual vector quite good.

PLS1 Regression

- t_2 can be written as: $t_2 = w_{21}x_{11} + \ldots + w_{2k}x_{1k}$

- We perform a multiple regression of Y on t_1, t_2:

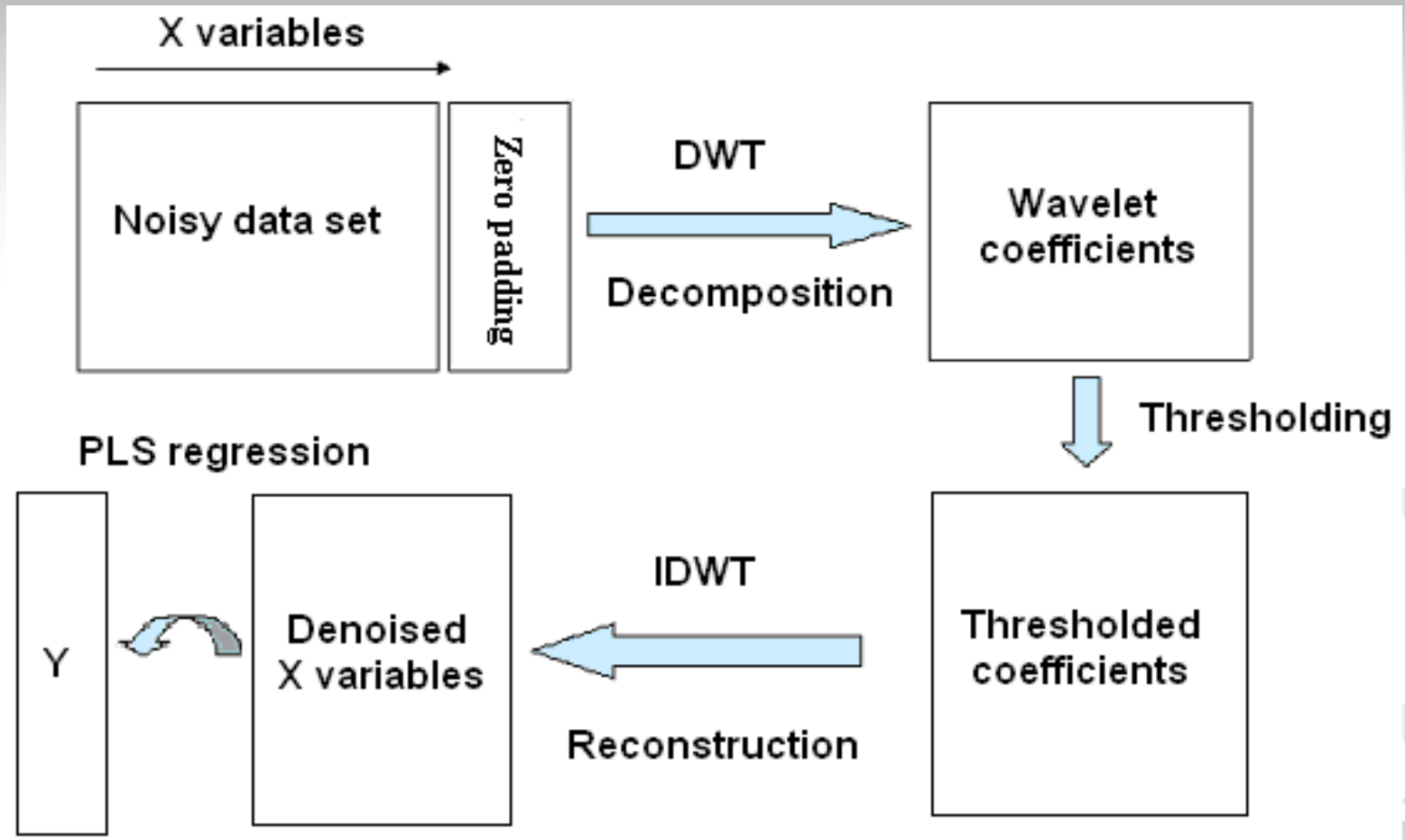
$$\hat{Y} = c_1 t_1 + c_2 t_2 + Y_2$$

regression coefficients

residual vector

- The number of components t_h to be retained is determined by cross validation

Wavelet-PLS


X variables
Noisy data set
Zero padding
DWT
Decomposition
Wavelet coefficients
Thresholding
PLS regression
Y
Denoised X variables
IDWT
Reconstruction
Thresholded coefficients

Application

- The response variable Y: the crude oil (petroleum) daily production in barrels in a given oil field composed of four wells during the period from May 1, 2003 to March 31, 2006 i.e.1024 observations.

- The data measurements are made on a daily basis
- The response variable Y depends on 16 explanatory variables:

  - ✓ Choke i : the choke valve position in the oil well i; i = 1,…, 4.

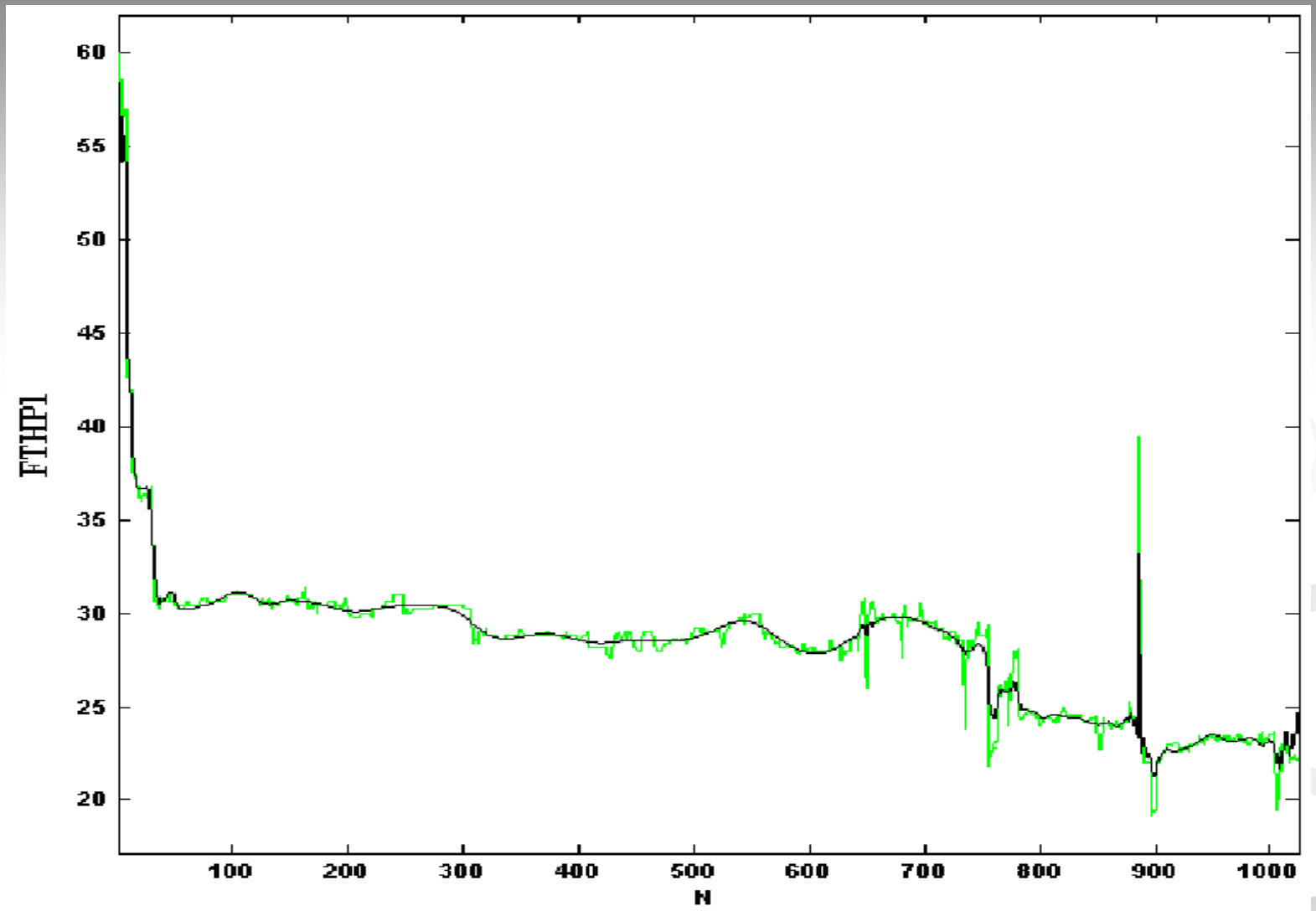
  - ✓ FTHP i : Flowing Tubing Head Pressure of the well i (in Bars); i = 1,…, 4.

- Pres at Choke i : pressure on the level of the choke in the well i (in bars); i = 1,…, 4.

- WCi : (Water cut) Percentage of water. It is the ratio of water produced to the volume of total liquids extracted from the well i; i = 1,…, 4.

Wavelet thresholding set-up

- We use a Daubechies compactly supported wavelet with 5 vanishing moments.
- The Discrete wavelet Transform is curtailed at scale j=5
- We opt for a soft thresholding



Signal before (green) and after thresholding (black)

Specification of the number of components by cross validation

PLS1

Number of components	Q^2_h	limits
1	0.742	0.0975
2	0.319	0.0975
3	0.393	0.0975
4	0.186	0.0975
5	0.0663	0.0975

Wavelet-PLS

Number of components	Q^2_h	limits
1	0.734	0.0975
2	0.287	0.0975
3	0.237	0.0975
4	0.266	0.0975
5	0.038	0.0975

- The PLS1 equation before thresholding:

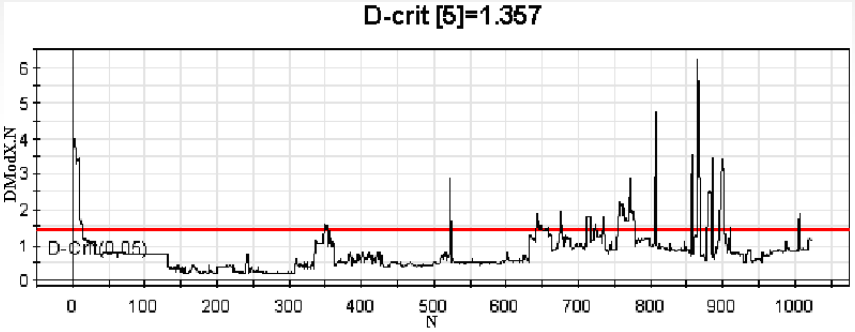
$$\hat{y} = 0{,}14745477\, x_1 + 0{,}12351255\, x_2 + 0{,}29458188\, x_3 + 0{,}16206525\, x_4 - 0{,}27695889\, x_5 + 0{,}03891265\, x_6 - 0{,}1728005\, x_7 - 0{,}14108841\, x_8 + 0{,}28230372\, x_9 + 0{,}27352113\, x_{10} + 0{,}23676341\, x_{11} + 0{,}08288938\, x_{12} + 0{,}01417857\, x_{13} - 0{,}19398681\, x_{14} - 0{,}00272167\, x_{15} + 0{,}00767741\, x_{16}$$

- The PLS1 equation after thresholding:

$$\hat{y} = 0{,}076750638\, x_1 + 0{,}073312704\, x_2 + 0{,}314558779\, x_3 + 0{,}116011568\, x_4 - 0{,}268962544\, x_5 + 0{,}002680218\, x_6 - 0{,}124656262\, x_7 - 0{,}254468339\, x_8 + 0{,}338198727\, x_9 + 0{,}317734483\, x_{10} + 0{,}277136406\, x_{11} + 0{,}053406536\, x_{12} - 0{,}028291771\, x_{13} - 0{,}13101302\, x_{14} - 0{,}028039241\, x_{15} - 0{,}01063908\, x_{16}.$$

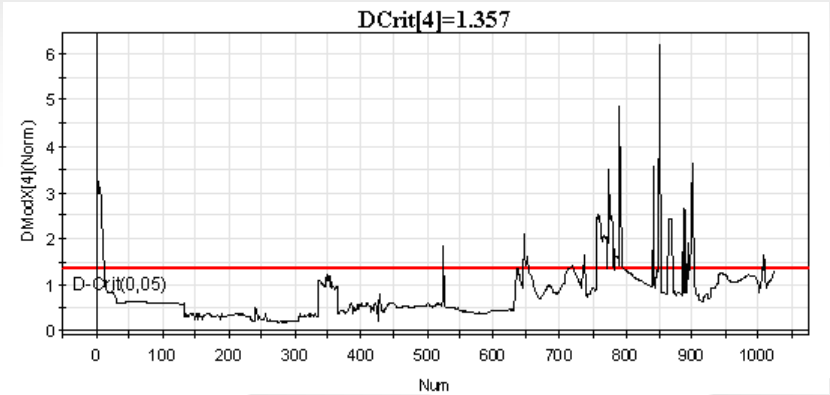
Outliers

PLS1 before thresholding



9.6% of the total sample are regarded as outliers

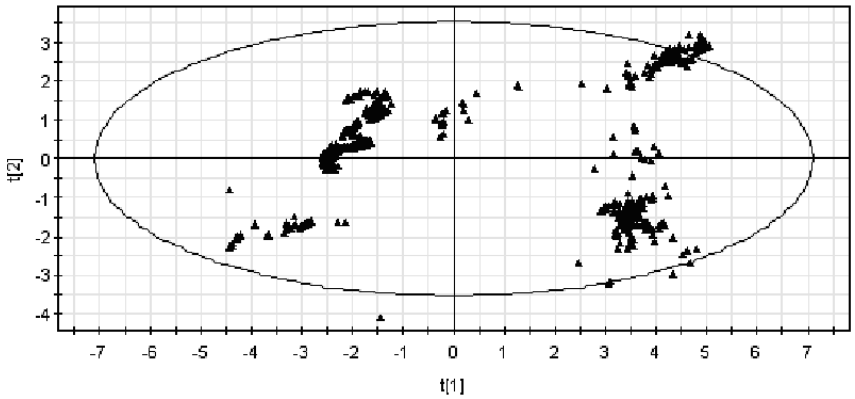
PLS1 after thresholding



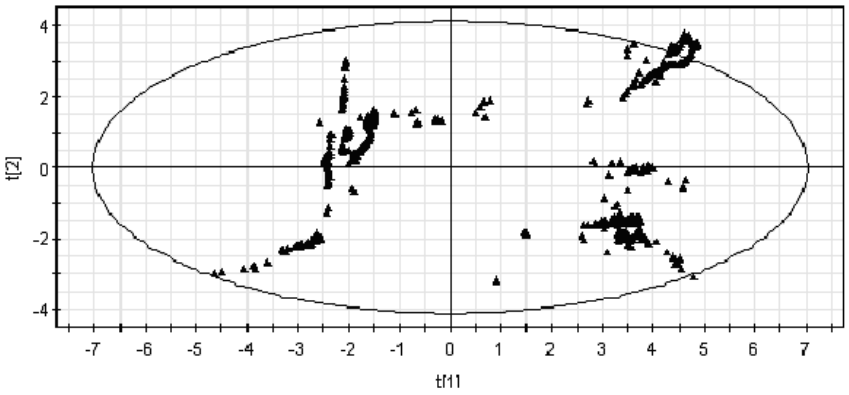
8.7% of the observations are regarded as outliers

Confidence ellipsoids

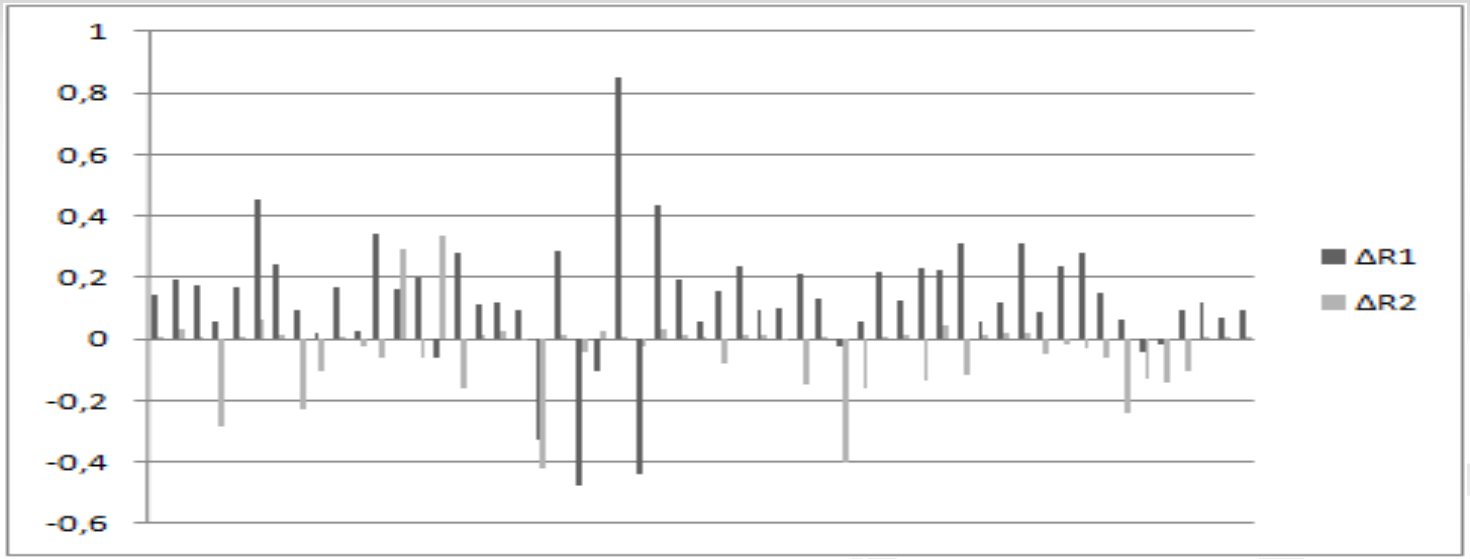
Raw data



Denoised data



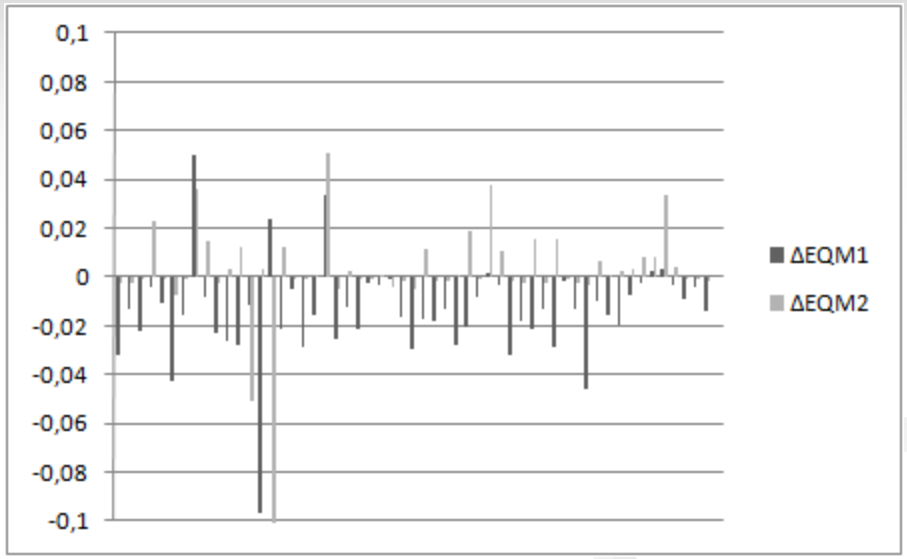
Goodness of fit



The ΔR_1^2 values are much closer to zero than the ΔR_2^2.

This shows the effectiveness of the wavelet techniques for noise removal.

Mean Squared Errors



It is clear that the $\Delta MSE2$ *are much smaller than* those of $\Delta MSE1$

⟶ *This confirms the relevance of the Wavelet-PLS method.*

Conclusion

The Wavelet-PLS approach allowed us to:

- reduce the number of outliers
- reduce the Mean Square Error
- correct the observations in the score plot
- ameliorate the goodness of fit of the model

Thanks