Wavelet-PLS Regression: Application to Oil Production Data

Salwa Benammou[1], Kacem Zied[1], Hedi Kortas[1], and Dhifaoui Zouhaier[1]

[1] Computational Mathematical Laboratory, *saloua.benammou@yahoo.fr*
[2] *ZiedKacem2004@yahoo.fr*
[3] *kortashedi@yahoo.fr*
[4] *dh.zouhaier@yahoo.fr*

Abstract. This paper is devoted to the study of PLS regression in the presence of noise that could affect the quality of the results. To solve this problem, we suggest a hybrid approach which combines PLS regression and wavelet-based thresholding techniques. The proposed method is validated via a simulation study and subsequently applied to petroleum data. Empirical results show the relevance of the selected approach and contribute to a better modelling of the series of study.

Keywords: PLS regression, thresholding, minimax, wavelet-PLS

1 Introduction

In numerous data analysis applications, statisticians are confronted with 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. To cope with these problems, several statistical approaches have been developed, among them, a data analysis method initially proposed by Wold and al. (1983). It is known as Partial Least Squares (PLS) regression.

Although PLS regression has proven to be of great performance in a wide range of applications, the model variables are usually corrupted by noise which may adversely affect the results drawn from the PLS regression in terms of modelling and prediction accuracy (Tenenhaus and al. (1995)).

To deal with this problem, we discuss, in this paper, a hybrid data analysis method based on the combination of wavelet thresholding techniques and PLS regression.

2 The Wavelet-PLS method

The Wavelet-PLS regression entails several steps. As a first step, we preprocess the variables in the following manner: if the explanatory data vectors are not of dyadic lengths (i.e. powers of 2), we extend the data samples by applying a so called "zero-padding" method. This method consists in adding zeros to the beginning and/or end of each time domain sequence in order to attain the next dyadic length. This pre-processing step is needed for the implementation of the Discrete Wavelet Transform (DWT) algorithm (Mallat (2000)) which requires a dyadic length time series. It should be stressed here that the wavelet coefficients relating to the zero-padding operation are subsequently eliminated while performing the Inverse Discrete Wavelet Transform (IDWT) signal reconstitution.

In the second stage, the DWT is performed to the exogenous variables. This requires a precise choice of the wavelet system to be used. In this work, we rely on Daubechies wavelet bases possessing attractive properties such as vanishing moments, orthogonality and especially support compactness which results in significant computational gains (Daubechies (1992)).

In the third step, the obtained wavelet coefficients are thresholded by means of the wavelet-based denoising techniques (Donoho and Johnstone (1998)). In the fourth step, we carry out an IDWT to reconstruct the set of explanatory variables which are now practically noise-free. Finally, the conventional PLS regression is applied to the new set of regressors. The Wavelet-PLS approach is illustrated in Fig.1:

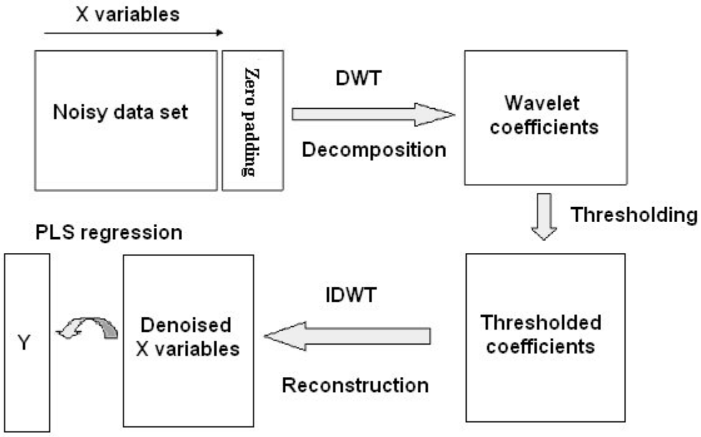

Fig. 1. Overview of the Wavelet-PLS regression.

3 Application

In order to assess the relevance of the Wavelet-PLS regression scheme, we consider a real world data set. The response (dependent) variable Y represents the crude oil (petroleum) production in barrels denoted ”oil” in a given oil field composed of four wells. The data measurements are made on a daily basis during the period from May 1, 2003 to March 31, 2006 thus totalling 1024 observations. Here the response variable Y depends on 16 explanatory variables corresponding to the features of the wells. The independent variables are:

- (Choke i), $i = 1, \ldots, 4$: the choke valve position in the oil well $i, i = 1, \ldots, 4$. This variable takes integer values ranging from 1 to 64. In fact, the choke valve is variably positionable defining 64 incremental positions allowing to regulate the flow rate.
- (FTHP i), $i = 1, \ldots, 4$: Flowing Tubing Head Pressure of the well i (in bars). Actually, oil extraction is assured by the difference between the underground pressure in the oil reservoir and the pressure at the top of the well. The pressure at the top of the well, which is the extraction pressure, is a key parameter and is defined as the Flowing Top Head Pressure (FTHP).
- (Pressure at Choke i), $i = 1, \ldots, 4$: pressure on the level of the choke in the well i (in bars).
- (WC i), $i = 1, \ldots, 4$: (Water cut) Percentage of water. It is the ratio of water produced to the volume of total liquids extracted from the well i.

4 PLS Regression on the raw data set

Using the cross validation technique, we retrain a PLS model with four components. The regression of y on t_1, t_2, t_3 and t_4 gives the following equation:

$$\hat{y} = (0.29716)t_1 + (0.199871)t_2 + (0.34454)t_3 + (0.167767)t_4$$

The normalized distances to the model in the X space are reported on Fig.2. Remember that the observations with $NDmodX$ exceeding the critical limit at the 95% are regarded as outliers in the X space.

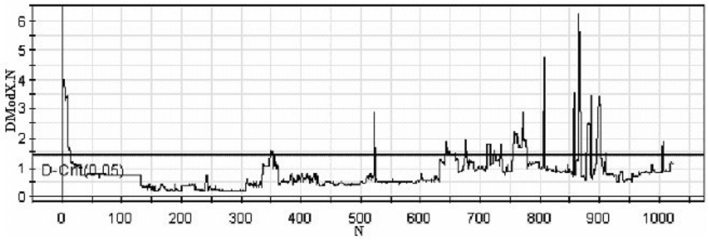

Fig. 2. Normalized distances to the model NDModX.

It is remarkable to note here that 98 observations i.e. 9.6 % of the total sample are regarded as outliers.

5 Wavelet-PLS regression results

5.1 Wavelet-based denoising

In order to eliminate the noise from the set of predictors, we first apply a DWT curtailed at the resolution level $j = 5$ to each exogenous variable using the $D(8)$ Daubechies compactly supported wavelet. The DWT results in five levels of wavelet detail coefficients and a single level approximation coefficients. Next, the obtained detail coefficients are subject to a wavelet thresholding operation. In our case, we opt for a soft thresholding function. The choice of the threshold value is done according to the Minimax procedure. This is due to the fact that the raw variables' series exhibit several discontinuities and abrupt changes. The Minimax procedure is well adapted for handling such features. It should be noted that associating the soft thresholding and the Minimax criterion defines the so called "Risk-Shrink" procedure (Donoho and Johnstone (1994)).

5.2 Wavelet-PLS regression on the denoised variables

In the following, we carry out a PLS regression using the obtained denoised regressors. According to the cross validation estimation results, we choose to retain four PLS components. The regression equation is then given by:

$$\hat{y} = (0.29816)t_1 + (0.165487)t_2 + (0.261839)t_3 + (0.319801)t_4$$

Estimation results for the Wavelet-PLS regression show a slight improvement for the determination coefficient with an R^2 value of 0.919 compared to $R^2 = 0.916$ for the PLS model performed on the raw data set.

Another interesting remark is that the Mean Square Error has decreased by 10.3% when carrying out a wavelet thresholding on the regressors' vectors. Fig.3 report the $NDmodX$ values for the Wavelet PLS procedure. Observe the effect of noise removal on the regression results.

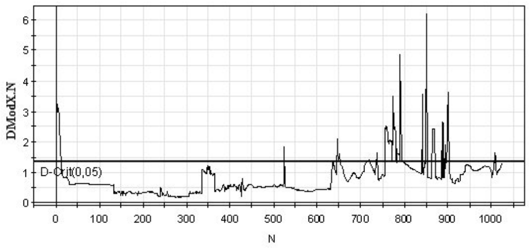

Fig. 3. Normalized distances to the model NDModX for the Wavelet-PLS regression.

The obtained results show that 90 observations among 1024 are outliers representing 8.7 % of the total sample size. Thus we can state that the denoising procedure has reduced the number of outliers yielding better modelling results.

6 Simulation

In this simulation study, we apply the Wavelet-PLS regression procedure to multidimensional fractional Brownian motions with noisy components.

In order to present the n-dimensional fractional Brownian, we restrict ourselves to the simple case $n = 2$. The generalization is straightforward.

The process $B_t = (B_t^1, B_t^2)^T$ is a correlated 2-dimensional fractional Brownian motion if:

- The increments $B_1(t) - B_2(t)$ et $B_2(t) - B_2(s), t > s$ are independent of $B_1(y)$ and $B_2(y), \forall 0 \leq y \leq s$.
- $cov(B_1(t), B_2(t)) = E(B_1(t)B_2(t)) = \rho t$ where $-1 \leq p \leq 1$. This implies that: $corr(B_1(t), B_2(t)) = \rho$. Besides, we have: $\forall t \neq s, cov(B_1(t), B_2(s)) = \rho min(t, s)$.

In this section, we synthesize 55 realisations of an n-dimensional fractional Brownian $B_t = (B_t^1, B_t^2, \ldots, B_t^n)^T$ with correlated components. Actually, this is a positive definite matrix whose elements are formally given by: $corr(B_t^{(i)}, B_t^{(j)}) = \rho_{ij}$ where ρ_{ij} is the $(i,j)^{th}$ entry of the matrix of Σ.

Issues related to construction and simulation of the multidimensional fractional Brownian motion are treated in details in the works of Haugh (2004) and Glasserman (2003).

The simulation study can be divided into three steps: In the first step, we apply the PLS regression scheme to the 55 realizations of the multidimensional fractional Brownian motion process. In the second phase, we add noise components to the simulated trajectories and we apply the PLS regression to the noisy dataset. It should be stressed here that the amount of the simulated noise is pre-specified by the "signal to noise ratio". This is the ratio of a signal power P_S to the noise power P_N present in the signal. Formally the SNR is defined as:

$$SNR = 10 \log_{10}(\frac{P_S}{P_N})$$

We impose 17 different levels of additive Gaussian noise to the components of the simulated multidimensional fractional Brownian motions. It should be remarked here that choosing different SNR levels allows us to assess the robustness of the denoising technique to a change in the noise amplitude. The third step consists in applying the Wavelet-PLS method to the 55 noisy matrices so as to test the relevance of the proposed scheme.

Fig.4 shows the values of $\triangle R_1^2$ which are the differences between the goodness of fit values for the PLS regression performed on the initial dataset and those

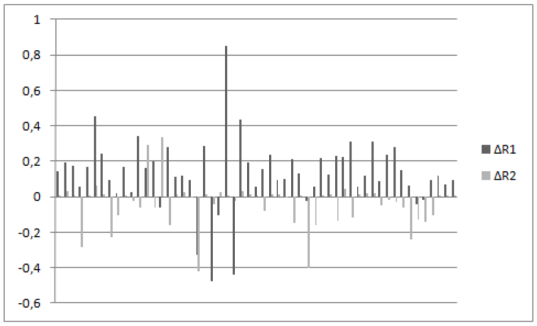

Fig. 4. Normalized distances to the model NDModX for the Wavelet-PLS regression.

of the PLS model performed on the noisy data. On the same figure, we have also plotted associated with simulated cases provided by the the values of $\triangle R_2^2$ which are the differences between the R^2 values for the PLS regression performed on the initial dataset and those of the PLS model performed on the denoised data. ($\triangle R^2$) before and after introduction of noise

Remark that, overall, the $\triangle R_1^2$ values are much closer to zero than the $\triangle R_2^2$. This shows the effectiveness of the wavelet techniques for noise removal.

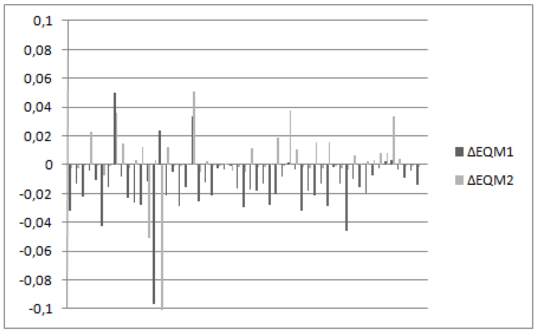

Fig. 5. ($\triangle MSE$) before and after introduction of noise

Fig.5 shows the values of $\triangle MSE_1$ which are the differences between the Mean Squared Errors (MSE) for the PLS regression performed on the initial dataset and those of the PLS model performed on the noisy data. For the sake of comparison, we have also plotted the values of $\triangle MSE_2$ which are the differences between the MSE values for the PLS regression performed on the initial dataset and those of the PLS model applied to the denoised data.

In view of these results, it is clear that the $\triangle MSE_2$ are much smaller than those of $\triangle MSE_1$. This confirms the relevance of the Wavelets-PLS method.

7 Conclusion

In this work, a novel data analysis method has been proposed and discussed. It consists of utilizing wavelet based thresholding techniques in association with PLS regression.

By applying the Wavelet-PLS approach to oil production data sets, we were able to improve the modelling performance of the PLS regression model. Indeed, we succeeded in:

- diminishing the number of outliers
- reducing the Mean Square Error
- correcting the observations in the score plot
- ameliorating the model goodness of fit (R^2)

References

AMINGHAFARI M., CHEZE N. and POGGI J.M. (2006). Multivariate denoising using wavelets and principal component analysis. *Computational Statistics & Data Analysis 50 (9), 2381-2398*

DAUBECHIES I. (1992). *Ten lectures on wavelets*, SIAM, Philadelphia

DONOHO D. (1995). De-noising by soft-thresholding. *IEEE Trans. Inform. theory, 41 (3), 612-627.*

DONOHO D. and JOHNSTONE I. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist, 26 (3), 879-921.*

DONOHO D. and JOHNSTONE I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika, 81, 425-455.*

MALLAT S. (2000). *Une exploration des signaux en ondelettes.* Les éditions de l'école Polytechnique, Ellipses edition.

HAUGH M. (2004). "The Monte Carlo Framework, Examples from Finance and Generating Correlated Random Variables". Course Notes. *www.columbia.edu/mh2078/MCS04/MCS framework FEegs.pdf.*

GLASSERMAN P. (2003). *Monte Carlo methods in financial engineering.* Springer-Verlag.

TENENHAUS M. (1998). *La régression PLS : Théorie et Pratique.* Technip, Paris.

TENENHAUS M. (1995). *Nouvelle Méthodes de Régression PLS.* Les cahiers de recherche, CR540.

TENENHAUS M., GAUCHI J. P. and MENARDO C. (1995). Régression PLS et Applications. *Revue de Statistique Appliquée, (3), 7-63.*

VIGNERON V., PARASCHIV-IONESCU A., AZANCOT A., JUTTEN C. and SIBONY O. (2003). Fetal electrocardiogram extraction based on non-stationary ICA and wavelet denoising. *Seventh IEEE International Symposium on Signal Processing and its applications, (2), 69-72.*

WOLD S., MARTENS H. and WOLD H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. *In Proc. Conf. Matrix Pencils, Ruhe A. and Kastrom B. (Eds), March 1982, Lecture Notes in Mathematics, Springer Verlag, Heidelberg, 286-293.*