Contribution to the methodology for principal components selection when developing NIR calibration models

Jesús Pérez Aparicio[1] and Carmen Miranda Losa[1]

[1]Centro de Investigación y Formación Agraria, CIFA .Dirección General de Investigación y Formación Agraria y Pesquera, Avda Rodríguez de la Fuente s/n, Apdo. 29-14700, Palma del Río, Córdoba, Spain

Introduction

An important step in the statistical model development is the selection of principal components and their number. The most commonly used methodology (Naes and Martens) adds principal components in order of explained variance until validation shows that there is no significant improvement in the prediction. This procedure is developed during the cross validation.[1]

This technique can be very useful in data problems involving minimal distributional assumptions.[2]But we necessarily must analyse our prediction model overfitting phenomenon, otherwise our model would underestimate the error expected prediction rate with excessive components number, like Herwig Friedl and Erwin Stampfer explained.[2]

Methods

Calibration process

For obtaining reliable models we have developed a strategy (see Figure 1) for principal components number determination by PLS regression and cross validation with Burman recommendations on building a cross-validatory method also called "corrected h-block cross-validation".[2]

On this strategy we have made a script program with Matlab 6.3 version and the Fastmcd,[3] Rapca,[4] and Savitsky–Golay algorithms, and it has been proved with data from the spectra obtained to predict the fat percentage of commercial sliced sausages from a meat industry in Córdoba. (For more details about this data, you can find the proceeding entitled: "Utilization of NIR to predict the fat percentage on commercial sliced salchichón").

Independent test set creation (Burman criterion)

The test set is a fixed fraction from either side of the total population distribution. The results for this criterion are shown in Figures 2 and 3 with different components number.

Results

Some results obtained by this strategy are given in Figures 2 and 3.

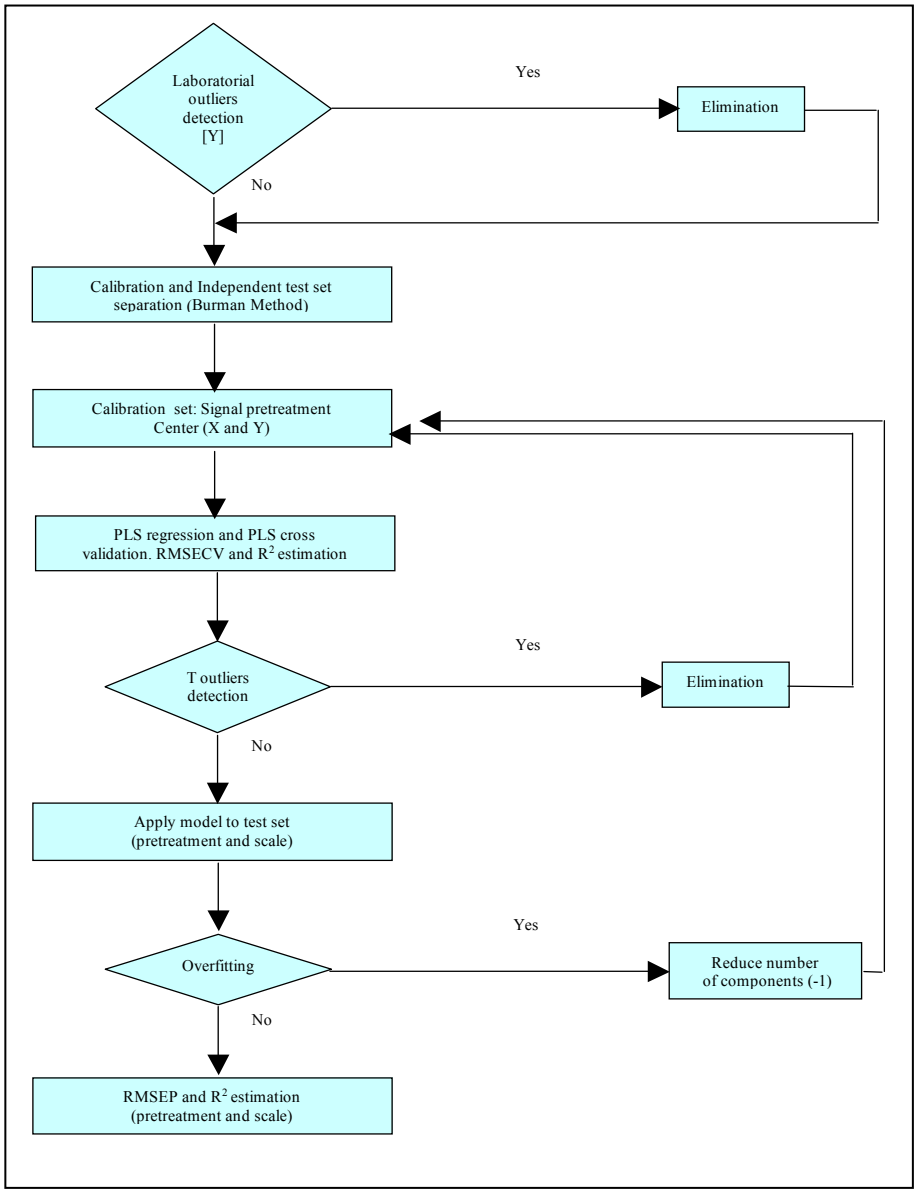

Figure 1. Calibration process diagram for principal components determination.

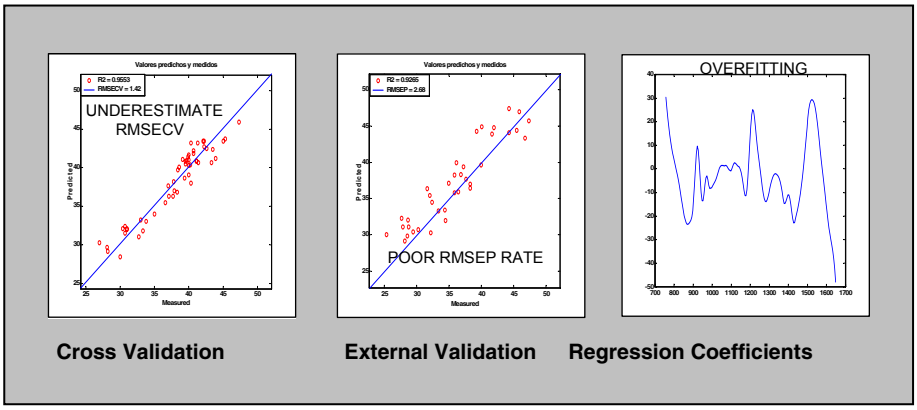

Figure 2. Overfitting phenomenon with four principal components, therefore we must reduce components.

The overfitting phenomenon appears when we choose more components than we should do and is noticeable through the regression coefficients graphic and the difference between both error rate values (cross validation and external validation). When the overfitting is difficult to see, we must decide according to the best *RMSEP*.

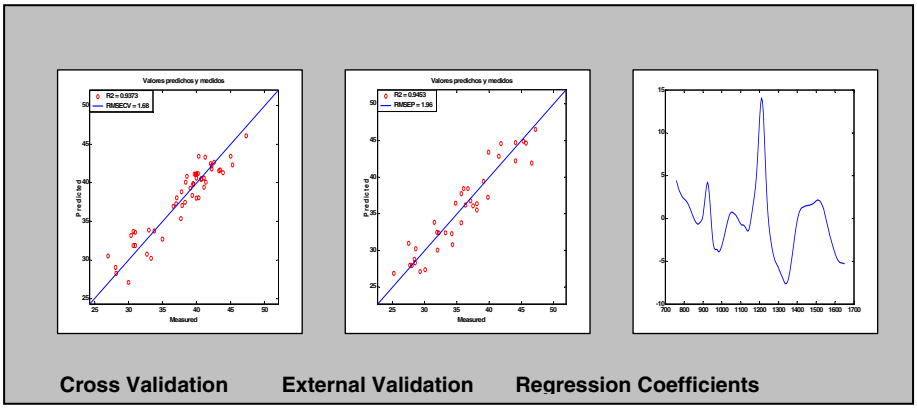

Figure 3. The best results are when we choose three principal components.

Discussion

The same data should never be used for training, optimising and validating the model. If you do not have external validation data to build your model you should use a strategy to obtain almost independent training and test sets, like Burman recommendations on Herwig Friedl and Erwin Stampfer article.[2]

References

1. R. de Maesschal, F. Estienne, J. Verdu-Andres, A. Candolfi, V. Centner, F. Despagne, D. Jouan-Rimbaud, B. Walczak, D.L. Massart, S. De Jong, O.E. de Noord, C. Puel, B.M.G. Vandeginste, *The development of calibration models for spectroscopic data using principal component regression.*
2. Herwig Friedl and Erwin Stampfer, *Cross-Validation*. Technical University Graz, Austria, (2001).
3. P.J. Rousseeuw and K. Van Driessen, "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics* **41,** 212 (1999).
4. M. Hubert, P.J. Rousseeuw and S. Verboven, "A fast method for robust principal components with applications to chemometrics", *Chemometr. Intell. Lab. Syst.* **60,** 101 (2002).