SPARSE-BASED EXPLORATION OF HYPERSPECTRAL IMAGES

R. Calvini[1], J.M. Amigo[2], A. Ulrici[1]

[1]*Department of Life Sciences, University of Modena and Reggio Emilia, Pad. Besta Via Amendola 2, Reggio Emilia, Italy*
[2]*Department of Food Science, Faculty of Sciences, University of Copenhagen, Rolighedsvej, Copenhagen, Denmark*

One of the main problems related to hyperspectral image analysis is to extract the relevant information from the huge amount of data contained in each image. To this aim, Principal Component Analysis (PCA) is generally used as a first approach for data exploration before the application of more complex calibration or classification algorithms. In particular, PCA allows evaluating the main sources of variability and identifying similarities and dissimilarities between the objects depicted in the images. However, sometimes the identification of relevant spectral regions based on the interpretation of the PCA loading vectors is not a trivial task.

In this context, the use of sparse-based PCA (sPCA) [1, 2] allows to exploit the advantages of PCA for image exploration and, at the same time, to simplify the chemical interpretation of the results by forcing to zero the loading coefficients of noisy or uninformative variables. The level of sparsity, (i.e., the amount of spectral variables that are set to zero), is a user-defined parameters which has to be carefully tuned in order to avoid the deletion of features bringing relevant information. Since sPCA is an unsupervised technique, the choice of the proper combination of the model parameters (i.e., sparsity and optimal number of principal components) is a major issue and, therefore, it is necessary to define a response which relates the sparsity level to the desired performance of the model.

In the present work, frequency distribution curves calculated form the score vectors of different sPCA models were used as an effective tool for the choice of the proper sparsity level. In this manner, it was possible to evaluate the influence of different sparsity levels on the effectiveness of the resulting models and, therefore to identify the optimal conditions for the problem at hand.

In particular, two practical applications were considered in this study: the separation between groups of homogeneous samples and the identification of outlier pixels in the space domain. These two case studies are related to different issues, which may occur in the analysis of hyperspectral images.

The first application concerned the discrimination between homogeneous samples of Arabica and Robusta green coffee. For this situation the best sparsity level is the one leading to the higher separation between the frequency distribution curves of the two considered categories. The second application concerned the identification of outlier pixels, related to the presence of pieces of different plastic polymers among pieces of polystyrene. In this case, it is necessary to find a sparsity level which enhances the amount of pixels detected as outliers.

For both the case studies, the identification of the proper sPCA model allowed to highlight the relevant spectral variables and to enhance the chemical interpretation of the considered problems.

Furthermore, the stability of variable selection performed with sPCA was evaluated by assessing the convergence of the variables selected with different sparse models. As a matter of fact, for both applications the most frequently selected variables converged to those selected using the optimal sPCA model.

References:

[1] M.A. Rasmussen, R. Bro, *A tutorial to the Lasso approach to sparse modelling*, Chemom. Intell. Lab. Syst., 119, 21 (2012).

[2] R. Calvini, A. Ulrici, J.M. Amigo, *Practical comparison of sparse methods for classification of Arabica and Robusta coffee species using near infrared hyperspectral imaging*, Chemom. Intell. Lab. Syst., 146, 503 (2015).