Assessing the impact of variable scaling on projection methods in the analysis of spectral data

Jean-Pierre Labuschagne*
University of South-Africa, Roodepoort, South-Africa - labusj@unisa.ac.za

René Pellissier
University of Massachusetts, United States of America - renepellissier@gmail.com

Data pre-processing performs a core function in the analysis of spectral data. The choice of the pre-processing method not only impacts the final model, but also affects the extent of influence that variables play in estimating the final model. In this research the impact that different variable scaling methods have on principal component analysis (PCA) and variable importance on projection (VIP) was investigated. Auto, Pareto, level, power, vast and range scaling was applied to eight spectral datasets. Results from PCA indicated that centred and log scaled data consistently extracted the largest amount of variation with the lowest number of components throughout all the datasets. VIP scores were heavily affected by the use of different scaling methods, furthermore, patterns pertaining to the ranking of the VIP scores between the different scaling methods were observed. This research has confirmed to what extent data pre-processing influences analyses when using projection methods, and care should therefore be taken in the selection of the data pre-treatment method. This result could change the way previous results were interpreted and it is hoped that it will contribute to setting a standard for pre-processing data in a set way as well as enhance comparisons in studies were different scaling methods are employed.

Keywords: data pre-processing; variable scaling; VIP; PCA; PLS.