Automated fitting process using robust reliable weighted average on near infrared spectral data analysis

ABSTRACT

With the complexity of Near Infrared (NIR) spectral data, the selection of the optimal number of Partial Least Squares (PLS) components in the fitted Partial Least Squares Regression (PLSR) model is very important. Selecting a small number of PLS components leads to under fitting, whereas selecting a large number of PLS components results in over fitting. Several methods exist in the selection procedure, and each yields a different result. However, so far no one has been able to determine the more superior method. In addition, the current methods are susceptible to the presence of outliers and High Leverage Points (HLP) in a dataset. In this study, a new automated fitting process method on PLSR model is introduced. The method is called the Robust Reliable Weighted Average—PLS (RRWA-PLS), and it is less sensitive to the optimum number of PLS components. The RRWA-PLS uses the weighted average strategy from multiple PLSR models generated by the different complexities of the PLS components. The method assigns robust procedures in the weighing schemes as an improvement to the existing Weighted Average—PLS (WA-PLS) method. The weighing schemes in the proposed method are resistant to outliers and HLP and thus, preserve the contribution of the most relevant variables in the fitted model. The evaluation was done by utilizing artificial data with the Monte Carlo simulation and NIR spectral data of oil palm (Elaeis guineensis Jacq.) fruit mesocarp. Based on the results, the method claims to have shown its superiority in the improvement of the weight and variable selection procedures in the WA-PLS. It is also resistant to the influence of outliers and HLP in the dataset. The RRWA-PLS method provides a promising robust solution for the automated fitting process in the PLSR model as unlike the classical PLS, it does not require the selection of an optimal number of PLS components.

Keyword: Near infrared spectral data; Robust; Partial least squares regression; Average-weighted; Number of components; Reliability coefficients