EXPLORING THE IMPACT OF SCALING TECHNIQUES ON CLASSIFICATION ACCURACY OF NMR METABOLOMIC DATA USING PLSA-DA AND OPLS-DA

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

Analysis of NMR data is highly sensitive to the method of scaling used. The classification accuracy of models as well as model significance is highly influenced by the chosen data pre-processing method. There is a lack of consistency in articles that report on results in using different scaling techniques and vague guidelines as to which techniques are more suited for different analysis objectives. Prior studies have compared relatively newly introduced scaling techniques such as variable stability scaling (vast) and generalised logarithm (glog) transform to commonly used techniques such as pareto- and unit variance scaling, but cross comparisons between these techniques have not been conducted with respect to implications on PLS and OPLS-DA classification models. This presentation will evaluate the impact of various scaling techniques on the classification accuracy of the two most common supervised methods used in the multivariate analysis of NMR data, namely PLS-DA and OPLS-DA. The tendency of over fitting and the influence of sample size will also be explored. Model validation tools used will include cross-validation, jack-knifing model parameters, permutation tests and bootstrapping.

Keywords: generalised logarithm transform, pareto scaling, variable stability scaling