L6. Defining Multivariate Calibration Model Complexity for Model Selection and Comparison

John Kalivas
Idaho State Univeristy, Pocatello, USA

In the analysis and comparison of multivariate calibration models, the concept of degrees of freedom (fitting degrees of freedom, prediction rank, model complexity, etc.) has an important role. This concept is often related to the number of respective basis vectors (latent vectors, factors, etc.) when using principal component regression (PCR) or partial least squares (PLS). Comparisons between PCR and PLS models for a given data set are often made with the prediction rank to determine the more parsimonious model, ignoring the fact that the values have been obtained using different basis sets. Additionally, it is not possible to use this approach for determining the prediction rank of models generated by other modeling methods such as ridge regression (RR). Measures are presented of what will be called the effective rank for a given model that can be applied to all modeling methods, thereby providing inter-model comparisons in terms of model complexity. With a proper definition of effective rank, a better assessment of degrees of freedom for statistical computations is possible. Additionally, the true nature of variable selection for improved parsimony over full variable models can be properly assessed. Spectroscopic and quantitative structure activity relationship (QSAR) data sets are used as examples with PCR, PLS, and RR.