L04. Cross-validation — a chemometric dinosaur going extinct

Kim H. Esbensen
ACABS research group, Aalborg University, campus Esbjerg, Denmark

Validation is concerned with assessing the performance of a specific data analytical model, be it for prediction, classification, time-series forecasting, or similar... In statistics, data analysis and chemometrics, a much favoured method of validation is cross-validation, in which a subset of the training set apparently performs as an 'independent test set' in a sequential manner. Depending on the fraction of training set samples (N) temporarily held out of the contemporary modelling, a range of no less than (N–1) potential cross-validation segments will always exist, the specific number of segments falling in the interval [2,3,4,...,(N–1), N]. Various 'schools of thought' of cross-validation have developed within chemometrics, some favouring 'full cross-validation' (one object per segment; N segments in total), some defining 10 (or 4) as the canonical number of segments — with still other, more complex schemes are also offered. Usually however there is more focus on *strict adherence* to some form of cross-validation procedure, or other, than openness to investigate what exactly are the precise assumptions and prerequisites behind cross-validation. This comprehension has hitherto been mostly lacking — as has indeed also been a desire to discuss and comprehend the full suite of in-depth issues involved in a debate which has ranged throughout the entire history of chemometrics (sometimes a very heated discussion, often more emotional than rational).

This contribution discusses these issues in depth. The general conclusion arrived at is that cross-validation is only a *simulation* of test set validation, in form strikingly similar but not with regard to the essential characteristics. The crucial fact is that there is only one data set involved in cross-validation, namely the training set. This precludes any possibility for more than one realization of the sampling errors involved; there are both statistical as well as physical sampling errors in the general case, the latter collectively termed TSE (Total Sampling Errors). Given the fact that the physical sampling errors overwhelmingly dominate, typically 10–50–100 X analytical errors alone, it is evident that any singular N-object data set constitutes only one specific realization of these sampling error materializations. The main lesson from TOS' more than 50 years of practical experience is that there is no such thing as a constant sampling bias - the physical sampling bias changes with every new sampling from heterogeneous materials as well as from similar measurement scenarios. From this it follows that there can never be any guarantee that one specific training set realization will also be representative of all future similar data sets. Taking a second data set, the test set, becomes an absolute necessity. Incorporating this second data set is the only way in which to incorporate information from both TSE materializations, for example, in prediction performance validation. From this discussion, one can conclude with complete generality that all variants or schematics of the cross-validation type are inferior, indeed unscientific. Cross-validation must logically be discontinued — unless absolutely no option for test set validation can be demonstrated. It should be stated that there does exist a (very) minor class of scenarios (only) in which cross-validation still has merit, but absolutely no generalizations can be made on this basis. Only test set validation can stand up to the logical demands of the general validation imperative.