T02. Critical levels in projection techniques

A. Pomerantsev
Institute of Chemical Physics RAS, Moscow, Russia

Keywords: PCA, SIMCA, leverage distribution, residual variance distribution, type I error

The critical levels in projection are used intensively in chemometrics; particularly, in the following areas. The first one is SIMCA method; a popular chemometric tool for the supervised pattern recognition. The second area is the multivariate statistical process control (MSPC) that employs critical levels for tracing the behavior of a process. The third application is the outlier detection in the multivariate calibration with the influence plot as a specific tool.

After projecting initial data onto a score subspace, each sample can be presented as the sum of the vector that lies in the subspace (a projection) and transversal vector (a residual). The lengths of these vectors characterize the sample position with respect to the model (subspace). They are termed as the score distance (SD, aka leverage) and the orthogonal distance (OD, aka residual variance), correspondingly. The SDs and ODs obtained for the known class members constitute two samplings, which represent the population. By exploring these samplings the critical membership levels can be established. Further, when a new candidate object x is considered it can be projected onto the model subspace, and its own SD and OD values can be compared with the critical levels to make a decision on the membership of the class.

In such a context, several statistical problems are of vital importance. The first is the form of the SD and OD distributions. Moreover, in each specific case, the distribution parameters are to be evaluated using the training data set. It is further important to set up the rules that reveal the extremes and outliers in the data. At last, the acceptance area in the SD-OD plot should be defined for a given type I error. Such issues have already been discussed in the numerous publications but they are still topical.

This research aims at an open theoretical discussion on these topics. Several real-world examples will be shown at presentation T03, which is closely connected with this one