L4. Simple View on Simple Interval Calculation (SIC)

Alexey Pomerantsev, Oxana Rodionova
Institute of Chemical Physics, Moscow, Russia

Keywords: multivariate calibration, interval estimations, SIC-method, linear programming, object status classification.

Simple Interval Calculation (SIC) is a method for linear modeling and for prediction interval estimation in the multivariate calibration (MVC) problem

$$\mathbf{y} = \mathbf{Xa} + \boldsymbol{\varepsilon}$$

It is shown that SIC leads to results that are in a convenient interval form, and which account for all uncertainties present (X measurement errors, y measurement errors, bilinear modeling errors). The SIC approach also provides wide possibilities for leverage-type object status classification. This method is based on the single assumption that all errors involved in MVC are finite. In this aspect SIC differs from traditional chemometric methods used for multivariate data analyses and therefore is hardly apprehended by analytics.

In the presentation we discuss the finiteness of error. It is known, that assumption of normal error distribution is a commonplace in the conventional data analysis. Sometimes this is expressed implicitly, but often this is assumed by default. However, the researchers do not connect the normality of error distribution with its unboundedness. Does anybody take into account the data points, which are located beyond four standard deviations (4σ)? On the other hand, the amount of data in the modern data mining is often greater than 10+6. Therefore, from the statistical point of view, there should be 20-30 values that lie beyond 4σ. Where are they? The answer is that if such values occurred, they are excluded before data processing. We consider that in a real case study it should be assumed that all error distributions are truncated on 4σ or may be even on 3σ. Just this simple idea leads to the very drastic outcomes that throw a new light on the old MVC problem.

We are explaining the SIC approach using the simplest and familiar examples. Our goal is to introduce the SIC method in parallel with the traditional regression approach in order to emphasize both common and extraordinary SIC features. In the description we use a simulated data set that illustrates the SIC technique and adumbrate the main SIC concepts and postulates. We also demonstrate the SIC application to a well-known real world data set that is the octane rating procedure based on the near infrared spectroscopy.