Weight selection and rotation ambiguity in multivariate curve resolution of absorption spectra with weighted least squares

Skvortsov A.N.

Saint Petersburg Polytechnic University, St. Petersburg, Russia

Multivariate curve resolution (MCR) is set of chemometric techniques, which aim to resolve data matrices into spectra and concentrations of individual chemical components, given a set of physicochemical constraints. Given a data matrix X an MCR method searches for a best bilinear model CS^T, where C and S are constrained matrices of spectra and concentrations. In most MCR methods this is done by minimizing the sum of squares $\sum_{mn} e^2_{mn}$ of the elements of the residual matrix $E=X-CS^T$. Such approach implies that the residuals are distributed normally, similarly, and independently. This is not always the case in real data, where the uncertainties may differ from point to point and/or outliers may occur. One of the classic solutions is to use weighted least squares (WLS) approach, which minimizes $\sum_{mn} e^2_{mn} w_{mn}$, where $\{w_{mn}\}$ is a weight matrix. Typically, the weight matrix is constructed from the maximum likelihood principle. However, WLS is a very flexible technique, it can be adapted for MCR of sparce matrices, robust MCR (by iterative reweighting procedure), and some nonlinear MCR tasks.

A complicacy of MCR with WLS is the inapplicability of principal components for finding MCR factors, so many good MCR methods cannot be used with WLS. The problem can be solved by weighted alternative least squares (termed MCR-WALS). Also a faster positive matrix factorization (PMF) algorithm was proposed (*Paatero, Tapper, 1994*), which is currently a widely used WLS method. The possibilities of MCR-WALS, PMF and some other methods were recently revisited in (*Stanimirova, Tauler, Walczak; 2011*). MCR-WLS solutions, similarly to common MCR, may have residual rotation ambiguity (RA). While RA in common MCR was extensively studied, RA in MCR-WLS and its relation to weight matrix is much less known.

In the present work we focused on MCR of 'saturated' absorption spectra (optical density). Optical density *D*, which is bilinear with rellllspect to spectra and concentration, is a log-ratio measurement. Together with the nature of uncertainties in light intensity measurements this results in strong dependence of measurement uncertainty δD on *D* with typical sharp increase of δD at large *D*. So large *D* values are unstable, effectively nonlinear (hence the term 'saturated spectra') and compromise the MCR model. Common MCR is unable to handle this situation, but MCR-WLS may succeed if proper weights are selected, which would damp or even zero out the contribution of large *D* values. Still, the analysis of residual RA of the solution is required.

To test the relative efficiency of different weighting schemes and estimate RA in MCR- WLS several simulated examples and a real-life data matrix were analyzed. MCR-WLS factors and analogs of principal components (for RA analysis) were calculated by MCR-WALS and weighted version of NIPALS respectively. RA was estimated by Borgen plots, tracking, and charged particle optimization cPSO. Simple threshold weighting ($w=0 \mid D > D_{max}$) or accurate weighting based on uncertainty function and maximum likelihood principle were used.

Both weighting schemes displayed better performance in MCR of saturated spectra than unweighted MCR (regarding the determination of the number of components, retrieving spectra and concentrations, and satisfying the constraints). Most of the existing methods of RA estimation can be generalized to MCR-WLS. RA in simple MCR and MCR-WLS for slightly saturated spectra showed similar behavior, although the shapes of the feasible areas were different. Strong saturation in the spectra increased rotation ambiguity; it corresponded well to the decrease in the relative count of 'reliable' and 'extreme' data points. It was also shown that accurate weighting of measurements, which are close to zero, can expand the power of Borgen plot RA analysis (as an alternative to generalized Borgen plot algorithm).