Lecture for CMA4CH 2010, Mediterraneum Meeting

Application of Multivariate Analysis and Chemometry to Cultural Heritage and Environment
3rd ed., Taormina, Sicily island, Italy, Europe, 26-29 September 2010

Using Clutter to Improve Pattern Recognition, Calibration and Classification Models

B.M. Wise, J.M. Shaver and N.B. Gallagher
Eigenvector Research, Inc., 3905 West Eaglerock Drive, Wenatchee, WA 98801 USA

Abstract

Over the past 10 years, a number of powerful spectral analysis methods have been published which make use of orthogonalization (*i.e.* projection followed by weighted subtraction) of interferences or "clutter." These filtering methods provide a means to mitigate the effect of interferences arising from background chemical or physical species, instrumental artifacts, systematic sampling errors and instrument or system drift. They have been used very effectively with complex biological systems, remote sensing applications, chemical process monitoring and calibration transfer problems.

This class of methods includes Orthogonal Partial Least Squares (O-PLS) [1], External Parameter Orthogonalization (EPO) [2], Dynamic Orthogonal Projection (DOP) [3], Orthogonal Signal Correction (OSC) [4], Constrained Principal Spectral Analysis (CPSA) [5], Generalized Least Squares Weighting (GLSW) [6-7], and Science Based Calibration (SBC) [8] among others. All are based on the orthogonalization premise and each touts a unique ability to improve model performance, robustness, and/or interpretability.

However, in spite of their unique claims, these methods are all highly related and some are basically identical to or special cases of each other. Furthermore, they are related to many older methods, including Weighted Least Squares (WLS) [9], Extended Least Squares (ELS) [10], and the Extended Mixture Model (EMM) [11]. This heritage is often ignored.

This paper will discuss how these methods are related and how they perform in classification and calibration applications. The actual differences in implementation and performance will be discussed and presented for example applications. Emphasis will be placed on the methods that use clutter for defining filters which may be applied in pattern recognition, calibration and classification models.

Introduction

When developing chemometric models, whether for pattern recognition, calibration or classification, data preprocessing is often the critical step, and often determines the overall success of the model. The goal of preprocessing is to eliminate, or at least mitigate, the effects of variation which are unrelated to the problem of interest so that related variation may be more easily seen. Spectroscopic systems that produce light scattering effects are an example of this: light scattering is a physical phenomena that produces variation in the data that makes it harder to see the variation due to chemical effects, which is generally the goal.

Orthogonalization filters remove spectral patterns from data that are “interfering” with the signal of interest. The interfering species are historically called “clutter” and include background components, noise, and chemical species other than the targets of interest. Filters return spectra with the clutter features removed. This often has the desired effect of improving and/or simplifying subsequent modelling.

One possible way of dividing orthogonalization filters is between those that do a soft subtraction, *i.e.* a de-weighting (GLS, SBC, WLS) and those that do a hard subtraction which totally eliminates the subspace spanned by the clutter (O-PLS, EPO, OSC, CPSA, ELS, EMM). Another way to divide the methods is in the way that the clutter components are derived. In O-PLS and OSC the clutter is derived directly from the calibration data using the predicted variable (**y**) as a guide. In the other methods, the clutter can be derived from many different sources, including 1) background

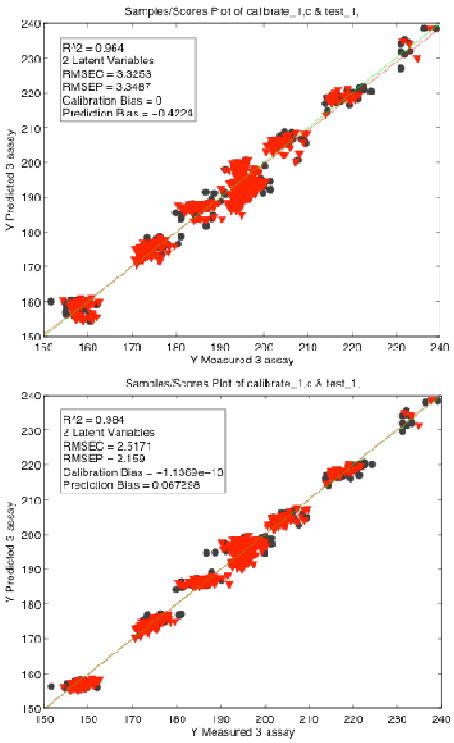
samples that vary but do not contain the target analyte, 2) calibration samples weighted by the inverse of the **y**-block gradient and 3) pure component spectra of the interfering components.

Materials & Methods

Examples of the orthogonalization filters considered were generated using publicly available data sets. Computation was done using MATLAB, PLS_Toolbox and MIA_Toolbox.

Results

The effectiveness of the orthogonalization methods, are demonstrated on several data sets, with a single example shown here. Figure 1 shows the calibration for the IDRC 2002 Shootout data, which is the NIR transflectance spectra of pharmaceutical tablets and associated assay values. The initial calibration is done with only Multiplicative Scatter Correction (MSC) and mean centering. It has an RMSEC of 3.33 (black points) and RMSEP of 3.35 on an independent test set (red points). Figure 2 shows the same data with GLS weighting based on y-gradient. This model has an RMSEC of 2.52 and RMSEP of 2.16, a considerable improvement.



Conclusions

It is demonstrated that many of the method produce similar results. This is not surprising given that they are all using the same basic information in similar ways. Compared to O-PLS and OSC, the methods which derive the clutter from external sources offer the advantage of additional flexibility in the way the filters are derived and applied.

References

1) J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), *J. Chemometr.*, 16(3), (2002), 119-128
2) J.M. Roger, F. Chauchard, V. Bellon Maurel, EPO–PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits, *Chemometr. Intell. Lab.*, 66(2), (2003), 191-204
3) M. Zeaiter, J.M. Roger,V. Bellon-Maurel, Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations, *Chemometr. Intell. Lab.*, 80(2), (2006), 227-235
4) S. Wold, H. Antti, F. Lindgren, J. Öhman, Orthogonal signal correction of near-infrared spectra, *Chemometr. Intell. Lab.*, 44(1-2), (1998), 175-185
5) J.M. Brown, Method for correcting spectral data for data due to the spectral measurement process itself and estimating unknown property and/or composition data of a sample using such method, Patent Number: 5121337, Jun 9 (1992)
6) A.C. Aitken, On Least Squares and Linear Combinations of Observations, *Proceedings of the Royal Society of Edinburgh*, 55, (1935), 42-48
7) H. Martens, M. Høy, B.M. Wise, R. Bro, P.B. Brockhoff, Pre-whitening of data by covariance-weighted pre-processing, *J. Chemometr.*, 17(3), (2003), 153-165
8) R. Marbach, A new method for multivariate calibration, *J. Near Infrared Spectrosc.*, 13(5), (2005), 241-254
9) G. Strang, *Linear Algebra and Its Applications, 2nd Ed.*, Academic Press, Orlando, (1980), 148-149
10) N.B. Gallagher, in *Techniques and Applications of Hyperspectral Image Analysis*, H.F. Grahn, P. Geladi, Eds. John Wiley & Sons, England, (2007), 181-201, ISBN 978-0470010860
11) H. Martens, T. Næs, *Multivariate Calibration*, John Wiley & Sons, NY, (1992), 978-0471930471