EIGENVECTOR RESEARCH INCORPORATED
Research, Training, and Software

3905 W. Eaglerock Dr.
Wenatchee, WA 98801
www.Eigenvector.com

Development and Implementation of Effective Multivariate Calibrations for Process Analytical Applications

Charles E. Miller

Key words: FTIR, NIR, Process Analytical, PLS, sample selection, variable selection

Introduction: One of the greatest cost-saving applications of multivariate modeling techniques is in the calibration of multivariate analytical instruments. Such instruments are latently capable of providing rapid analyses of multiple analytes in a process, thus enabling advanced monitoring and control schemes that can lead to greatly improved process yields, process safety, environmental compliance and product uniformity.

The development and implementation of effective multivariate calibrations often involves a wide array of different modeling *tasks*, each of which addresses critical needs in the final calibration model. Some of these tasks are listed below:

- Outlier Filtering
- Data Preprocessing
- Sample Selection
- Variable Selection
- Model Building *and Interpretation*
- On-Line Model Monitoring

This paper provides a brief demonstration of some of these tasks, using routines that are readily available in the PLS Toolbox for Matlab.

Experimental: The data used for this study was obtained from an on-line FT-NIR spectrometer, which takes a continuous side-stream of sample from the feed of a polymerization reactor. This analyzer provides a reactor feed composition measurement every 20 seconds, and is relied upon for advanced reactor composition controls. There are a total of nine different known chemical constituents that can be present in the reactor feed, although only 2 to 4 known constituents are present at any given time. More details regarding this application can be found in other references [1,2].

Calibration data were obtained from two different sources: 1) spectra of actual process samples, and 2) spectra of designed calibration standards, generated by a specially-designed sample injector apparatus. In the former case, reference constituent concentrations were obtained from slower, redundant on-line analytical methods, and in the latter case, they were calculated from measured component masses and flows.

Results and Discussion:

Outlier Filtering: Real process data typically contains a large number of outliers, and the inclusion of such data can adversely affect the performance of a calibration model. Many outliers can be readily observed by simply plotting the raw data. However, there might be more subtle outliers that require more elaborate detection. PCA and PLS model residuals and leverage statistics can be useful for detecting such outliers (see Figure 1).

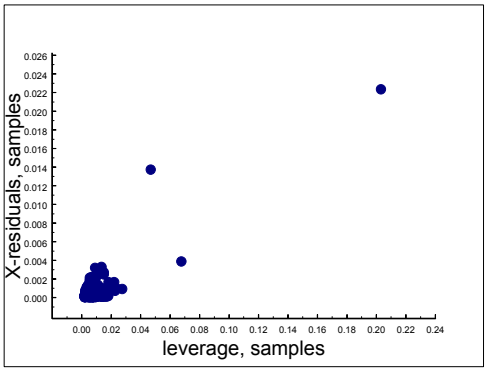

Figure 1: Outlier Filtering via PCA: The residuals versus the leverages of data points of a calibration data set, before outlier removal.

In practice, once such outlier candidates are detected, it is prudent to obtain more information on them, to help determine whether it is appropriate to exclude them from the model data[1-3].

Sample Selection: The calibration samples that were collected from actual process data were simply collected over a fixed time period, and no care was taken to select specific samples for calibration. However, it is very important to ensure that all process sample states are equally-well represented in the calibration data. Otherwise, the resulting model will tend to perform much better for some sample states and not for others. Furthermore, in cases where a very large number of such process calibration samples are collected, sample selection has the added practical benefit of reducing the data processing burden on the

3905 W. Eaglerock Dr.
Wenatchee, WA 98801
www.Eigenvector.com

subsequent modeling software. In this work, a selection method that relies on Single Linkage Nearest Neighbor Cluster Analysis [1-4] was used to define a subset of the process-collected calibration samples (see Figure 2).

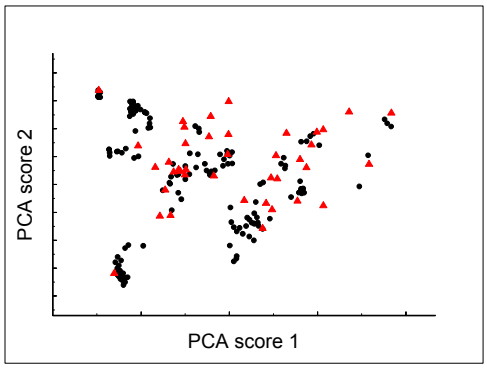

Figure 2: Sample Selection via Cluster Analysis: PC1 vs. PC2 scores for the process calibration samples. Red triangles denote selected samples.

It should be noted that Figure 2 shows only the first 2 dimensions of the data space, and that the selections were based on clustering in a higher-dimensional space.

Variable Selection: In many process spectroscopy applications, access to a very large number of variables (i.e., wavelengths) often tempts one to use irrelevant variables that can only be harmful to the analysis. In this situation, variable selection techniques can be used to simplify the calibration model, thus making it less susceptible to unforeseen disturbances and interferences during its application. One of the techniques that is available in the PLS Toolbox is the Genetic Algorithm technique [3,5]. Figure 3 shows the subset of 45 variables selected by this method (from a total of 290 variables) for this specific application, and Figure 4 illustrates the effect of this selection on the model validation error for this application. Note that selection not only reduces the number of PLS latent variables required for the model (from 10 to 5) and the resulting model validation error (from 0.38 to 0.32 weight%), but it also results in a more dramatic "breakpoint" in the validation error vs. number of PLS latent variables plot.

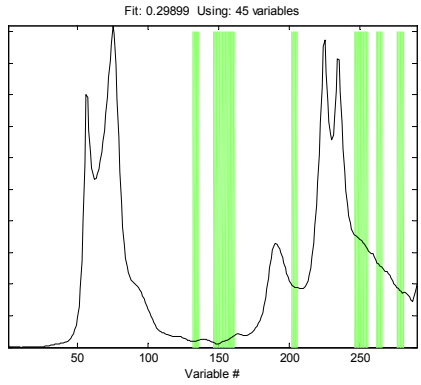

Figure 3: Subset of variables selected using the Genetic Algorithm method, compared to the mean spectrum.

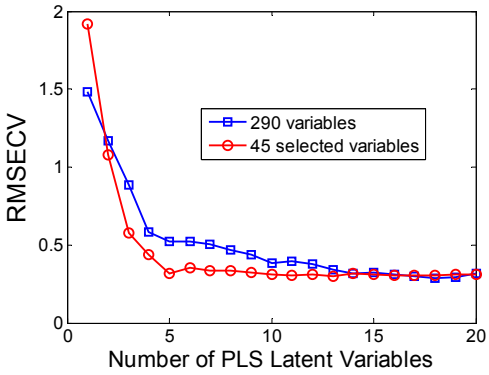

Figure 4: Effect of variable selection, via the Genetic Algorithm method, on the model validation error.

References:

[1] Charles E. Miller, "The Use of Chemometric Techniques in Process Analytical Method Development and Operation", Chemom. Intell. Lab. Syst., 30, 1995, 11-22.

[2] Charles E. Miller, "Chemometrics for On-Line Spectroscopy Applications- Theory and Practice", J. Chemom., 14, 2000, 513-528.

[3] Charles E. Miller, "Chemometrics in Process Analytical Chemistry", in Process Analytical Chemistry, Blackwell Publishing, Oxford, 2004.

[4] T. Isaksson, T. Naes, "Selection of Samples for Calibration in Near Infrared Spectroscopy. Part II: Selection Based on Spectral Measurements", Appl. Spectrosc., 44, 1990, 1152.

[5] D. Jouan-Rimbaud, D.-L. Massart, R. Leardi, O. E. deNoord, "Genetic Algorithms as a Tool for Wavelength Selection in Multivariate Calibration", Anal. Chem., 67, 1995, 4295-4301.