RSC Advances

Threshold Concentration Monitoring Based on Pattern Recognition Analysis of Differential Near-Infrared Spectra

Journal:	*RSC Advances*
Manuscript ID:	RA-ART-07-2014-007579
Article Type:	Paper
Date Submitted by the Author:	24-Jul-2014
Complete List of Authors:	Karunathilaka, Sanjeewa; University of Iowa, Department of Chemistry Small, Gary; University of Iowa, Optical Science and Technology Center

Threshold Concentration Monitoring Based on Pattern Recognition Analysis of Differential Near-Infrared Spectra

Sanjeewa R. Karunathilaka and Gary W. Small*

Department of Chemistry & Optical Science and Technology Center

University of Iowa

Iowa City, IA 52242 USA

*E-mail: gary-small@uiowa.edu; Fax: +1-319-353-1115; Tel: +1-319-335-3214

Abstract

A threshold concentration monitoring procedure based on near-infrared (near-IR) spectroscopy is described for use in continuous process monitoring applications. The method is based on collecting an off-line reference sample and obtaining a near-IR spectrum and corresponding reference concentration at the start of the monitoring period. Subsequently, spectra are collected continuously and ratios are taken to the reference spectrum. The resulting spectra in absorbance units are differential spectra whose effective analyte concentration (termed the differential concentration) is the difference in concentration relative to the concentration in the reference sample. By knowing the reference concentration and a user-specified threshold, a critical concentration can be defined that specifies the threshold in terms of differential concentrations. Determining whether the analyte concentration is within specification can then be addressed as a pattern classification problem and a qualitative classification model can be used to discriminate differential spectra that reflect the two possible states: within or outside of specification. A simulated biological process is used to test the methodology in which a dynamic system of glucose, lactate, urea, and triacetin in the mM range in phosphate buffer is monitored continuously to detect occurrences when the glucose concentration drops below a threshold of 3.0 mM. With the use of three sets of prediction data, one of which was collected 2.5 years after the calibration data, the monitoring algorithm is implemented with 100% successful detections and no false detections.

Introduction

Near-infrared (near-IR) spectroscopy has found significant use in industrial monitoring applications involving process control and quality assurance.[1-7] Within these applications, an area of potential interest is threshold monitoring. In threshold monitoring, rather than track the absolute concentration of a component of interest, the goal is to identify whether that concentration is too low, too high, or outside of a desired range. By not requiring an absolute concentration to be determined, the monitoring task is simplified and potentially made more robust.

In this paper, a threshold monitoring methodology is developed on the basis of the continuous collection of near-IR spectra and the real-time application of a classification model to assign spectra to one of two data categories: (1) spectra meeting a concentration specification or (2) spectra not meeting that specification. In the context of a process control application, these two data classes correspond to situations in which an operator would be notified that the process is out of specification (alarm class) or the corresponding case in which no notification is necessary (non-alarm class). For the example used here, the scenario implemented is a case in which an alarm is sounded if an analyte concentration drops below a pre-determined threshold. The same methodology could be used for identifying the occurrence of a concentration above a threshold or the case in which a concentration is required to be between a specified minimum and a maximum.

In the work presented, a dynamic system based on varying mM concentrations of glucose, urea, lactate, and triacetin in phosphate buffer is implemented. While this chemical system was not designed to mimic a specific industrial or biological process, the compounds are either present in living systems or are spectroscopically similar to such compounds. Thus, the results obtained in this work could be relevant for applications as diverse as continuous monitoring of analytes such as blood glucose in a hospital environment or industrial monitoring

of a biotechnology process. In addition, the compounds used in this study all have absorptivities on the order of 10^{-4} absorbance units (AU)/mm-mM[8] and exhibit a high degree of spectral overlap. This system thus provides a challenge to the sensitivity and selectivity of near-IR spectroscopy for use in continuous monitoring applications.

Theory

The objective of this research was to evaluate the potential for implementing a threshold concentration monitor with near-IR spectroscopy. The key elements of this algorithm are: (1) collection of a calibration database of near-IR spectra and associated reference analyte concentrations that can be used subsequently in the construction of classification models that allow spectra to be assigned membership in one of two data classes: spectra meeting a desired threshold concentration (non-alarm) or spectra not meeting this standard (alarm), (2) use of a reference measurement to determine the analyte concentration at the start of the monitoring period, (3) collection of a reference near-IR spectrum at the same time the reference analyte measurement is made, (4) use of this reference spectrum as the spectral background in the calculation of absorbance values for spectra collected subsequently, (5) definition of a critical concentration that specifies the change in analyte concentration relative to the reference concentration that will cause the process being monitored to be out of specification and thereby trigger an alarm, (6) use of the calibration database to construct a classification model that allows spectra to be grouped into the alarm and non-alarm classes on the basis of the critical concentration, and (7) collection of spectra continuously during the monitoring period and classification of these spectra in real time as belonging to the alarm or non-alarm data classes. If a spectrum is placed into the alarm class, the process is judged to be out of specification and the operator is alerted.

As indicated above, the spectra submitted to the classification algorithm are in absorbance units relative to the reference spectrum collected at the start of the monitoring period.

The motivation for the use of this reference spectrum as the absorbance background is to remove common spectral features that originate from the sample matrix. A complication that arises, however, is that the analyte is present in both spectra and thus the absorbance spectrum that results from taking the ratio no longer has an analyte signal intensity that corresponds to its original concentration. We term this generated spectrum a differential absorbance spectrum and specify its new effective concentration as the differential concentration.

The effective analyte concentration in the differential spectrum is equal to the concentration differences of the two original spectra that are used in the absorbance calculation. This concept can be explained using the derivation shown below. Given two single-beam intensities, I_1 and I_2, collected for samples 1 and 2, respectively, the absorbance for the two samples can be calculated using a background single-beam intensity, I_0, according to the Beer-Lambert law

$$-log_{10}\left(\frac{I_1}{I_0}\right)=abc_1 \quad (1)$$

$$-log_{10}\left(\frac{I_2}{I_0}\right)=abc_2 \quad (2)$$

In Eqs. 1 and 2, the terms c_1 and c_2 correspond to the concentrations of the analyte in samples 1 and 2, respectively, and a and b denote the absorptivity and path length. The wavelength dependence of I_1, I_2, I_0 and a is omitted for simplicity. For the sake of this derivation, these equations further assume that the samples contain only a single absorbing species.

The difference between the two equations is given by:

$$-log_{10}\left(\frac{I_1}{I_0}\right)+log_{10}\left(\frac{I_2}{I_0}\right)=abc_1-abc_2 \quad (3)$$

Expanding the expressions containing logarithms, canceling the terms containing I_0 and rearranging yields:

$$-log_{10}\left(\frac{I_1}{I_2}\right) = ab(c_1 - c_2) \quad (4)$$

As shown in Eq. 4, taking the negative logarithm of the ratio of two single-beam spectra computes a differential spectrum in absorbance units and the concentration is equal to the differences in concentrations of the corresponding spectra (i.e., numerator concentration – denominator concentration).

A characteristic of this calculation is that differential concentrations can be either positive or negative. For glucose as an example analyte in a monitoring application, differential spectra corresponding to differential glucose concentrations of 20.0 mM and –20.0 mM in phosphate buffer are shown in Fig. 1. Both positive and negative spectral features can be observed depending on the differences in the concentrations of the two spectra whose ratio is computed.

This calculation can be extended to multicomponent systems that follow a linear mixture model. In this case, the negative logarithm of the ratio of two single-beam spectra will compute a differential spectrum corresponding to the sum of differences in concentrations of each absorbing species in the two samples. For example, if the sample whose differential absorbance spectrum is displayed in Fig. 1 also contained lactate, both glucose and lactate bands would appear in the differential spectrum according to the differences in concentrations of glucose and lactate between the numerator and denominator single-beam spectra used in the absorbance calculation.

The differential spectral calculation is based on several assumptions. The derivation assumes that the optical path length b does not change across the data collection. This suggests that care should be taken to minimize path length variation over time. The derivation further assumes that the background information is the same for all the spectra collected. If this is not the case, instead of a single I_0 term in Eqs. 1-3, there would be $I_{0,1}$ and $I_{0,2}$. A differential background term would then be introduced into the computed absorbance spectrum. For example, with aqueous samples, if the solution temperature were different between the

numerator and denominator spectra used to compute the absorbance, a temperature-dependent baseline artifact would be introduced into the differential spectrum because of the temperature sensitivity of the background water absorbance.

Spectra can be used interchangeably in the numerator and denominator in the calculation of the differential spectrum. In the work described here, the monitoring scenario was to detect concentration excursions below a set threshold. If one assumes the process is in specification at the start of the monitoring period when the reference measurement is made, the alarm condition will correspond to a negative differential concentration and differential concentrations on both sides of the alarm threshold will be negative. For this reason, calibration differential spectra were computed such that only negative concentration differences resulted.

For example, consider the case in which the alarm threshold concentration is 3.0 mM and the reference concentration obtained at the start of the monitoring period is 5.0 mM. If the reference spectrum corresponding to the reference concentration is used in the denominator of the absorbance calculation, differential spectra just below and just above the monitoring threshold will have differential concentrations near 3.0 – 5.0 = -2.0 mM.

The steps used in building the calibration database are shown in the left flowchart in Fig. 2. This would involve running the process under conditions in which the analyte concentration is caused to change over a specified range, conventional reference measurements are made at fixed time intervals, and near-IR spectra are collected continuously. Single-beam spectra are collected at a specified level of signal averaging and stored in blocks, which are contiguous groups of spectra corresponding to a selected time window. The block size specifies a time window in which the background variation is assumed to be negligible. Differential spectra computed within a block are assumed to have matching backgrounds and thus constant background features will have been reduced to zero absorbance.

Because of the need for potentially costly and time-consuming reference concentration measurements, the number of spectra in which the actual analyte concentration is known will be limited. Depending on the process, it may be possible to interpolate concentrations in order to assign pseudo-reference concentrations to additional spectra. In general, however, the calibration database acquired will have fewer spectra, fewer analyte levels, and fewer reference analyte measurements than would be desirable from the standpoint of experimental design.

The use of differential spectra has an additional advantage for such a case in which the calibration database is limited. By computing differential spectra from all combinations of the single-beam spectra collected within a time block, the calibration database is expanded to fill in additional levels of analyte concentration, as well as some additional variation in the non-constant background components that are not removed by the differential absorbance calculation.

To compress the calibration database, a dimensionality reduction algorithm is applied. For the work described here, the partial least-squares[9] (PLS) algorithm was used to reduce the dimensionally of the original spectra to an *h*-dimensional PLS score matrix. Other possible techniques for use in this task are principal component analysis[10, 11] (PCA) and the discrete wavelet transform.[12] The reduction in dimensionality reduces the time required for the steps required in building the classification model used to identify alarm and non-alarm spectra. The computed PLS spectral loadings and loading weights are saved for the calculation of PLS scores for spectra collected in the future when the classification model is put into operation.

Once the calibration database is assembled, the next step is to calibrate the alarm algorithm. The steps of this procedure are summarized in the right flow chart in Fig. 2. The alarm threshold concentration, C_{alarm}, is user-selected and specific to the process being monitored. In the minimum threshold example employed here, if a spectrum represents an analyte concentration that is equal to or lower than the threshold, an alarm would trigger to alert the operator that the process has gone out of specification.

Once the alarm threshold concentration is defined, the next step is to partition the calibration database (i.e., as represented by the PLS scores) into alarm and non-alarm groups. The differential concentrations can be used to identify the alarm and non-alarm patterns. As discussed previously, the calculation of the differential spectra is performed to yield negative differential concentrations. To identify the alarm and non-alarm spectra within the differential concentrations that comprise the calibration database, a negative threshold concentration needs to be defined. As defined in Eq. 5, the difference between the alarm threshold concentration and the reference concentration (C_{ref}) measured at the start of the monitoring period is defined as the critical concentration, C_{crit}:

$$C_{crit} = C_{alarm} - C_{ref} \quad (5)$$

Assuming that C_{ref} is not already in the alarm state, C_{crit} will always be negative in sign for the low threshold application. This critical concentration identifies the alarm point in the context of the future differential spectra computed with respect to the reference spectrum. As an example, if C_{ref} is 4.0 mM, C_{crit} = 3.0 mM – 4.0 mM = -1.0 mM. Any differential spectrum having a differential concentration below -1.0 mM will trigger an alarm.

The calibration database is partitioned into alarm and non-alarm classes on the basis of C_{crit}. If any differential concentration is lower than this critical concentration, the corresponding PLS score vector (pattern) is placed into the alarm class; otherwise it is placed into the non-alarm class. The alarm decision is thus a classification problem in which patterns are classified into either the alarm or non-alarm classes.

Once the calibration database is partitioned, the next step is to compute a classification model that can mathematically discriminate the alarm and non-alarm classes. Many choices exist for building classification models, including artificial neural networks,[13-17] support vector machines,[16, 17] and discriminant analysis methods.[18, 19] In this work, the technique of piecewise linear discriminant analysis (PLDA) was used to build classification models.

The PLDA method uses multiple linear discriminant functions to approximate a nonlinear separation boundary between two data classes. We have used this method in a number of applications in our laboratory,[20-22] and found it to be quick to implement in an automated manner without the requirement for extensive optimization of its architecture. Because the optimization of the discriminant functions is iterative in nature, three replicate classification models were constructed to allow the incorporation of variability in the positioning of the separating boundary between the data classes.

The steps in the operation of the alarm algorithm are summarized in Fig. 3. Spectra are collected continuously over time while the process evolves. The ratio of each spectrum to the collected reference is taken, forming a differential spectrum corresponding to the signed difference in concentration relative to the reference. After projecting each differential spectrum collected at time, *t*, onto the previously computed PLS factors, an *h*-dimensional spectral pattern (i.e., the PLS score vector), $\mathbf{t}_{diff,t}$ is obtained. Using the previously computed classification model, the pattern $\mathbf{t}_{diff,t}$ is classified into either the alarm or non-alarm classes.

Application of the PLDA method produces a discriminant score that determines the class membership of the pattern, $\mathbf{t}_{diff,t}$. If the discriminant score is higher than zero, the corresponding pattern belongs to the alarm side of the separating boundary while a zero or negative discriminant score corresponds to a pattern on the non-alarm side. As noted above, this research computed three replicate classifiers. To be classified as an alarm, two of the three replicate classifiers had to place the pattern in the alarm class.

Experimental

Near-infrared spectra used for this study were collected during 14 one-day data collection sessions by using two dynamic systems (DS 1 and DS 2), each consisting of four chemical components in phosphate buffer. The individual data groups will be termed runs 1 to 14. If run 1

is defined as time zero, runs 2-14 were conducted approximately 1, 2, 2, 3, 4, 21, 26, 28, 30, 56, 57, 58 and 175 weeks later. Runs 1-5 corresponded to DS 1 and runs 6-14 derived from DS 2.

Reagents

Phosphate buffer (0.1 M) was prepared in 18.2 MΩ water purified by a Labconco water purification system (Labconco, Kansas City, MO). A buffer pH of 7.4 was achieved by titrating monobasic sodium phosphate (ACS reagent, Fisher Scientific, Fair Lawn, NJ) with 50 % w/w sodium hydroxide (Fisher Scientific). The buffer contained sodium benzoate (ACS reagent, Fisher Scientific) at 5 g/L as a preservative.

For DS 1, stock solutions of α-D-glucose (ACS reagent, Fisher Scientific), triacetin (ACS reagent, Sigma-Aldrich, St. Louis, MO) and urea (ACS reagent, Fisher Scientific) were prepared in the phosphate buffer. Each of the stock solutions in DS 1 contained 10 mM sodium L-lactate (ACS reagent, Sigma-Aldrich). Dynamic system 2 was composed of α-D-glucose, triacetin, and L-lactate prepared in phosphate buffer. Each of the stock solutions in DS 2 contained 10 mM urea. The data collected from these two dynamic systems were concatenated to create an overall data set for use in testing the threshold monitor.

Apparatus and Procedures

To simulate concentration excursions that might occur during an industrial process, three stock solutions maintained at ~55 °C in a water bath were mixed in different ratios using three peristaltic pumps (Rabbit-Plus and Dynamax Models, Rainin Instrument Co., Woburn, MA) operating under the control of Rainin Pump Control software (Version VI, Waterville Analytical, Waterville, MA). The individual solutions exiting the three pumps were connected by Y-connectors and flowed through an in-line mixer (Cole-Parmer Instrument Co., Vernon Hills, IL) to achieve a homogeneous output solution.

By changing the pump speeds, and therefore the flow rates of each of the solutions, the concentrations of the solutions exiting the mixer were varied. The concentration of each component of the solution exiting the mixer can be calculated as shown in Eq. 6:

$$C_2 = \frac{(C_1 R_1)}{(R_1 + R_2 + R_3)} \quad (6)$$

For the case of a prepared glucose concentration, C_2, C_1 in Eq. 6 is the concentration of the glucose stock solution, R_1 is the pump speed for the glucose pump, and R_2 and R_3 are the pump speeds for the other two stock solutions. The stock solution concentration, total pump speed (R_1+R_2+R_3) and glucose pump speed thus determine a desired glucose concentration. For the work presented here, the total pump speed was always 10 rpm.

Runs 1-11 were used to form the calibration database, and runs 12-14 were used as prediction sets to simulate the operation of the threshold monitor. The concentration values for each component per sample were assigned to minimize correlations between the constituents. For the calibration data, pairwise correlation coefficients between the chemical components ranged from 0.26 to -0.34. For prediction sets 1, 2, and 3, respectively, the ranges of correlation coefficients were -0.42 to -0.62, -0.40 to -0.65, and -0.43 to -0.57. These levels of correlation were judged to be acceptable from the standpoint of preventing fortuitous results based on chance correlations.

The solution exiting the mixer was flowed through a 20 mm-diameter circular aperture transmission cell (Model 118-3, Wilmad Glass, Buena, NJ). The sample cell employed sapphire windows (Meller Optics, Providence, RI) and had a path length of 1.26 mm. The transmission cell was placed in the sample compartment of a Nicolet 6700 Fourier transform spectrometer (Nicolet Analytical Instruments, Madison, WI). The spectrometer employed a tungsten-halogen source, CaF_2 beam splitter, and a liquid-nitrogen-cooled InSb detector. A K-band optical interference filter (Barr Associates, Westford, MA) was placed before the sample to isolate the

region of 5000-4000 cm^{-1}. To ensure detector linearity, an aperture setting of 100 was used and the source was further attenuated by placing a nominal 63 % neutral density filter (Rolyn Optics, Covina, CA) before the sample.

The temperature of the samples exiting the sample cell was monitored with a copper-constantan thermocouple probe and digital meter (Omega Engineering Inc., Stamford, CT) inserted into a port in the vinyl tubing. For the entire study, the temperature range of the flowing liquid was maintained in the range of 36.6-37.2 °C.

After the sample exited the sample cell, fractions were continuously collected at a rate of 1 min/tube using a Gilson FC 203B fraction collector (Gilson, Inc., Middleton, WI). The glucose concentrations of each of the fractions were verified each day with a YSI Model 2300 STAT PLUS glucose-lactate analyzer (YSI Inc., Yellow Springs, OH) which had an estimated instrumental error of ± 0.2 mM according to the YSI product specifications.

The software used for the data collection and subsequent Fourier processing was Omnic (Version 7.1, Nicolet Analytical Instruments) operating on a Dell OptiPlex GX280 computer (Dell Computer Corp., Austin, TX) running under Windows 7 (Microsoft, Inc., Redmond, WA). Spectra for the liquid flowing through the sample cell were collected continuously as 64 co-added (~ 1 min) asymmetric scans consisting of 4097 points. The Fourier processing steps included one level of zero filling, Happ-Genzel apodization, and Mertz phase correction. The computed spectra had a point spacing of 1.93 cm^{-1}. This corresponded to 519 resolution elements over the range of 4000-5000 cm^{-1}.

The starting time for the spectral collection, starting time for the fraction collector and the time for the solution to flow through the tubing from the sample cell to the fraction collector were used to assign a glucose concentration value to each of the collected spectra. Spectra collected while the pump speed was changing had partially equilibrated glucose concentrations and were omitted from the data analysis.

After collection and Fourier processing, spectra were transferred to a Dell Precision 670 workstation (Dell Computer Corp.) running under Red Hat Linux (Version 5.3, Red Hat, Inc., Raleigh, NC). All subsequent calculations were performed on this computer using the Matlab development environment (Version 7.4.0 (R2007a), The MathWorks, Inc., Natick, MA). Software for the calculation of piecewise linear discriminants used in-house code written in Fortran and compiled with the Intel Fortran Compiler for Linux (Version 10.0, Intel Corp., Santa Clara, CA).

Results and Discussion

Overview of Collected Data

The complete glucose concentration profile for runs 1-14 is given in Fig. 4. The horizontal line at 3.0 mM in indicates the concentration used in this work to define the monitoring threshold. The calibration set was split into training and monitoring sets. The monitoring set was used as a pseudo-prediction set to help in the evaluation of parameters pertaining to the use of PLS and PLDA. Vertical lines in Fig. 4 denote the training (runs 1-10), monitoring (run 11), and prediction (runs 12-14) sets. Across all the data sets, there were 1088 single-beam spectra.

For each concentration level, short-term noise was evaluated by computing 100 % lines from each pair of consecutive spectra. These 100 % lines were converted to AU, and the wavenumber region of 4300-4500 cm^{-1} was fitted to a third-order polynomial model. The RMS noise was then computed about the polynomial fit to obtain the intrinsic measurement noise. The polynomial model corrects for systematic offsets in the 100% lines. The average RMS noise values calculated across the spectra in each of the 14 runs ranged from 1.5 to 4.3 μAU.

Assembly of Calibration Database

For this study, all spectra collected during a single run were treated as having a constant background and placed into a single data block. Differential spectra were calculated by taking

the ratios of all combinations of single-beam spectra within each block. Each ratio was oriented to produce a negative differential concentration. Those combinations that produced a differential glucose concentration of 0.0 mM were not used. This procedure yielded a total of 51,269 differential spectra in the calibration database.

Optimization of Calibration Parameters

The PLS algorithm was used to reduce the multidimensional spectral information in the calibration database into a series of PLS scores. Two parameters that must be optimized for the implementation of PLS are the spectral region submitted to the algorithm and the number of latent variables to be computed.

For the optimization of these two parameters, the training and monitoring sets were used. The training set consisted of 936 single-beam spectra collected over 10 days. The calculation of differential spectra for the training set led to 45,844 spectra. The optimization of the spectral range and the number of PLS factors was performed in two steps: (1) a grid search analysis and (2) a PLDA-based optimization.

The grid search was based on sliding a window of fixed spectral width in 50 cm^{-1} increments across the 4900-4100 cm^{-1} range in the differential spectra. The starting spectral width of 100 cm^{-1} was incremented in 50 cm^{-1} increments up to 700 cm^{-1}. At each step, PLS models for differential glucose concentration were constructed using 3-16 latent variables. This produced a total of 1386 parameter combinations. The performance of each model was assessed by use of cross-validation. Individual cycles in the calculation involved withholding 10% of the calibration subset in contiguous blocks, building a PLS model with the remaining data and then using the model to predict the differential glucose concentrations for the spectra withheld. Pooling the errors in predicted concentration over 10 cycles produced a standard error of cross-validation (SECV).

The computed SECV values were sorted, and an *F*-test was performed at the 95% level to identify the optimal number of latent variables for each spectral range. The optimal model size for a given spectral range was set as the number of latent variables that produced a value of SECV that was not statistically different from the minimum SECV found for that range. Table 1 summarizes the four optimal wavenumber ranges and the corresponding numbers of latent variables.

Fig. 5 plots the values of SECV with respect to the number of latent variables for the spectral range that produced the overall lowest SECV (4650-4250 cm^{-1}). While the minimum SECV occurs at 16 latent variables, the trace is only decreasing very slowly past 11. No benefit to extending the optimization past 16 latent variables is apparent. In addition, despite the results of the *F*-test, further evaluation of model sizes less than 16 is suggested.

The performance of the monitoring set with PLDA was tested with the top four spectral ranges found through the grid search. For each range tested, the number of latent variables was varied from 6 to 11. This selection was made on the basis of plots such as Fig. 5 that suggested little improvement in modeling performance was obtained past 11 latent variables.

To simulate the implementation of the threshold monitor, the first spectrum in the monitoring set was taken as the reference spectrum, and the corresponding glucose concentration (5.3 mM) was used as C_{ref} in Eq. 5. Thus, using an alarm concentration of 3.0 mM and according to Eq. 5, $C_{crit} = 3.0 – 5.3 = -2.3$ mM. The remaining 105 single-beam spectra in the monitoring set were used to compute differential spectra by taking the ratio to the reference spectrum.

The PLS loading weights and spectral loadings previously computed from the training set were then used to compute the scores that defined the pattern vectors corresponding to each differential spectrum. For a given spectral range under consideration, the loading weights and spectral loadings computed from that range were employed in the calculation of the PLS scores.

The critical concentration was used to partition the 45,844 PLS score vectors in the training set into alarm and non-alarm classes. There were 21,364 alarm patterns and 24,480 non-alarm patterns in the training set. For each combination of spectral range and number of latent variables, three replicate piecewise linear discriminants were computed on the basis of using the training set in conjunction with three sets of training parameters. Each replicate classifier was based on a single linear discriminant function.

One measure of the discriminating ability of the patterns is the total number of alarm patterns separated by the discriminant function. For each combination of spectral range and number of latent variables, Table 2 summarizes the percentage (average ± standard deviation) of alarm patterns separated across the three replicate classifiers.

Each replicate classifier was applied to predict the class assignment for the 48 alarm and 58 non-alarm patterns in the monitoring set. Table 2 further summarizes the percentage (average ± standard deviation) of missed and false alarms for each of the parameter combinations studied.

None of the combinations of spectral range and latent variables produced missed or false alarms with the monitoring set. To define a criterion for selecting a classifier for use in subsequent testing with the three prediction sets, the smallest model (i.e., the model based on the lowest dimensional patterns) that achieved an acceptable degree of separation of the training set was selected. A level of 95% separation of the training data was chosen as the criterion for acceptable performance. Through the use of this criterion, the classifier based on a spectral range of 4650-4300 cm^{-1} and eight latent variables was chosen as optimal for use in subsequent testing.

The results for this parameter combination are shown in bold in Table 2. This spectral range is logical as it encompasses the glucose C-H combination band at 4400 cm^{-1} (see Figure 1). The glucose concentration profile for the monitoring set is given in Fig. 6 (right y-axis). Overlaid on the plot are the discriminant scores produced by the optimal classifier when applied to the monitoring data. The discriminant scores plotted represent the result of the classification rule in

which two positive discriminant scores from among the three replicate classifiers signal an alarm decision. For patterns placed in the alarm class, the average positive discriminant score is plotted. For patterns classified as non-alarm, the average of the negative discriminant scores is plotted. As seen in the figure, no missed or false alarms were observed, and the trace of the discriminant scores about the 0.0 decision threshold matches the trace of the concentrations about the alarm threshold of 3.0 mM.

Classification Performance with Prediction Sets

Classifiers were next developed to test the implementation of the alarm algorithm with the three prediction sets. The parameter settings of 4650-4300 cm^{-1} and eight latent variables selected from the work with the monitoring data were again employed. The full set of calibration data based on runs 1-11 was used to define the calibration database for the development of the classifiers.

The reference concentration for prediction set 1 was 4.6 mM and the critical concentration was -1.6 mM. Differential spectra were generated relative to the reference spectrum and the PLS factors previously computed with the calibration data were used to compute the corresponding score vectors. The calibration patterns were partitioned on the basis of the critical concentration into a training set containing 28,469 alarm patterns and 22,800 non-alarm patterns. Three replicate classifiers were computed with the training set. As with the monitoring set, each classifier contained a single discriminant function. Across the three replicate classifiers, an average of 85 ± 0.01 % of the alarm patterns in the calibration set were separated. Discriminant scores were then computed for the differential spectra in prediction set 1, producing no missed or false alarms.

The same procedures were used for prediction set 2. The reference glucose concentration was 3.9 mM, and the corresponding critical concentration was -0.9 mM. The data partitioning based on this critical concentration resulted in 35,062 alarm patterns and 16,207 non-alarm

patterns. Approximately 80 ± 0.03 % of the alarm patterns of the calibration set were separated with a single discriminant function. Applying the decision rule of two out of three classifiers signaling an alarm, there were no missed or false alarms for this prediction set.

The critical concentration for prediction set 3 was -1.1 mM, and the calibration database was partitioned into 31,575 alarm patterns and 19,694 non-alarm patterns. Approximately 77 ± 0.01 % of the alarm patterns of the calibration set were separated with a single discriminant function. Fig. 7 shows the glucose concentration profile for prediction set 3 in which there were 62 non-alarms and 71 alarm patterns. The first and the third classifiers predicted no missed or false alarms, while the second classifier predicted one missed alarm and no false alarms. Applying the alarm decision rule for the combined use of the three replicate classifiers gave no missed or false alarms. The trace of the combined discriminant scores is also shown in Fig. 7 (left y-axis).

Fig. 8 plots the three replicate discriminant scores with respect to the corresponding differential glucose concentrations for prediction set 3. A clear relationship between the discriminant scores and differential concentrations is observed, and the intersection of the discriminant score threshold of 0.0 with the critical concentration of -1.1 mM can be seen from the plotted reference lines. The advantage of using a discriminant approach rather than a formal quantitative concentration model for implementing the alarm can also be observed. There is no need to build a precise predictive model for concentration when the question being addressed is one of classification (i.e., whether the concentration is above or below a threshold value).

Finally, Fig. 9 addresses the utility of the calculation of differential spectra in helping to maintain the viability of the calibration data with time. The first three PLS scores for the calibration data and prediction set 3 are plotted in the figure. The calibration and prediction patterns cluster together in the same data space. When one considers the data from prediction set 3 were collected 2.5 years after the end of the collection of the calibration data, the lack of

evidence of instrumental drift is directly attributable to the use of differential spectra in developing the methodology.

Conclusions

In this paper, a threshold monitor algorithm based on differential near-IR spectra was tested with a simulated process. A synthetic sample matrix was constructed from glucose, urea, lactate, and triacetin in phosphate buffer to provide a challenge to the ability to extract glucose information selectively from near-IR spectra in the combination region.

This study provided a first test of one of the key components of the alarm algorithm, the use of differential spectra computed relative to a glucose-containing reference spectrum. Within the calibration data, the calculation of all combinations of spectral ratios within time blocks served to expand the concentration data space. Further, the calculation of differential spectra served to simplify the resulting absorbance spectra by removing constant features of the spectral background.

The successful use of the PLDA method to implement the alarm decision provided verification that a pattern classification approach can be employed to identify concentration levels within near-IR spectra. The iterative nature of the training of the classifiers raises the possibility that the optimization may become trapped in local maxima. By training three replicate discriminants and using them together to implement the alarm decision, the overall robustness of the alarm algorithm was improved and the potential problem of training variance was effectively addressed.

The developed alarm algorithm was tested with three external prediction sets. The results obtained were very promising. No missed or false alarms were observed for any of the prediction sets. The robustness of the methodology was also tested by collecting prediction set 3 two years and six months later than the last day of the calibration data. The result for this prediction set (i.e., no missed or false alarms) clearly demonstrates the excellent robustness of the methodology

to changes in instrumental characteristics with time. The use of differential spectra computed relative to a same-day reference was considered to be a key component of the observed robustness of the methodology.

While the results presented in this paper were extremely promising, several qualifying considerations must be underscored. First, even though the calculation of differential absorbance values helped to simply the background contributions present in the resulting spectra, all of the conventional requirements for a successful multivariate calibration remained. For example, the calibration data must match the future data to which the method will be applied. Unpredictable results will likely be obtained if an unknown component with varying concentration (i.e., varying relative to the reference spectrum) is introduced. Second, the chemical system used here was relatively straightforward and thus did not offer the complexities of a real biological process that might be monitored. Thus, there was no need to incorporate spectral outlier detection or to adopt a more sophisticated decision-making algorithm regarding when to identify the alarm state. These represent areas of future investigation.

Finally, while the methodology presented here was based on a low threshold application, the same procedures could be used to detect when a concentration exceeds a high threshold. In this case, the differential spectra would be computed to yield positive differential concentrations and the critical concentration would be positive. All of the other procedures would be the same as described here. Implementing classification models for both high and low thresholds could also be done to develop an alarm system for maintaining concentrations between specified upper and lower limits.

Acknowledgements

Cynthia Medford and Wei Wang are acknowledged for performing initial feasibility studies that led to the work presented in this paper.

References

1. A. M. Brearley, in *Process Analytical Technology*, ed. K. A. Bakeev, Blackwell Publishing Ltd., Oxford, UK, 2005, pp. 392-423.
2. N. W. Broad, R. D. Jee, A. C. Moffat and M. R. Smith, *Analyst (Cambridge, U.K.)*, 2001, 126, 2207-2211.
3. W. Camacho, A. Valles-Liuch, A. Ribes-Greus and S. Karlsson, *J. Appl. Polym. Sci.*, 2003, 87, 2165-2170.
4. T. Davies, in *Analusis*, 1998, vol. 26, pp. M17-M19.
5. R. A. Forbes, M. L. Persinger and D. R. Smith, *J. Pharm. Biomed. Anal.*, 1996, 15, 315-327.
6. C. P. S. Kuda-Malwathumullage and G. W. Small, *J. Appl. Polym. Sci.*, 2014, 131, in press.
7. A. S. Zidan, M. J. Habib and M. A. Khan, *J. Pharm. Sci.*, 2008, 97, 3388-3399.
8. A. K. Amerov, J. Chen and M. A. Arnold, *Appl. Spectrosc.*, 2004, 58, 1195-1204.
9. D. M. Haaland and E. V. Thomas, *Anal. Chem.*, 1988, 60, 1193-1202.
10. I. T. Jollife, *Principal Component Analysis*, Springer-Verlag, New York, 1986.
11. S. Wold, K. Esbensen and P. Geladi, *Chemom. Intell. Lab. Syst.*, 1987, 2, 37-52.
12. F. Ehrentreich, *Anal. Bioanal. Chem.*, 2002, 372, 115-121.
13. C. Cheng, W. Xiong and Y. Tian, *Chin. J. Chem. FIELD Full Journal Title:Chinese Journal of Chemistry*, 2009, 27, 911-914.
14. C. L. Hammer, G. W. Small, R. J. Combs, R. B. Knapp and R. T. Kroutil, *Anal. Chem.*, 2000, 72, 1680-1689.
15. Y. Shao, Y. He, Y. Bao and J. Mao, *Int. J. Food Prop. FIELD Full Journal Title:International Journal of Food Properties*, 2009, 12, 644-658.
16. L. Liang, B. Wang, Y. Guo, H. Ni and Y. Ren, *Vib. Spectrosc.*, 2009, 49, 274-277.

17. Y. Zhang, Q. Cong, Y. Xie, J. Yang and B. Zhao, *Spectrochim. Acta, Part A FIELD Full Journal Title:Spectrochimica Acta, Part A: Molecular and Biomolecular Spectroscopy*, 2009, 71A, 1408-1413.

18. T. F. Kaltenbach and G. W. Small, *Anal. Chem.*, 1991, 63, 936-944.

19. R. E. Shaffer and G. W. Small, *Chemom. Intell. Lab. Syst.*, 1996, 32, 95-109.

20. R. E. Shaffer and G. W. Small, *Anal. Chim. Acta*, 1996, 331, 157-175.

21. Y. Sulub and G. W. Small, *Appl. Spectrosc.*, 2008, 62, 1049-1059.

22. B. Wan and G. W. Small, *Analyst*, 2011, 136, 309-316.

Table 1 Results of grid search optimization of spectral range and latent variables

Spectral range ($cm^{-1)}$	Latent variables	SECV[a] (mM)
4650-4250	16	0.32_2
4650-4300	15	0.32_4
4700-4250	16	0.32_4
4700-4300	15	0.32_5

[a]Standard error of cross-validation obtained from the grid search optimization.

Table 2 Average percentages of missed and false alarms for the monitoring set

Spectral range (cm^{-1})		Number of Latent Variables					
		6	7	8	9	10	11
4650-4250	AM (%)[a] ± S.D	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	AF (%)[b] ± S.D	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	DS (%)[c] ± S.D	82.5 ± 0.0	85.9 ± 6.2	95.1 ± 1.4	97.7 ± 1.4	98.1 ± 0.7	99.7 ± 0.0
4700-4250	AM (%)[a] ± S.D	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	AF (%)[b] ± S.D	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	DS (%)[c] ± S.D	75.5 ± 0.9	89.0 ± 0.1	88.2 ± 3.4	98.4 ± 0.3	97.9 ± 1.3	98.4 ± 0.9
4650-4300	AM (%)[a] ± S.D	0.0 ± 0.0	0.0 ± 0.0	**0.0 ± 0.0**	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	AF (%)[b] ± S.D	0.0 ± 0.0	0.0 ± 0.0	**0.0 ± 0.0**	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	DS (%)[c] ± S.D	85.4 ± 0.6	94.8 ± 0.6	**96.0 ± 0.0**	99.0 ± 0.1	99.7 ± 0.1	99.8 ± 0.0
4700-4300	AM (%)[a] ± S.D	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	AF (%)[b] ± S.D	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	DS (%)[c] ± S.D	83.2 ± 0.3	88.5 ± 0.1	93.8 ± 0.0	93.0 ± 1.7	98.7 ± 0.6	99.6 ± 0.2

[a]Average percentage of missed alarms (AM) ± standard deviation.

[b]Average percentage of false alarms (AF) ± standard deviation.

[c]Average percentage of separated alarm patterns with a single discriminant (DS) ± standard deviation.

Figure Captions

Fig. 1 Differential spectra of glucose in 0.1 M, pH 7.4 phosphate buffer for both positive (dashed line) and negative (solid line) concentrations of 20.0 mM, computed by taking the negative logarithm of the ratio of two single-beam spectra containing glucose. The spectral features can be either positive or negative depending on the concentrations corresponding to the spectra used in the numerator and denominator of the absorbance calculation. Glucose combination bands near 4300 (C-H), 4400 (C-H), and 4650 (O-H) cm^{-1} are visible in the spectra.

Fig. 2 Flow charts describing the steps used in building the calibration database (left) and the calibration procedure used at the start of the monitoring period (right). The calibration database consists of a PLS score matrix ($n \times h$) computed from using n differential spectra and corresponding differential concentrations to produce h PLS latent variables. In the right diagram, C_{ref} is the reference analyte concentration obtained at the start of the monitoring period. The alarm threshold concentration, C_{alarm}, is 3.0 mM for these experiments. The difference between C_{alarm} and C_{ref} is termed the critical concentration, C_{crit} (Eq. 5).

Fig. 3 Flow chart of the operation of the alarm. A spectrum is collected at time t and the ratio is taken to the reference spectrum to compute a differential spectrum. Projection of the differential spectrum onto the calibration PLS factors yields a pattern (i.e., $\mathbf{t}_{dif,t}$) which is classified using the previously computed discriminants. If the pattern is classified into non-alarm class, the process repeats. If the pattern is placed into the alarm class, an alarm is sounded to alert the operator.

Fig. 4 Glucose concentration profiles for the study. The labels denote the subdivision of the data into groups for calibration, calibration testing (monitoring), and external prediction. The horizontal dashed line denotes the threshold monitor concentration of 3.0 mM used in this work.

Fig. 5 Cross-validation results (SECV) vs. the number of latent variables for the optimal wavenumber range of 4650-4250 cm^{-1}. Calibration models were based on PLS analysis of differential spectra and concentrations in the calibration subset.

Fig. 6 The discriminant scores corresponding to the combined use of the three replicate classifiers with the monitoring data are shown (diamonds, left y-axis) with the reference glucose concentrations superimposed (solid trace, right y-axis). The horizontal line corresponds to the alarm/non-alarm thresholds for the discriminant scores (0.0) and reference concentrations (3.0 mM). There were 48 and 58 spectra in the alarm and non-alarm data classes, respectively. No missed or false alarms were observed.

Fig. 7 The discriminant scores corresponding to the combined use of the three replicate classifiers with prediction set 3 are shown (diamonds, left y-axis) with the reference glucose concentrations superimposed (solid trace, right y-axis). The horizontal line correspond to the alarm/non-alarm thresholds for the discriminant scores (0.0) and reference concentrations (3.0 mM). There were 71 and 62 alarm and non-alarm patterns, respectively. No missed or false alarms were observed.

Fig. 8 Discriminant scores are plotted with respect to differential glucose concentrations for prediction set 3. Circles, "+" symbols, and squares denote the discriminant scores produced by the three replicate classifiers. A clear relationship between discriminant scores and differential concentrations is noted. Reference lines allow the intersection between the critical concentration of -1.1 mM and the discriminant score alarm threshold of 0.0 to be seen.

Fig. 9 First three PLS scores plotted for the calibration data set and prediction set 3. Blue circles, red squares, and green triangles denote the calibration patterns from DS1, the calibration patterns from DS2, and the patterns from prediction set 3, respectively. Clear overlap of the patterns is noted. This verifies that the calibration and prediction data are consistent, even with a separation in time of 2.5 years.

Fig. 1

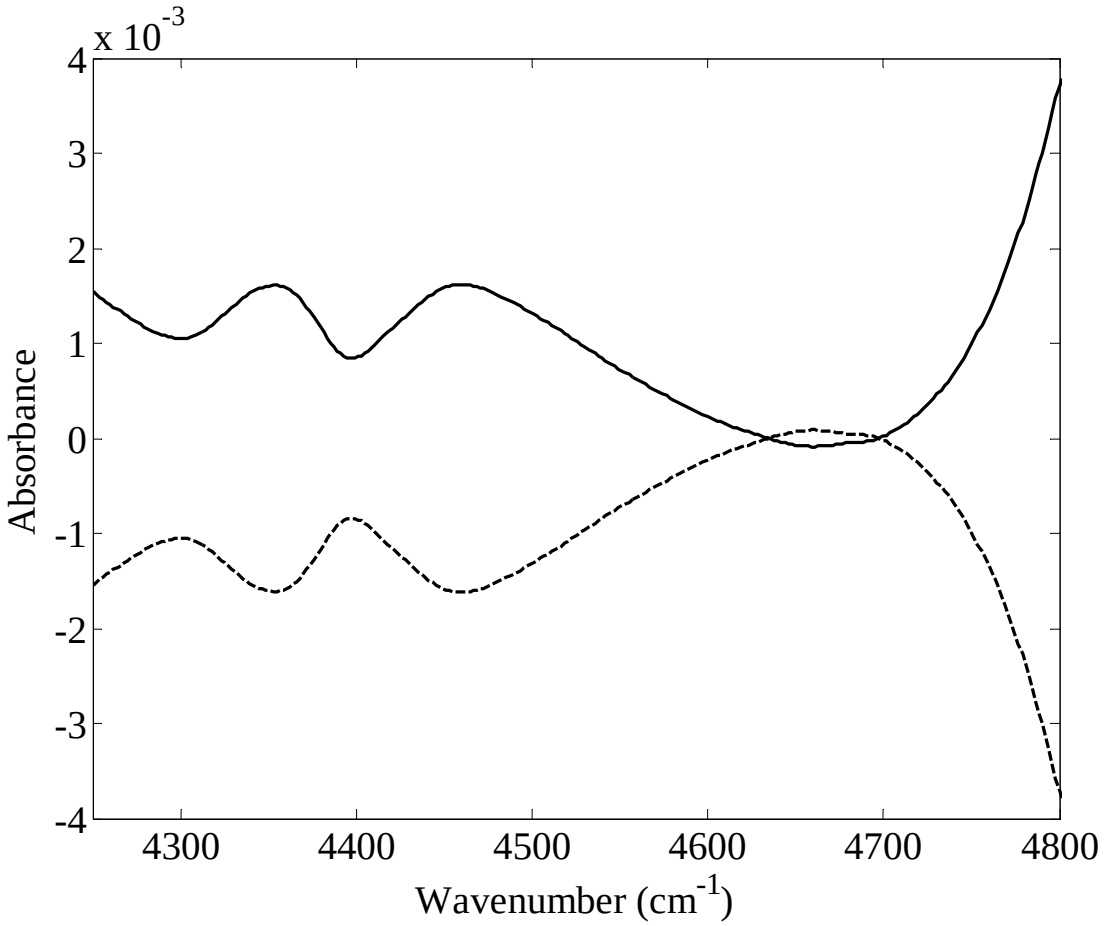

x 10^-3
4
3
2
1
0
-1
-2
-3
-4
Absorbance
4300
4400
4500
4600
4700
4800
Wavenumber (cm^-1)

Fig. 2

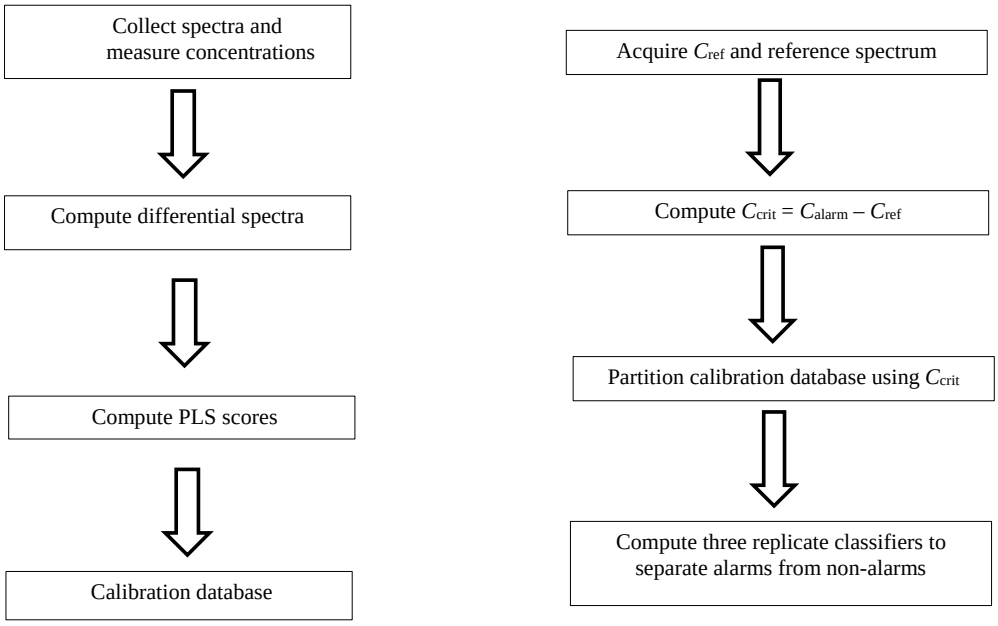

Collect spectra and measure concentrations
Compute differential spectra
Compute PLS scores
Calibration database
Acquire C_{ref} and reference spectrum
Compute $C_{crit} = C_{alarm} - C_{ref}$
Partition calibration database using C_{crit}
Compute three replicate classifiers to separate alarms from non-alarms

Fig. 3

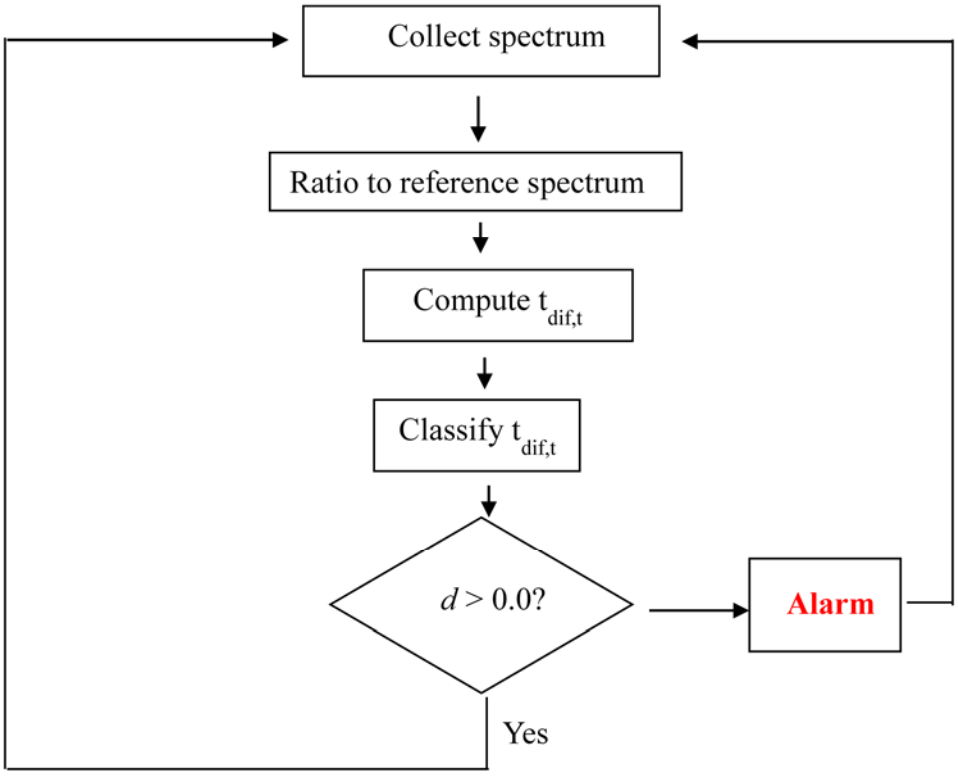

Collect spectrum
Ratio to reference spectrum
Compute t dif,t
Classify t dif,t
d > 0.0?
Alarm
Yes

Fig. 4

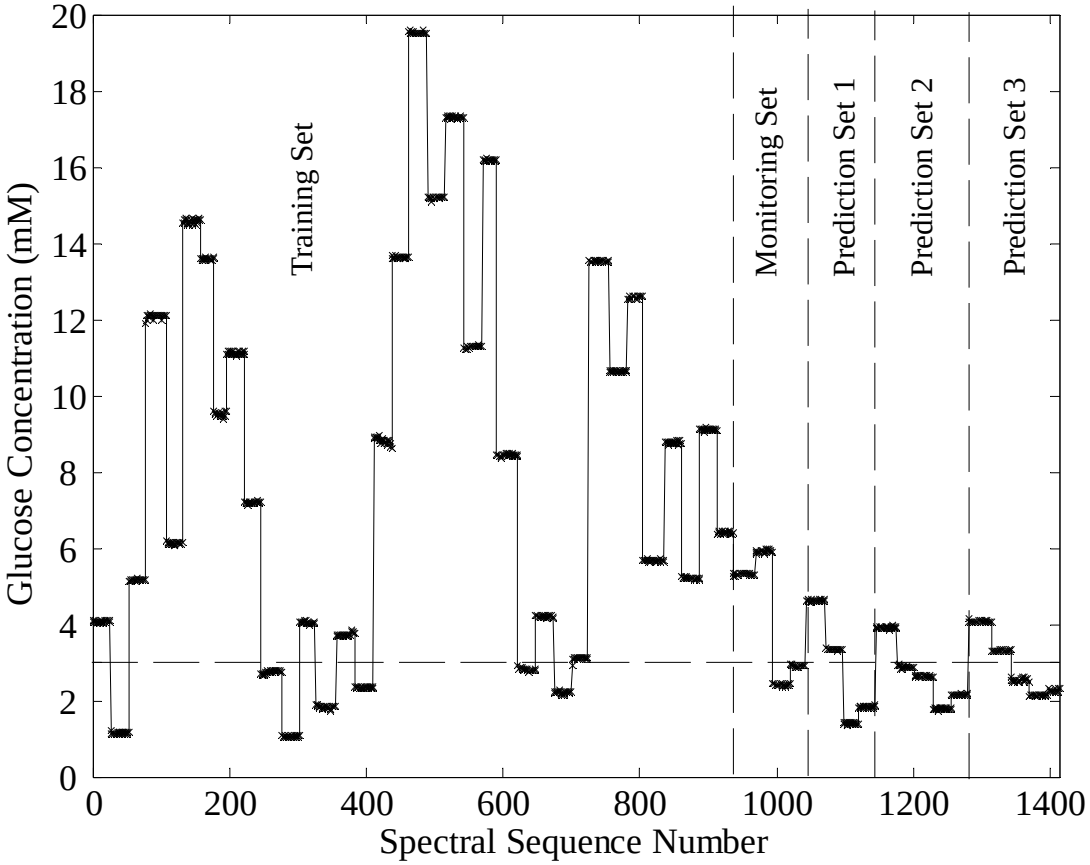

Glucose Concentration (mM)
Spectral Sequence Number
Training Set
Monitoring Set
Prediction Set 1
Prediction Set 2
Prediction Set 3
0
2
4
6
8
10
12
14
16
18
20
0
200
400
600
800
1000
1200
1400

Fig. 5

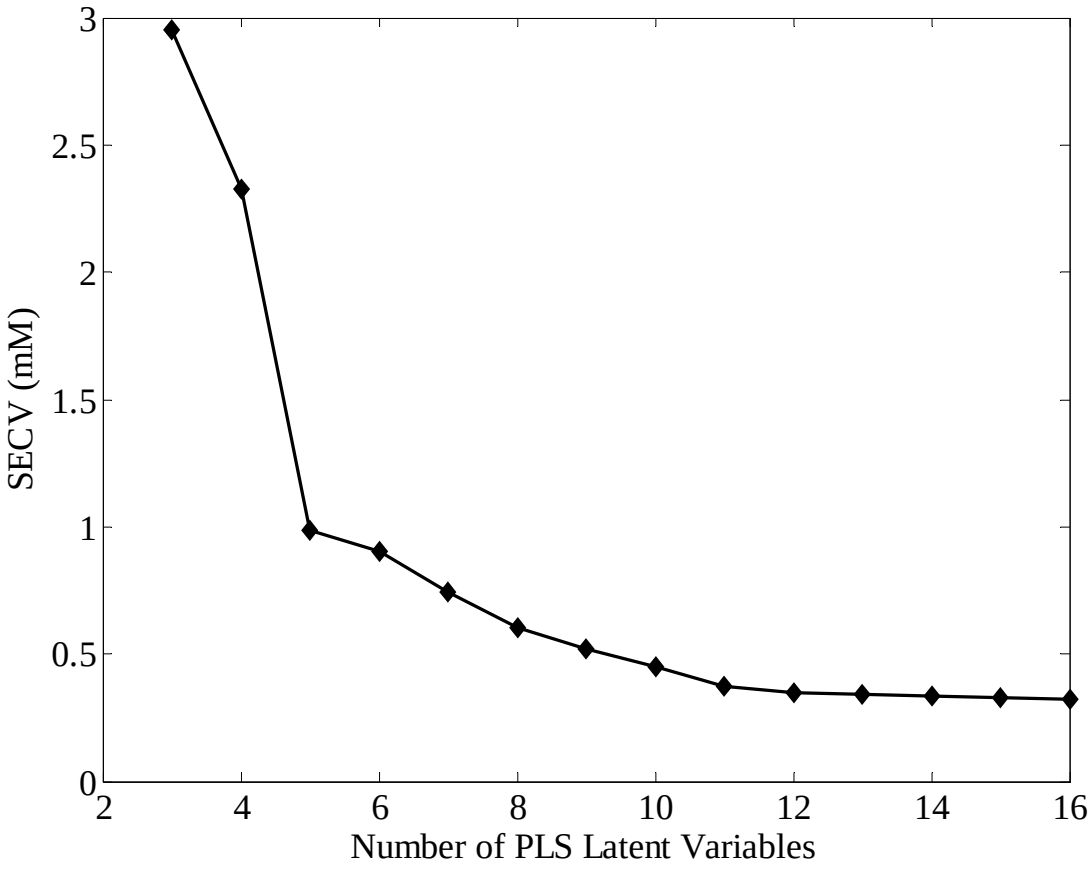

SECV (mM)
Number of PLS Latent Variables
0
0.5
1
1.5
2
2.5
3
2
4
6
8
10
12
14
16

Fig. 6

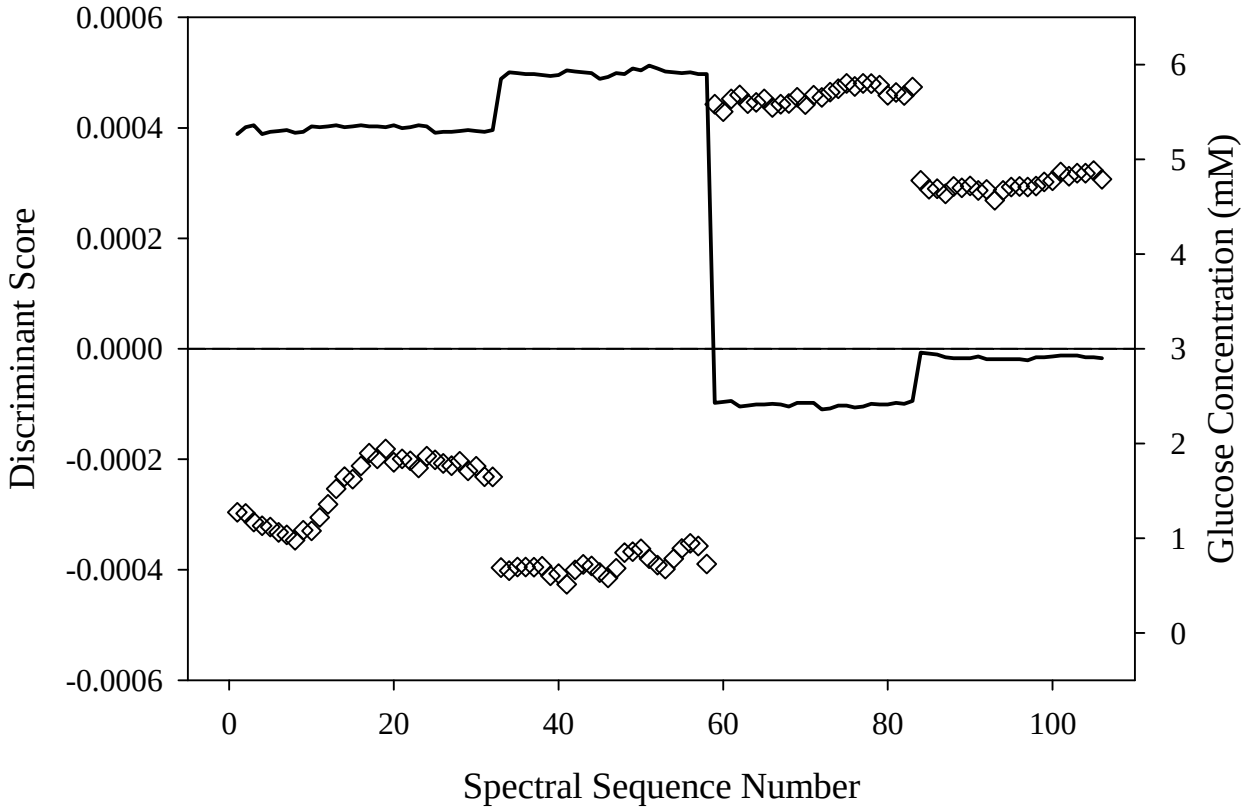

0.0006
0.0004
0.0002
0.0000
-0.0002
-0.0004
-0.0006
Discriminant Score
0
20
40
60
80
100
Spectral Sequence Number
6
5
4
3
2
1
0
Glucose Concentration (mM)

Fig. 7

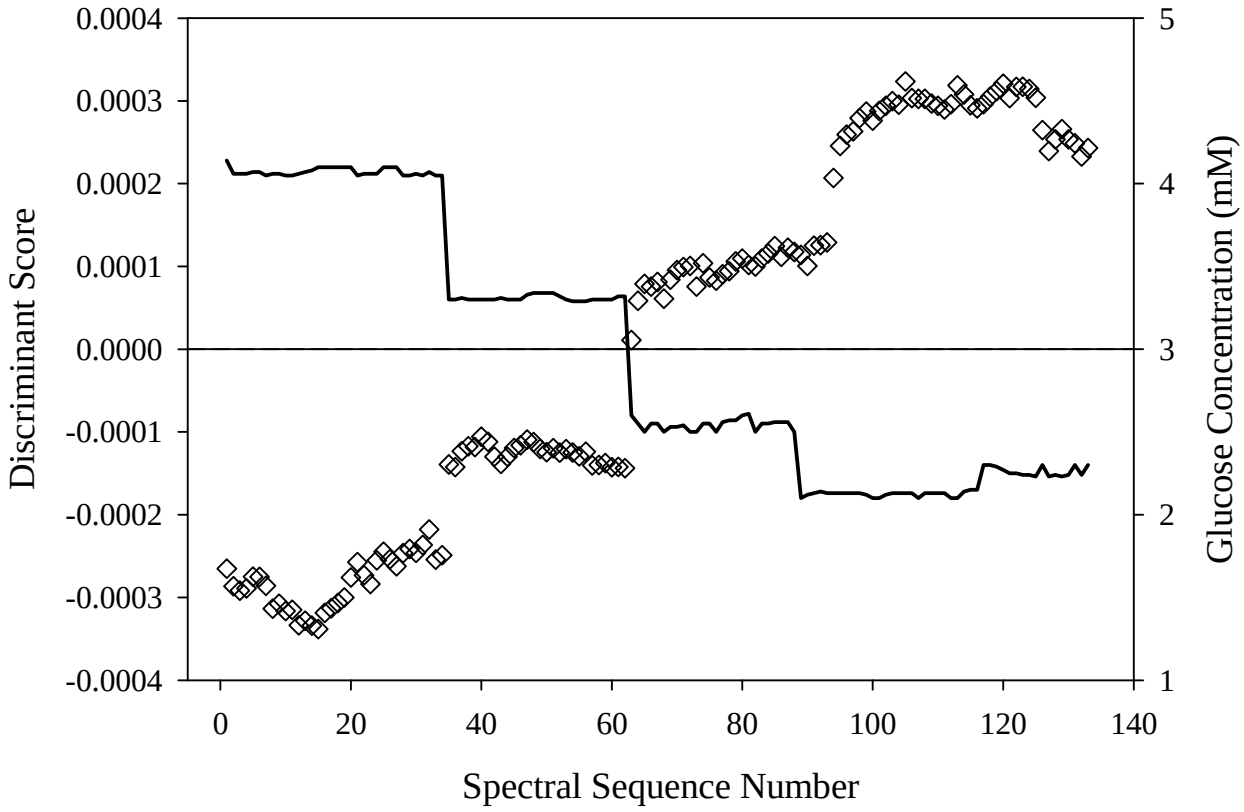

0.0004
0.0003
0.0002
0.0001
0.0000
-0.0001
-0.0002
-0.0003
-0.0004
Discriminant Score
5
4
3
2
1
Glucose Concentration (mM)
0
20
40
60
80
100
120
140
Spectral Sequence Number

Fig. 8

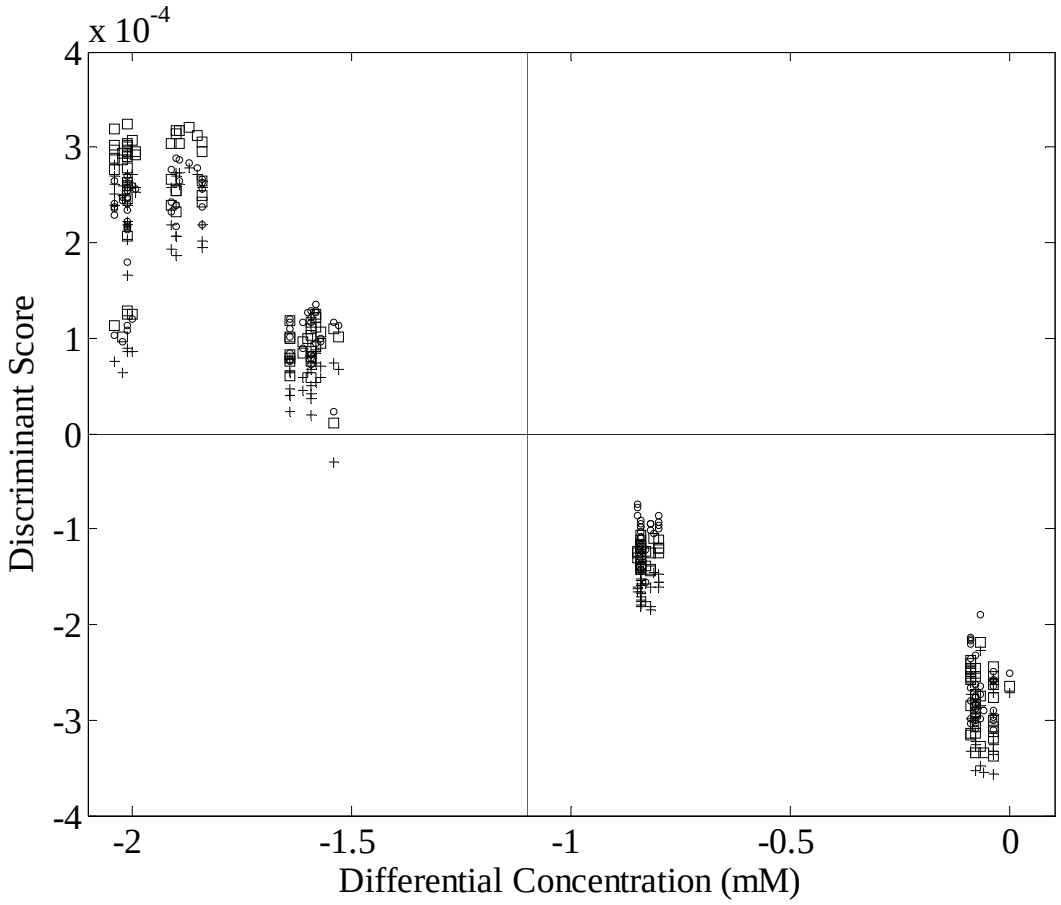
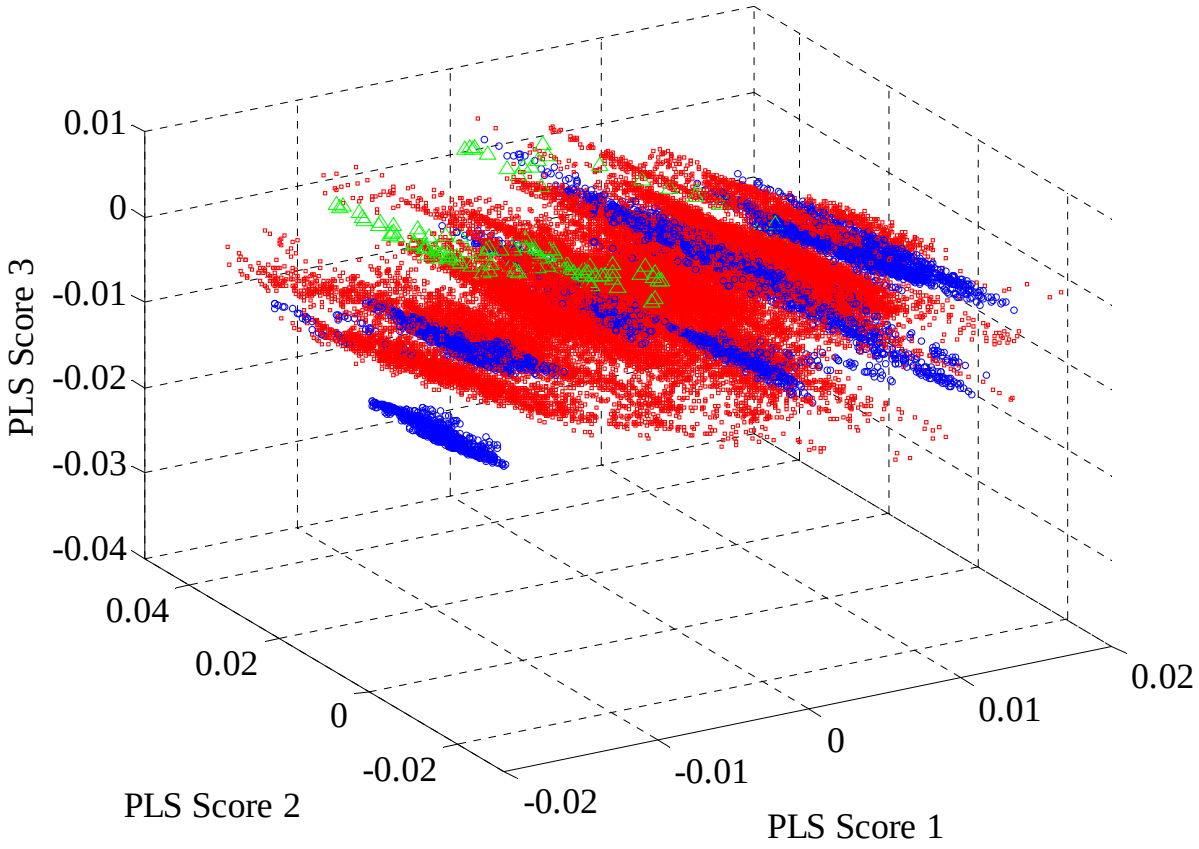
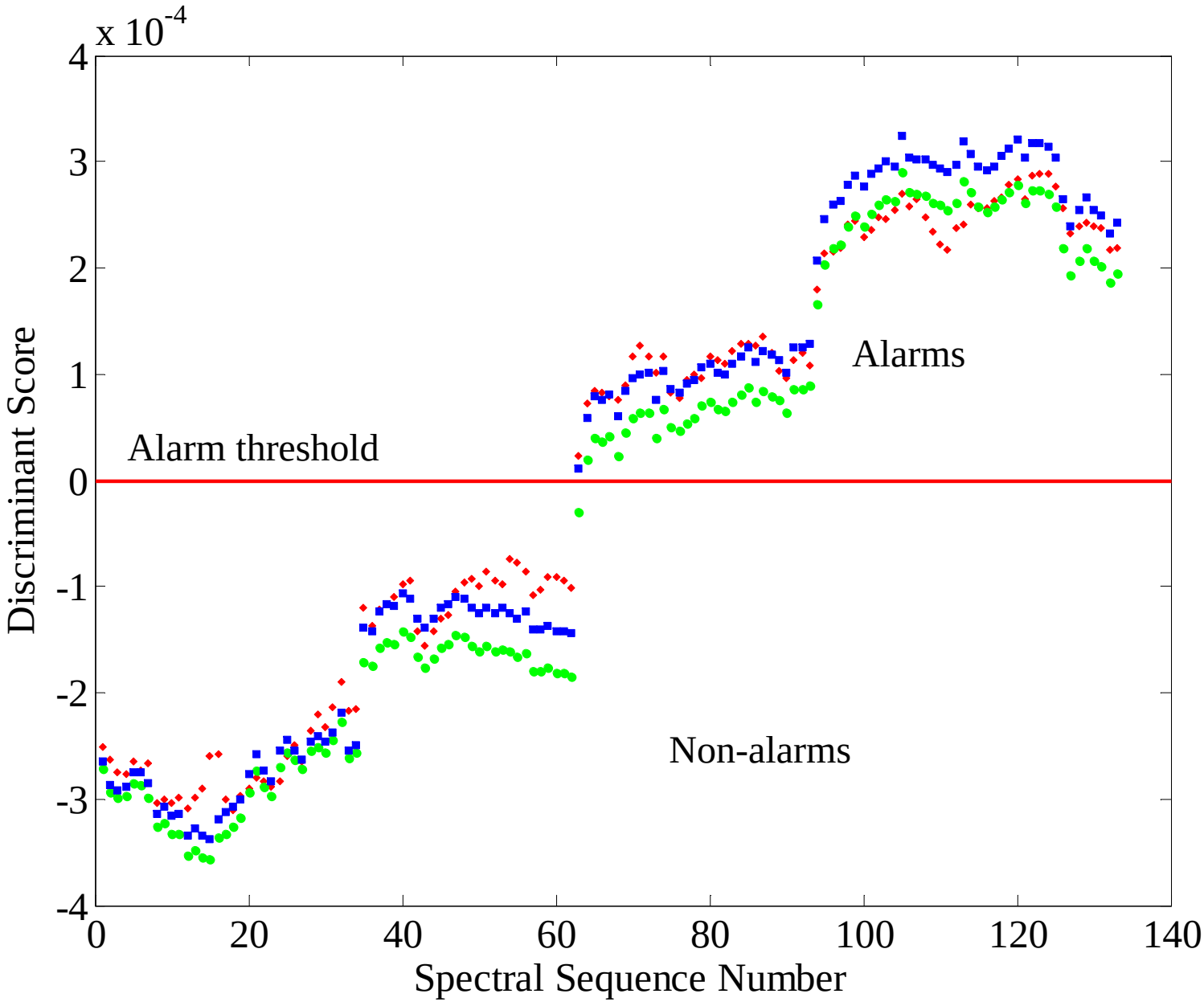

x 10-4
4
3
2
1
0
-1
-2
-3
-4
Discriminant Score
-2
-1.5
-1
-0.5
0
Differential Concentration (mM)

Fig. 9


0.01
0
-0.01
-0.02
-0.03
-0.04
PLS Score 3
0.04
0.02
0
-0.02
PLS Score 2
-0.02
-0.01
0
0.01
0.02
PLS Score 1


x 10-4
4
3
2
1
0
-1
-2
-3
-4
Discriminant Score
Alarm threshold
Alarms
Non-alarms
0
20
40
60
80
100
120
140
Spectral Sequence Number