Development of a non-linear statistical routine for optical filter optimization

Denis Lafrance and David H. Burns

Department of Chemistry, McGill University, Otto Maass Building, Montreal, Quebec, H3A 2K6, Canada.

Introduction

Optical measurements of light absorbing substances are the basis for many clinical and experimental analyses used in medical physiology and laboratory testing. Optically based chemical analysis relies on predictable interaction of light with chemical constituents in samples. Most analytical methods rely either upon fluorescence/phosphorescence emission intensity being linear functions of concentration of the emitting constituent, or upon transmitted light intensity being exponential functions of concentration of an absorbing constituent as described by the Beer–Lambert Law.

The key to produce low cost medical instrumentation based on spectroscopic techniques is the conversion of methods developed on high-resolution, high-cost scanning or diode-array spectrophotometers to low cost filter or diode spectrometers. This is particularly applicable to instruments designed for clinical laboratory assays, high speed sensitive physiological measurements and medical imaging methods using interference filters for definition of wavelength regions.

However, most of the work regarding the design of filter-based spectroscopic instruments for chemical analysis has relied on researchers' best estimates of relationships between desired properties and multiwavelength spectra. Spectra were analyzed for absorption maxima, minima and isosbectic points. If no isosbectic points existed in absorption spectra, wavelengths that exhibited minimal absorbance change and wavelengths of maximal absorbance change were used. Methods used to determine wavelengths and bandwidths were essentially "best guess".

A way to increase the objectivity in filter selection is the use of statistical regression. There are two general approaches to the statistical regression analysis—(i) methods that utilize full spectra for transformation such as classical least squares (CLS) or partial least squares (PLS), or (ii) methods that select individual wavelengths with the highest correlation coefficients for the quality factors, such as stepwise linear regression (SMLR). Since SMLR selects individual wavelengths, the correlation with center wavelength filter could be seen as easy to do.

However, SMLR has a disadvantage, since it selects only optimal center wavelength combinations for calibration. For quantitative work, both wavelength and bandwidth of filters are critical for maximum analytical accuracy with minimum influence from interfering absorbing or fluorescing species.

The filter shape is gaussian in intensity and is increasing linearly in transmittance. The I_0 / I ratio corresponds to the area under the curve. The sum of the dot product (**a . b**) of each wavelength of the spectrum, with the corresponding wavelength of the filter, will give a single value for each spectrum.

The aim of the project is to develop a statistical routine for the selection of optimal optical filters, to predict the temperature of a water sample, using near infrared (NIR) spectra. The water

From Near Infrared Spectroscopy: The Future Waves
© IM Publications Open LLP 1996

data set will be divided in two sets, a training set and a test set, and the quality factor will be the temperature at which the spectra were recorded.

Experimental

The proposed method is the following: raw spectra of a water sample at different temperature will be acquired at high resolution. Spectra will not be preprocessed by smoothing but the untreated training sets will be used to calculate optimal filters with varying center wavelengths and bandwidths. Information contents from sharp and broad spectral features are, therefore, compromised reducing the accuracy of prediction. In our approach, each predicted filter has an associated bandwidth which produces the most robust prediction.

Results and discussion

Water was chosen to develop the statistical routine, because the spectral properties of water have been studied extensively and its near infrared spectrum has simple absorption bands in the region of interest. The 1100–1300 nm region was chosen. Two peaks can be distinguished in the broad band seen in Figure 1. The first peak at around 1160 nm is attributed to free hydrogen-bonded O–H groups vibration that originate from the combination band of the first overtone of ν_1 (symmetric stretching) and ν_3 (asymmetric stretching), and ν_2 (bending). The second peak at around 1200 nm is attributed to hydrogen-bonded O–H groups vibration of the same combination band ($2\nu_{1,3} + \nu_2$). An isosbectic point is also observed between these two bands. One can see the 1160 nm peak increase with temperature, while the 1200 nm peak decrease (red spectrum compares to blue spectrum). This peak increase corresponds to the disruption of the intermolecular hydrogen bonding network between O–H groups and so, to the increase of free O–H groups vibration. Finally, a third band has also been identified around 1250 nm. This band originates from

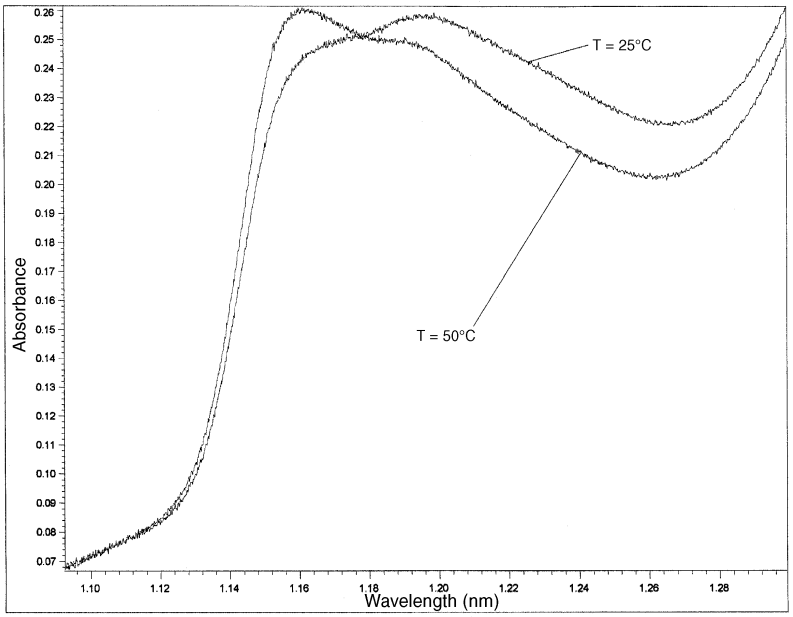

Figure 1. Water spectra.

From Near Infrared Spectroscopy: The Future Waves
© IM Publications Open LLP 1996

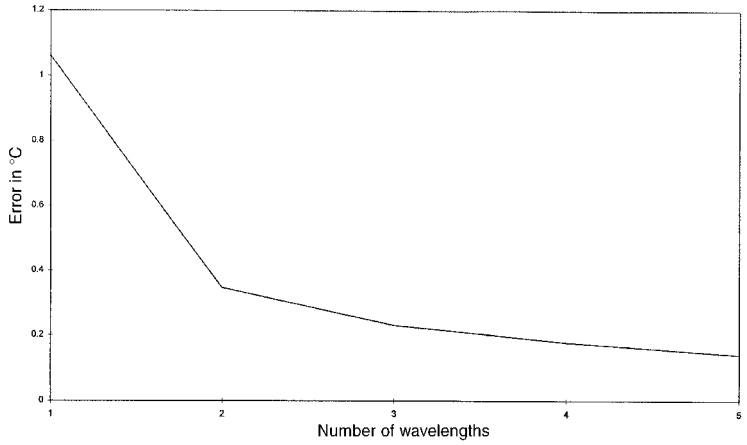

Figure 2(a). Error in °C vs number of wavelengths.

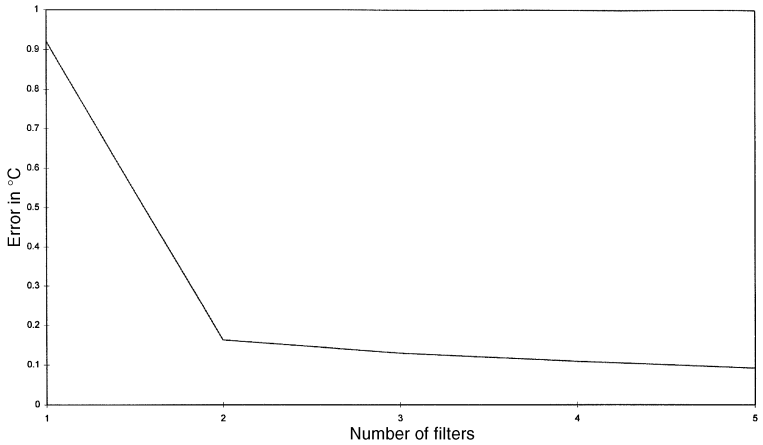

Figure 2(b). Error in °C vs number of filters.

the same combination band, but where the two O–H groups of the water molecule are hydrogen-bonded. This band is more likely seen in ice spectra than liquid water spectra.

Figure 2(a) shows the error graph with SMLR for the water sample. Two wavelengths make the most significant change for error minimization. By adding other wavelengths, the error continues to decrease, but not significantly. The two wavelengths choose by SMLR are 1153 and 1255 nm.

Figure 3 shows the temperature prediction of the test data set using equations built with the training data set. The correlation coefficient is good. However, the bandwidth parameter is not part of the training data set equations. The development of a non-linear statistical routine will introduce this parameter.

The analysis procedure used with the non-linear algorithm can be described as a series of steps. Absorbance raw data were first converted to transmittance data. Spectral data and correlated

From Near Infrared Spectroscopy: The Future Waves
© IM Publications Open LLP 1996

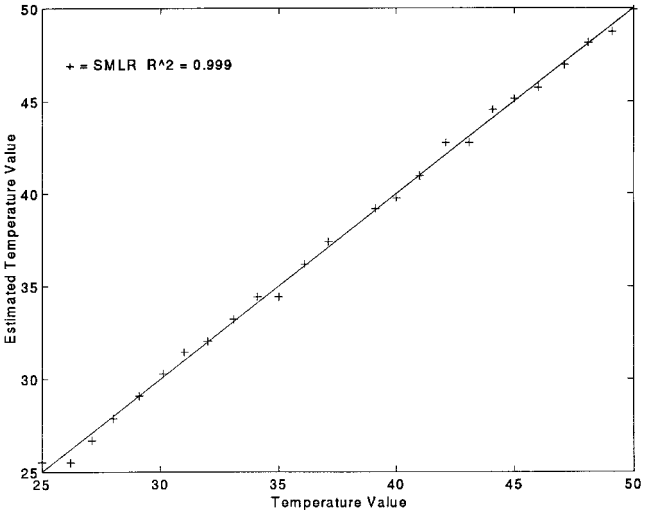

Figure 3. Temperature prediction.

quality factors were then read into a two-dimensional data array. Data were then randomly separated into two sets, a training and a test set. The number of filters required for prediction of quality factors was determined first. The training set was used for the development of prediction equations and the test set for evaluating the error of prediction. The initial filter center wavelengths were the values found using SMLR algorithm. The bandwidth was initially set arbitrarily at 30 nm, with an option to modify it if needed. The initial values for filter parameters were input into a simplex optimization procedure. Finally, a least-squared routine was used to minimize the error, using the log of transmittance responds, for the simulated filter instrument.

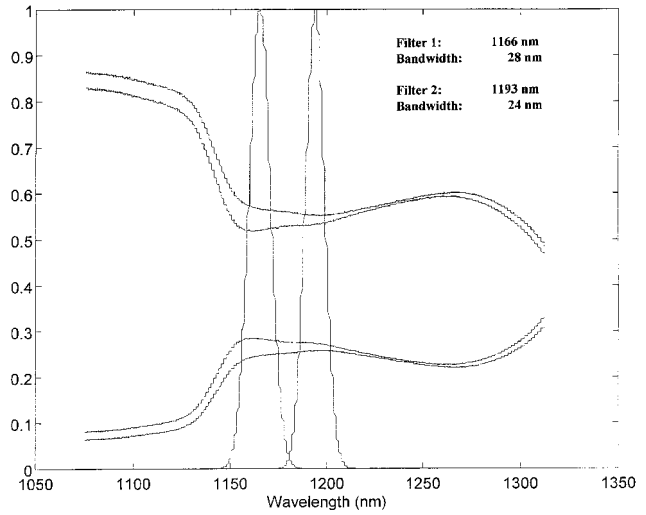

Figure 4. Optimal filters chosen.

From Near Infrared Spectroscopy: The Future Waves
© IM Publications Open LLP 1996

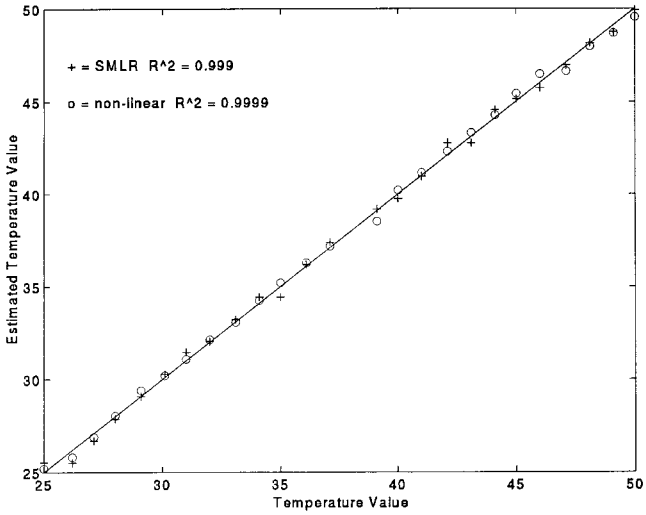

Figure 5. Temperature prediction.

Figure 2(b) shows the error graph using the non-linear routine for the water sample. Again, two filters make the most significant change for error minimization and describe most of the variation. By adding other wavelengths, the error continues to decrease, but not significantly.

Figure 4 shows the two filters chosen on top of both absorbance and transmittance spectra. If the first filter chosen by the non-linear routine is close to what was picked up by SMLR (1166 and 1153 nm), the second filter is different (1193 and 1255 nm). By SMLR, the filters were chosen where most of the change occurs (1153 nm) and where almost no change occurs (1255 nm). This is not the case with the non-linear routine. It should be seen also that the two filters that minimize error have different bandwidths.

Figure 5 shows the temperature prediction using the non-linear routine in comparison with SMLR. The correlation coefficient is slightly better. However, it does not mean the non-linear routine is superior to SMLR. The non-linear routine has allowed the introduction of the filter bandwidth parameter into the training set equation.

Conclusions

In conclusion, this project has demonstrated the feasibility of the filter optimization statistical engine by specifying filters for temperature determination using absorption spectroscopy of water. The system will be extended to the more physiologically important case where interfering substances are present. Analysis of complex biological fluids and tissues involves noise from light scatter, absorbance from many interfering substances as well as instrument noise. More filters, each adding another term to the prediction polynomial, may be required to minimize these interferences.

From Near Infrared Spectroscopy: The Future Waves
© IM Publications Open LLP 1996