Multivariate Analysis of Near-Infrared Spectra Using the G-Programming Language

Olusola O. Soyemi, Marianna A. Busch, and Kenneth W. Busch*

Center for Analytical Spectroscopy, Baylor University, P.O. Box 97348, Waco, Texas 76798-7348

Received February 21, 2000

"Real-time" chemometrics as envisioned by the union of instrument control, data acquisition, and chemometric analysis with a single software platform can provide substantial benefits to manufacturing concerns that require process control. Some of these benefits include faster generation of information and improved quality control. This paper describes a series of chemometric routines written in LabVIEW and demonstrates their use in predicting six properties of diesel fuel. In particular, near-infrared spectral data were used to predict the boiling point at 50% recovery, cetane number, density, freezing temperature, total aromatics, and viscosity for a series of diesel fuels.

INTRODUCTION

LabVIEW (Laboratory Virtual Instrument Engineering Workbench) is a graphical programming environment that uses the G-programming language to create data acquisition and analysis and visualization software that are collectively known as virtual instruments (VIs). LabVIEW programs are constructed from hierarchical combinations of VIs and sub-VIs (i.e., VIs at a lower level in a calling sequence), "wired" together graphically to exchange data values or control information.[1] All virtual instruments are composed of a front panel and a block diagram. The front panel provides the user interface for interactive operation, while the block diagram (which is visualized as lying behind the front panel) contains the source code. The source code consists of graphical icons of lower-level instruments (i.e., sub-VIs) and program control structures. The advantages of G-Programming over conventional text-based languages such as Fortran and C include the provision of standardized software modules (toolkits) that would otherwise need to be coded explicitly, and an attractive and easy to use development environment in which source code is represented graphically.

In applying LabVIEW to problems involving analytical instrumentation, its primary utility is usually thought to lie mainly in instrument control and data acquisition. However, frequent updates to the G-programming environment have led to the development of embedded tools/toolkits that have constantly opened up new possibilities as far as data analysis is concerned. A recent application of LabVIEW programming, where instrument control and analysis functions are intertwined, is the use of genetic algorithms written in LabVIEW for instrumentation control and optimization.[2] These algorithms have been applied to closed-loop control instruments, which have a variety of applications in the biomedical sciences, such as the regulation of physiological processes.

A program written in LabVIEW has also been applied in plasma diagnostics to determine the electron number density from Stark broadening measurements of the hydrogen β line.[3] Using this program, the electron number density in a glow discharge was calculated for two different operating conditions.

One of the data analysis techniques to which LabVIEW has been routinely applied in the recent past is Statistical Process Control (SPC).[4] SPC techniques are used to monitor the mean and variability of a process with the aim of keeping the various manufacturing variables within set (optimized) limits and have become an integral part of most manufacturing processes. The LabVIEW SPC toolkit consists of a comprehensive set of VI's that come with various data analysis/presentation techniques such as control charts, Pareto analysis, histograms and other statistical calculations. Special VI's compute process statistics and prepare data for display, and they can either accept historical data from a disk or maintain recent data in memory by storing it in a circular buffer.

While SPC is an important tool for the qualitative monitoring of various process parameters, there is also a need to quantitatively determine the components of a reaction/process stream before, during, and after the process has been completed. Near-infrared (NIR) spectroscopy coupled with fiber-optic technology is an analytical technique that is now routinely employed as a process analytical tool for on-line measurements because of its longer path length capabilities.[5] This technique is used in conjunction with multivariate calibration techniques to qualitatively and quantitatively model the desired component(s) in the presence of several interferences that are present in a typical reaction/process stream.

In combining the benefits of LabVIEW, NIR spectroscopy, and SPC, Hailey et al.[6,7] have described an automated system for the on-line monitoring of powder blending processes in a pharmaceutical plant. The system employs NIR spectroscopy using fiber-optics and a graphical-user interface (GUI) developed in the LabVIEW environment. The complete supervisory control and data analysis software (LabVIEW-based) controls blender and spectrophotometer operation and performs statistical spectral data analysis in real time. A data analysis routine using standard deviation was described to demonstrate an approach to the real-time determination of blend homogeneity. The second part of the application involved the extraction of the information content of NIR data using various chemometric approaches.[7]

10.1021/ci000447r CCC: $19.00
Published on Web 07/06/2000

As is typical of such applications, data acquistion and chemometric analysis were accomplished by different software packages.[6,7] In this case, LabVIEW was used for data acquisition and some statistical quality control, two different chemometric packages were used for the various statistical analyses, while a third program was used in converting the raw spectral files into a format suitable for both chemometric programs. If these different data acquisition and chemometric functions could be combined with a single software package, a marked reduction in process analysis time would result.

So-called "real-time chemometrics" as envisioned by the union of LabVIEW and chemometric analysis can provide substantial benefits to manufacturing concerns that require process control. While commercial chemometric toolkits, such as Charm Works 99 (Process Analysis & Automation Ltd., Hampshire, UK), are available for the LabVIEW environment, their extra expense (∼$2245) may be a consideration that could dissuade some potential users.

By employing the advanced analysis toolkit in LabVIEW, it is possible to write graphical-user interfaces that will not only control the instrument but acquire data and perform complex chemometric computations in a seamless manner without the use of other software packages. Because of the widespread popularity of LabVIEW and its multi-platform capability (Windows, MacOS, Unix, and more recently Linnux), the combination of chemometric analysis and real-time data acquisition using LabVIEW is of particular interest.

This paper describes the use of LabVIEW to perform common chemometric computations. In particular, an original LabVIEW program that implements Principal Component Analysis (PCA) was developed and will be described. In addition, two routines that implement Principal Component Regression (PCR) and Partial Least-Squares Regression (PLS) were also written. These routines will be used to predict six commonly determined properties of diesel fuel from NIR spectral data.

EXPERIMENTAL SECTION

The NIR spectral data sets were supplied by the Southwest Research Institute (SWRI), San Antonio, TX through Eigenvector Research, Inc. (Manson, Washington).[8] The data sets were used to determine the following physical properties of diesel fuel: boiling point at 50% recovery (°C), cetane number, density at 15 °C (g mL^{-1}), freezing temperature (°C), total aromatics (mass %) and viscosity (cSt). First derivative spectral data at 401 wavelengths in the NIR region were supplied. The spectral data files were provided in the form of 401 variables rather than in terms of actual wavelengths in nanometers. The lack of actual wavelength information in no way invalidates the use of these files for chemometric evaluation, however, since the files are all self-consistent with respect to the variable index. For each property, six data files were provided—three containing the spectral data and three matching ones for the given property value. One of the three spectral data sets consisted of high-leverage samples that spanned the range of values expected for future samples with respect to the given fuel parameter and was used to build the calibration algorithm for the given fuel parameter (along with the corresponding data set of property values). The calibration algorithm developed with the high-leverage data set was subsequently validated with the two remaining spectral data sets. In all cases, the data were thoroughly vetted, outliers were removed, and all samples were taken from the same class of fuels (all summer fuels, no winter fuels). The six properties for the same diesel fuel samples described above were independently determined by American Society of Testing and Materials (ASTM) standard procedures.[9−13] Prior to importing the data into LabVIEW, the data were converted from the MATLAB file format into text files. The LabVIEW code was written in version 5.1.

RESULTS AND DISCUSSION

When chemical and physical properties of refinery products are determined by traditional (standardized) methods, it is often necessary to employ a separate method for each property. This approach is inconvenient because standard methods are usually time-consuming, expensive, require large amounts of sample, and are not suited for rapid online measurements. Recent advances in process instrumentation (using NIR spectroscopy, in particular), data collection, and multivariate calibration have led to the use of NIR spectroscopy in the measurement of crude oil refinery products and certain qualitative indices of the refinery process.[14]

For oil fractions and diesel fuels, the NIR spectroscopic region (750−2500 nm) is especially attractive because most of the absorption bands observed in this region arise from overtones and combination bands of carbon−hydrogen (C−H) stretching vibrations of the hydrocarbon molecules. The sample spectra typically contain qualitative information (chemical and physical properties) about the whole molecule that is condensed through the complex overlapping of combination and overtone bands. The information can thus only be extracted through multivariate calibration techniques. This approach not only is faster than the conventional standardized techniques but also permits the simultaneous determination several properties from a single spectrum. Recent applications of NIR on-line analysis in the petrochemical industry include blending optimization,[15,16] steam cracker optimization, and crude oil distillation process control.[17]

DIESEL FUEL PARAMETERS

In this study, six diesel fuel parameters were determined from near-infrared spectral data with chemometric routines written in LabVIEW. The necessary NIR spectra and fuel parameters used to develop the models were obtained from data available from the Southwest Research Institute.[8] The significance of the different fuel parameters and the ASTM reference method used to determine them are described below.

Cetane Number. The cetane number provides a measure of the ignition characteristics of diesel fuel oil in compression ignition engines and is also used by engine manufacturers, petroleum refiners and marketers, and in commerce, as a primary specification measurement related to matching fuels and engines. The cetane number of a diesel fuel oil is determined by comparing its combustion characteristics in a test engine with those of blends of reference fuels of known

cetane number under standard operating conditions. Cetane (n-hexadecane), $C_{16}H_{34}$, is defined as having a cetane number of 100. 2,2,4,4,6,8,8-Heptamethylnonane (HMN), $C_{16}H_{34}$, which can be produced in high purity, is used as a low-value reference fuel with a cetane number of 15. Blends of cetane and HMN represent intermediate ignition qualities according to the formula:

$$\text{cetane number} = \%\ \text{cetane} + 0.15\ (\%\ \text{HMN}) \quad (1)$$

The ASTM test D613[9] determines the rating of diesel fuel in terms of an arbitrary scale of cetane numbers using a single cylinder, four-stroke cycle, variable compression ratio, indirect diesel engine. The cetane number scale covers the range from 0 to 100, but typical testing is in the range of 30 to 65 cetane number.

Total Aromatics. The ASTM method D5186-91[10] covers the determination of the total aromatic compounds in diesel motor fuels by supercritical fluid chromatography (SFC). The range of aromatics concentration to which this test method is applicable is from 5 to 75 mass %. Average aromatic levels in the United States are about 30%. In addition to having poor fuel quality, aromatics also contribute to exhaust emissions. The federal government began effectively limiting aromatic content to below 40% starting in October 1993 by specifying a minimum cetane index of 40.[18]

To determine total aromatics, a small aliquot of the fuel sample is injected into a packed silica adsorption column and eluted using supercritical carbon dioxide as the mobile phase. Aromatics in the sample are separated from nonaromatics and detected using a flame ionization detector. The chromatographic areas corresponding to the aromatic and nonaromatic components are determined, and the mass percent aromatic content of the fuel is determined by area normalization.

Viscosity. The ASTM method D445-94 specifies a procedure for the determination of the kinematic viscosity, ν, of liquid petroleum products, both transparent and opaque, by measuring the time for a volume of liquid to flow under gravity through a calibrated glass capillary viscometer.[11] Diesel kinematic viscosity is reported in units of $mm^2\ s^{-1}$. The desired viscosity is a function of fuel grade and ranges from a minimum of 1.3 $mm^2\ s^{-1}$ for 1-D to a maximum of 24 $mm^2\ s^{-1}$ for 4-D. In this study, the kinematic viscosity was reported in centiStokes (cSt), where 1 cSt = 10 $mm^2\ s^{-1}$. The dynamic viscosity, η, can be obtained by multiplying the measured kinematic viscosity by the density, ρ, of the liquid. Many petroleum products, and some nonpetroleum materials, are used as lubricants, and the correct operation of the equipment depends on the appropriate viscosity of the liquid being used. In addition, the viscosity of many petroleum fuels is important for the estimation of optimum storage, handling, and operational conditions. Thus the accurate measurement of viscosity is essential to many product specifications.

Density. Density is a fundamental physical property that can be used in conjunction with other properties to characterize both the light and heavy fractions of petroleum and petroleum products. Also, determination of density (or relative density of petroleum and its products) is necessary for the conversion of measured volumes to the volumes at the standard temperature of 15 °C.

The ASTM method D4052-95 covers the determination of the density, or relative density, of petroleum distillates and viscous oils that can be handled in a normal fashion as liquids at test temperatures between 15 and 35 °C.[12] A small volume (approximately 0.7 mL) of liquid sample is introduced into an oscillating sample tube and the change in oscillating frequency caused by the change in the mass of the tube is used in conjunction with the calibration data to determine the density of the sample.

Boiling-Point-at-50%-Recovery. The ASTM method D86-95 covers the distillation of diesel fuel and other petroleum products, utilizing either manual or automated equipment.[13] A 100-mL sample is distilled under prescribed conditions. Systematic observations of thermometer readings and volumes of condensate are made, and from these data, the boiling-point-at-50%-recovery of the condensate is determined. The distillation (volatility) characteristics of hydrocarbons often have an important effect on safety and performance, especially in the case of fuels and solvents. Volatility is the major determinant in the tendency of a hydrocarbon to produce potentially explosive vapors. It is also critically important for both automotive and aviation gasolines, affecting starting, warm-up, and the tendency to vapor lock at high operating temperatures or at high altitudes, or both. Also, the presence of high boiling point components in fuels can significantly affect the degree of formation of solid combustion products.

Freezing Point. Diesel fuel must be able to be pumped and flow through all filters and injectors at the lowest temperature that may be encountered in use. Despite this, there is no formal ASTM test for determining the freezing temperature of diesel fuel.

CHEMOMETRICS

Linear algebra is the language of chemometrics, and most chemometric software relies heavily on its use. The linear algebra toolkit, which the LabVIEW programming environment provides, comes with many of the sub-VIs that contain most of the basic routines needed for complex matrix computations. These can be used in conjunction with the extensive array tools (1D and 2D) also provided by LabVIEW in writing almost any chemometric routine. MATLAB (acronym for matrix laboratory) is the programming software that is routinely used by chemometricians for this purpose. MATLAB software can be leveraged by transferring existing chemometric routines written in MATLAB code into a LabVIEW measurement application.[19] The computational abilities of LabVIEW were tested with the following routines: PCA, PCR and PLS. Figure 1 shows a sampling of the NIR first-derivative spectra that were used in this study. In this discussion scalars will be denoted by italic lowercase letters, vectors by bold lowercase letters and matrices by bold uppercase letters. Vectors will be column vectors, and row vectors will be represented as a column vector transposed (using a superscript T).

Principal Component Analysis. The PCA routine is one that is central to most applications in chemometrics and has been thoroughly described in the literature.[20] The LabVIEW program for PCA was built around the NIPALS (Non-Iterative Partial Least Squares) algorithm[21] as shown in Figure 2. Here, the first principal component is calculated

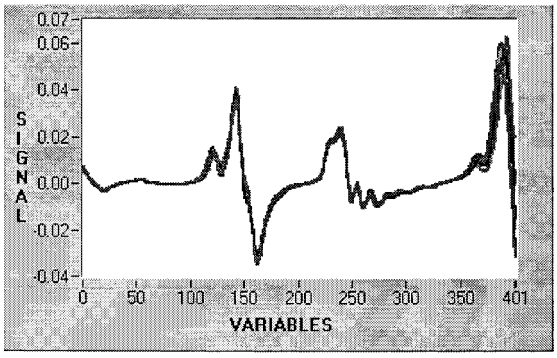

Figure 1. NIR spectra of selected diesel fuel samples.

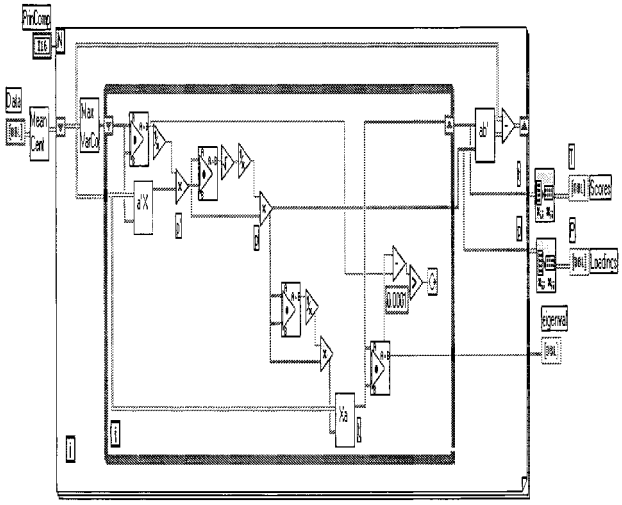

Figure 2. Non-Iterative Partial Least Squares algorithm (NIPALS) VI.

from the pre-scaled data matrix (**X**). This is done by first extracting each score (**t**) and loading ($\mathbf{p}^T$) vector from the matrix, calculating the outer product ($\mathbf{tp}^T$), and subtracting the outer product from **X** to give a residual matrix **E**. This is repeated for **E** until the final residual matrix is close to zero. Generally, PCA decomposes **X** into

$$\mathbf{X} = \mathbf{T}_A\mathbf{P}_A{}^T + \mathbf{E} \tag{2}$$

where *A* is the optimal number of principal components, **T** is the scores matrix, and **P** is the loadings matrix.

The NIPALS algorithm is an iterative procedure that computes $\mathbf{t}_a$ and $\mathbf{p}_a$ from $\mathbf{X}_{a-1}$ for factors $a = 1, 2, ..., A$. Each iteration in the algorithm is completed as follows: select start values, e.g., $\mathbf{t}_a$ = the column in $\mathbf{X}_{a-1}$ that has the highest sum of squares.

(*i*) Improve the estimate of loading vector $\mathbf{p}_a$ for this factor by projecting the matrix $\mathbf{X}_{a-1}$ on $\mathbf{t}_a$, i.e.,

$$\mathbf{p}_a{}^T = (\mathbf{t}_a{}^T\mathbf{t}_a)^{-1}\mathbf{t}_a{}^T\mathbf{X}_{a-1} \tag{3}$$

(*ii*) Scale the length of $\mathbf{p}_a$ to 1.0:

$$\mathbf{p}_a = \mathbf{p}_a(\mathbf{p}_a{}^T\mathbf{p}_a)^{-0.5} \tag{4}$$

(*iii*) Improve the estimate of score $\mathbf{t}_a$ for this factor by projecting the matrix $\mathbf{X}_{a-1}$ on $\mathbf{p}_a$:

$$\mathbf{t}_a{}^T = \mathbf{X}_{a-1}\,\mathbf{p}_a(\mathbf{p}_a{}^T\mathbf{p}_a)^{-1} \tag{5}$$

(*iv*) Improve the estimate of the eigenvalue τ_a:

$$\tau_a = \mathbf{t}_a{}^T\mathbf{t}_a \tag{6}$$

(*v*) Check convergence: If τ_a minus τ_a in the previous iteration is smaller than a certain small pre-specified constant, e.g., 0.0001 times τ_a, the method has converged for this factor. If not, go to step *i*.

Subtract the effect for this factor

$$\mathbf{X}_a = \mathbf{X}_{a-1} - \mathbf{t}_a\mathbf{p}_a{}^T \tag{7}$$

and repeat steps *i*-*v* for the next factor.

Each successive principal component (PC) describes less of the variation in the data so that the entire data matrix can be approximated by the first few principal components or factors. The maximum number of PCs is the lesser of the number of rows (samples) or columns (variables) of the data matrix. Figure 3 shows the PCA demo interface written in LabVIEW. Principal Components Analysis was performed on 118 NIR spectra (giving rise to a 118 × 401 data matrix). As is shown by the variance plot, the first two PCs describe 96% of the variation in the whole data set. A scores plot of PC2 versus PC1 is also shown. Here most of the variation in the spectra of 118 samples is reduced to a 2-dimensional space, so that sample similarities can be observed. LabVIEW's programmable front panel attributes have been used to create a cursor that allows the user to identify each sample on the graph. As shown in the figure, it appears that sample number 10 might be slightly different from the other samples (i.e., an outlier). The loadings plot shows the loadings for the first 4 principal components. Because of their high loadings values, it can be concluded that the contribution of wavelengths 370−401 to the variation described by all 4 PCs is significant. This suggests that those variables (wavelengths) might be particularly useful in predicting the value associated property of the spectra (e.g. cetane number).

Principal Component Regression and Partial Least Squares. These are the most common multivariate calibration techniques in the literature.[22] Both of these techniques are similar in that they assume the following linear inverse model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \tag{8}$$

where **y** is the vector of property variables, **X** is the $n \times m$ (n = samples, m = variables) instrument response matrix (NIR spectra), **b** is the vector of regression coefficients and **e** is the vector of property value residuals. The major difference between the two techniques lies in the manner in which **b** is estimated.

The principal component regression (PCR) algorithm utilizes the Singular Value Decomposition (SVD) function,[23] which is a part of the LabVIEW linear algebra toolkit and is used to decompose centered **X** into 3 matrices:

$$\mathbf{X} = \mathbf{USV}^T \tag{9}$$

In eq 9, **U** and **V** are the scores and loading matrices, respectively, while **S** is a diagonal matrix in which the diagonal elements (eigenvalues) are related to the amount of variance each PC describes. It should be noted that the scores and loadings matrices calculated by both the NIPALS and SVD algorithms are equivalent. The pseudoinverse ($\mathbf{X}^+$)

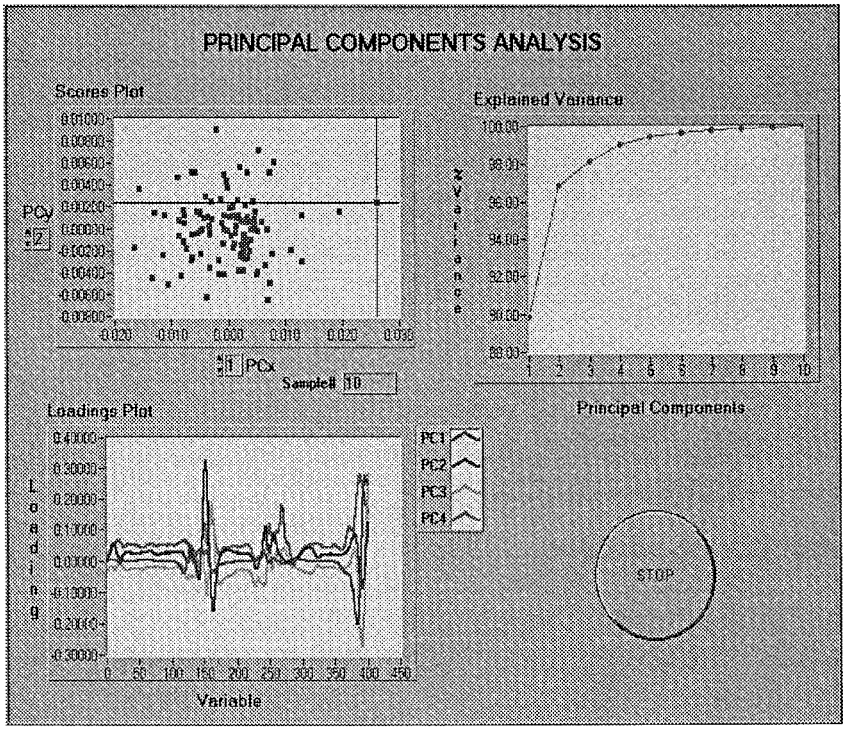

Figure 3. Principal Components Analysis (PCA) demo interface.

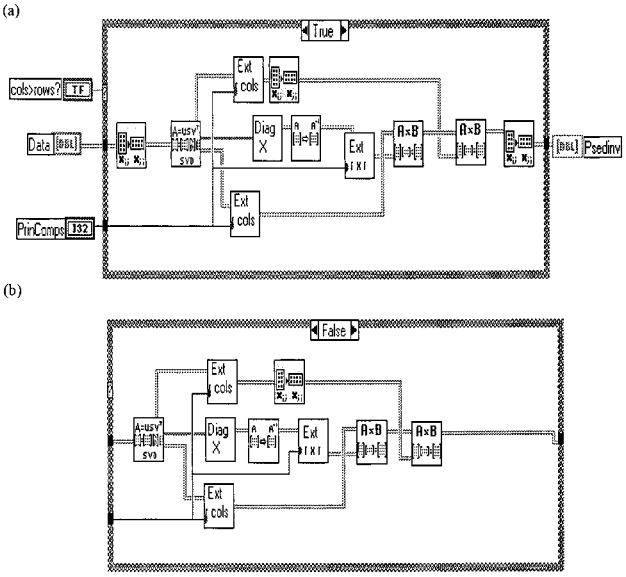

Figure 4. Pseudoinverse algorithm VI. (a) Computes the pseudoinverse when $n < m$. (b) Computes the pseudoinverse when $n > m$.

for an optimal number of factors is then calculated by means of the following equation:

$$\mathbf{X}^{+} = \mathbf{V}\mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}} \tag{10}$$

For a response matrix **X** in which $n < m$ (as is prevalent in spectroscopic techniques), and having r optimum factors, the LabVIEW algorithm described here computes $\mathbf{X}^{+}$ in the following sequence (Figure 4):

(*i*) Transpose **X**.

(*ii*) Determine **U**, **S** and **V** using the SVD function.

(*iii*) For r factors, extract r columns from **U**, r rows from **V** and an $r \times r$ matrix from **S** to give the following new matrices, **U***, **V*** and **S***.

(*iv*) Calculate $(\mathbf{X}^{+})^{\mathrm{T}}$ from the truncated matrices:

$$(\mathbf{X}^{+})^{\mathrm{T}} = \mathbf{V}^{*}\mathbf{S}^{*-1}\mathbf{U}^{*\mathrm{T}} \tag{11}$$

(*v*) Transpose $(\mathbf{X}^{+})^{\mathrm{T}}$.

The steps above are a summary of the LabVIEW PCR algorithm developed in this study. The full LabVIEW PCR program is available (including all the relevant subVIs) as Supporting Information. Calculating $\mathbf{X}^{+}$ from the truncated matrices generates a low-dimension approximation of the data in which relevant information is retained and noise is filtered out. The regression vector is then estimated as

$$\mathbf{b} = \mathbf{X}^{+}\,\mathbf{y} \tag{12}$$

and can be used to predict future samples.

Whereas PCR attempts to find factors that capture the greatest amount of variance in **X**, PLS attempts to find factors that will not only capture the greatest amount of variance in **X** but will also best correlate **X** to **y**. In other words, PLS attempts to maximize the covariance between **X** and **y**. The scores and loadings calculated in PLS are not the same as those calculated in PCA and PCR. They can be thought of, however, as PCA scores and loadings that have been rotated to be more relevant in predicting **y**. As shown in Figure 5, the LabVIEW PLS routine described here calculates the factor loadings for the **X**-variables (**P**), the factor loadings for the **y**-variable (**q**), and the loading weights for the **X**-variables (**W**). The entire LabVIEW PLS regression and prediction routine is described by the orthogonalized PLS algorithm for one **y** variable or PLS,[21] and is given by the following steps:

(*i*) Mean center the scaled input variables **X** and **y**. Choose A_{max} to be higher than the number of phenomena expected in **X**. For each factor $a = 1, ..., A_{max}$ perform steps *iia* to *iie*.

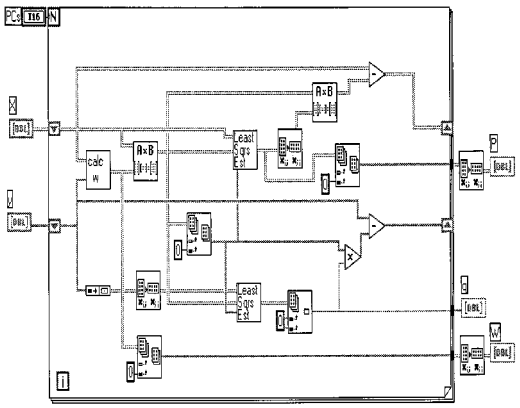

Figure 5. Orthogonalized PLS algorithm for one **y**-variable (PLS1) VI.

(*iia*) Determine the least-squares estimate of the loading weights $\mathbf{w}_a$, from the following "local model"

$$\mathbf{X}_{a-1} = \mathbf{y}_{a-1}\mathbf{w}_a^{T} + \mathbf{E} \tag{13}$$

The least squares solution is

$$\mathbf{w}_a = c\mathbf{X}_{a-1}^{T}\mathbf{y}_{a-1} \tag{14}$$

where c is the scaling factor that makes the length of $\mathbf{w}_a$ equal to 1, i.e.

$$c = (\mathbf{y}_{a-1}^{T}\mathbf{X}_{a-1}\mathbf{X}_{a-1}^{T}\mathbf{y}_{a-1})^{-0.5} \tag{15}$$

(*iib*) Estimate the scores $\mathbf{t}_a$ using the following "local model"

$$\mathbf{X}_{a-1} = \mathbf{t}_a\mathbf{w}_a^{T} + \mathbf{E} \tag{16}$$

The least-squares solution to eq 16 is (since $\mathbf{w}_a^{T}\mathbf{w}_a = 1$)

$$\mathbf{t}_a = \mathbf{X}_{a-1}\mathbf{w}_a \tag{17}$$

(*iic*) Estimate the spectral loadings $\mathbf{p}_a$ using the following "local model"

$$\mathbf{X}_{a-1} = \mathbf{t}_a\mathbf{p}_a^{T} + \mathbf{E} \tag{18}$$

The least squares solution to eq 18 is

$$\mathbf{p}_a = \mathbf{X}_{a-1}^{T}\mathbf{t}_a/\mathbf{t}_a^{T}\mathbf{t}_a \tag{19}$$

(*iid*) Estimate the chemical loading $\mathbf{q}_a$ using the "local model"

$$\mathbf{y}_{a-1} = \mathbf{t}_a\mathbf{q}_a + \mathbf{f} \tag{20}$$

where **f** is the vector of residuals for the local model. The least-squares solution is

$$\mathbf{q}_a = \mathbf{y}_{a-1}^{T}\mathbf{t}_a/\mathbf{t}_a^{T}\mathbf{t}_a \tag{21}$$

(*iie*) Create new **X** and **y** residuals by subtracting the estimated effect of this factor:

$$\mathbf{E} = \mathbf{X}_{a-1} - \mathbf{t}_a\mathbf{p}_a^{T} \tag{22}$$

$$\mathbf{F} = \mathbf{y}_{a-1} - \mathbf{t}_a\mathbf{q}_a \tag{23}$$

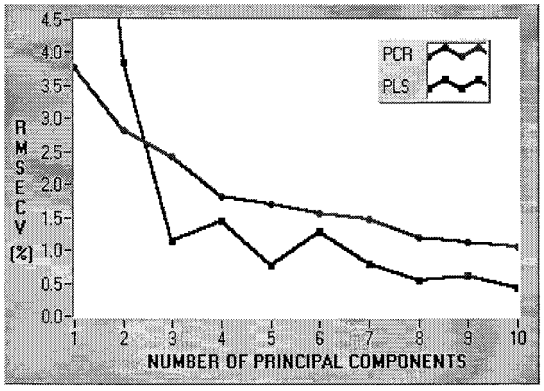

Figure 6. Plot of Root Mean Square Error of Cross Validation (RMDECV) versus the number of factors for PCR and PLS models of cetane number.

Replace the former $\mathbf{X}_{a-1}$ and $\mathbf{y}_{a-1}$ by the new residuals **E** and **f** and increase a by 1, i.e., set

$$\mathbf{X}_a = \mathbf{E} \tag{24}$$

$$\mathbf{y}_a = \mathbf{f} \tag{25}$$

$$a = a + 1 \tag{26}$$

(*iii*) Determine A, the number of valid PLS factors to retain in the calibration model.

(*iv*) Compute b_o and **b** for the A PLS factors for use in the prediction of the unknown y variables. **b** is the regression coefficient vector and b_o is the intercept.

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^{T}\mathbf{W})^{-1}\mathbf{q} \tag{27}$$

$$b_o = y_{av} - \mathbf{x}_{av}^{T}\mathbf{b} \tag{28}$$

where $\mathbf{x}_{av}$ is a vector with elements that are the average intensity values at each wavelength for all the spectra that make up the calibration set. y_{av} is the average of all the y (property) values in the calibration set. The prediction step is carried out as follows

$$y_i = b_o + \mathbf{x}_i^{T}\mathbf{b} \tag{29}$$

where y_i is the value of unknown i in the test or validation set.

For both PCR and PLS, selection of the optimum number of factors is essential to the calibration process to ensure that an adequate model is built. To this end, LabVIEW routines that implement the "leave-one-out" cross-validation procedure[24] were applied to the PCR and PLS models built from the diesel fuel NIR spectra. To illustrate, Figure 6 shows plots of the Root Mean Square Error of Cross Validation (RMSECV) versus the number of factors for the PCR and PLS models for diesel fuel cetane number obtained from NIR spectra. The optimum number of factors is chosen at the first curve minimum. As is expected, the PLS model has a lower optimum number of factors (3), compared with the PCR model (4 factors).

Figure 7 shows the front panel of the PLS regression VI. The PLS regression VI (see Supporting Information) requires the input of a calibration and validation data set. Both kinds of data set consist of a NIR spectral matrix and property value vector. From the user specified number of principal

Table 1. Results of the Multivariate Calibration of Six Diesel Fuel Properties from NIR Spectral Data Using PCR and PLS Algorithms Written in LabVIEW

	cetane number		boiling point at 50% recovery		total aromatics		viscosity		freezing point		density	
	PCR	PLS	PCR	PLS	PCR	PLS	PCR	PLS	PCR	PLS	PCR	PLS
factors[a]	4	3	4	3	2	2	4	2	3	3	2	2
R^2	0.991	0.985	0.975	0.991	0.994	0.989	0.976	0.979	0.936	0.984	0.992	0.985
RMSEP[b]	0.70	1.23	4.90	5.34	0.89	3.91	0.19	0.23	2.91	2.85	0.002	0.006

[a] Optimum number of factors. [b] Root-mean-square error of prediction.

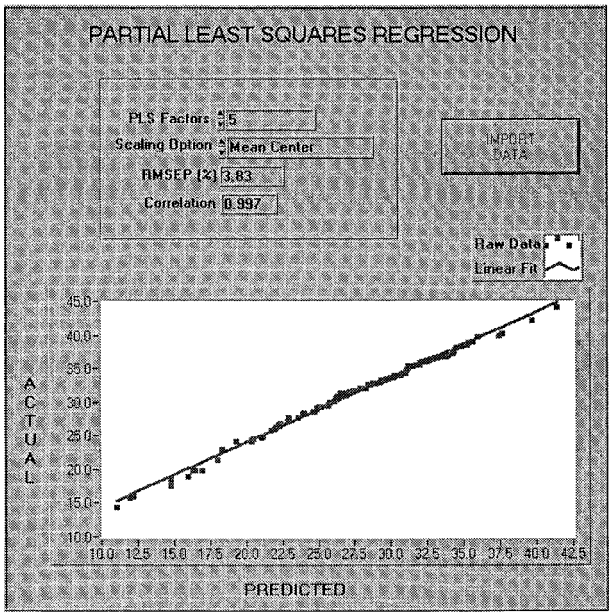

Figure 7. User interface for the PLS regression program.

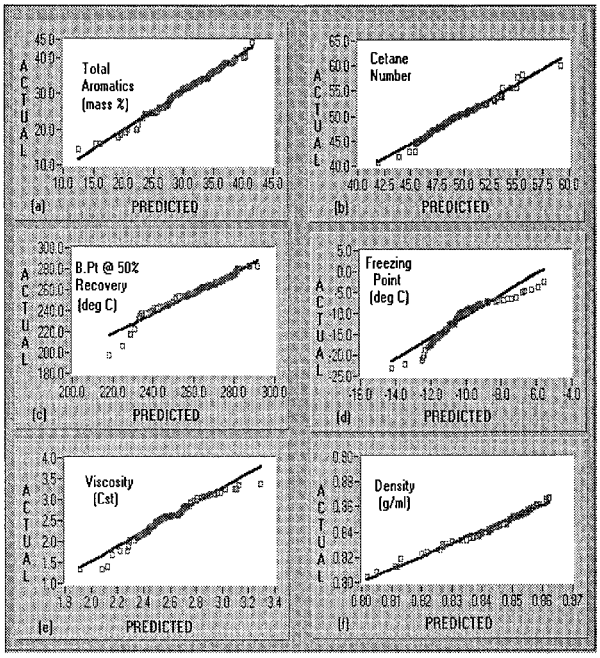

Figure 8. Plots of predicted versus actual values for the six diesel fuel properties. (a) Total aromatics; (b) Cetane number; (c) Boiling point at 50% recovery; (d) Freezing point; (e) Viscosity; (f) Density.

components, a regression model is built from the calibration set and then used to predict the property values of the validation set from the corresponding NIR spectral data. The predicted values and the actual values are compared on a plot as shown in the diagram. The program also determines the correlation coefficient and calculates the Root Mean Square Error of Prediction (*RMSEP*)

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{N}(y_i^p - y_i)^2}{N}} \qquad (30)$$

where N is the number of validation samples, y_i is the certified value for the i^{th} prediction sample and y_i^p is the predicted value for this sample. Figure 8a−f shows the actual values and the predicted values for all six properties. Excellent correlation was obtained between the actual and the predicted values for all six properties.

A summary of the results of the PCR and PLS calibration of the NIR data is shown in Table 1. The results show that PCR yielded lower prediction errors with *RMSEP* values from 0.002 to 4.9%, compared to 0.006−5.34% for the PLS calibration. Plots of the predicted versus actual values for all the diesel fuel properties gave correlation coefficients better than 0.94 for both calibration techniques.

CONCLUSIONS

Original chemometric routines written in the LabVIEW programming environment have been described. The performance of these routines with NIR data of diesel fuel samples was tested and evaluated with excellent results. These LabVIEW programs are flexible in two ways. First of all, they are fully functional as "stand-alone" executables with user-friendly interactive interfaces. Previously generated data can thus be imported into the programs for the purpose of building calibration models. Once the model has been built and the model parameters defined, the program can also operate at the level of a sub-VI in which it becomes a module of the LabVIEW-based instrument control and data acquisition program. The multivariate calibration routines can then be used to process recently acquired data in "real-time". Data from the instrument can be maintained in memory (perhaps in a circular buffer), prior to analysis and display.

Finally, it is important to note that the routines described above only demonstrate the utility of LabVIEW as a chemometric analysis tool. The full range of its capabilities in this field has yet to be explored and indeed the possibilities are endless.

ACKNOWLEDGMENT

The authors would like to acknowledge the Southwest Research Institute of San Antonio, Texas, and Eigenvector Research, Inc. (Manson, Washington) for the data sets used in this study. Support by the Baylor University Research Committee is also acknowledged (002-A98-URC). O. Soyemi would like to acknowledge the Department of Chemistry & Biochemistry, the Biology Department, and the Graduate School for travel funds to present this work at the 26th Federation of Analytical Chemistry & Spectroscopy Societies meeting in Vancouver, British Columbia, Canada, October, 1999.

Supporting Information Available: The VIs and subVIs used for the various multivariate analysis routines described in this paper are presented. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kirkman, I. W.; Buksh, P. A. Data Acquisition and Control using National Instruments' "LabVIEW" software. *Rev. Sci. Instr.* **1992**, *63*, 869−872.

(2) Moore, J. A. Artificial Intelligence Programming with LabVIEW: Genetic Algorithms for Instrumentation, Control, and Optimization. *Comput. Methods Prog. Biomed.* **1995**, *47*, 73−79.

(3) Starn, T. K.; Sesi, N. N.; Horner, J. A.; Hieftje, G. M. A LabVIEW Program for Determining Electron Number Density for Stark Broadening Measurements of the Hydrogen Beta-Line. *Spectrochim. Acta Part B* **1995**, *50*, 1147−1158.

(4) Johnson, G. *LabVIEW Graphical Programming: Practical Applications to Instrumentation and Control*; McGraw-Hill: New York, 1997.

(5) Hassell, C. D.; Bowman, E. M. Process Analytical Chemistry for Spectroscopists. *Appl. Spectrosc.* **1998**, *52*, 18A-29A.

(6) Hailey, P. A.; Doherty, P.; Tapsell, P.; Oliver, T.; Aldridge, P. K. Automated System for the Online Monitoring of Powder Blending Processes using Near Infrared Spectroscopy. Part I. System Development and Control. *J. Pharm. Biomed. Anal.* **1996**, *14*, 551−559.

(7) Sekule, S. S.; Wakeman, J.; Doherty, P.; Hailey, P. A. Automated System for the Online Monitoring of Powder Blending Processes using Near Infrared Spectroscopy. Part II. Qualitative Approaches to Blend Optimization. *J. Pharm. Biomed. Anal.* **1998**, *17*, 1285−1309.

(8) Data sets for the diesel fuels were supplied by Southwest Research Institute, San Antonio, TX, and the U.S. Army through the web site of Eigenvector Research, Inc., Manson, Washington (http://www.eigenvector.com/Data/SWRI).

(9) Standard Test Method for Cetane Number of Diesel Fuel Oil (D613-95). In *Annual Book of ASTM Standards*; Storer, P. A., Ed.; ASTM Press: West Conshohocken, PA, 1996; V. 5.04, pp 137−164.

(10) Standard Test Method for Determination of Aromatic Content in Diesel Fuels by Supercritical Fluid Chromatography (D5186-91). In *Annual Book of ASTM Standards*; Storer, P. A., Ed.; ASTM Press: West Conshohocken, PA, 1996; V. 5.03, pp 338−341.

(11) Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (the Calculation of Dynamic Viscosity) (D445-94). In *Annual Book of ASTM Standards*; Storer, P. A., Ed.; ASTM Press: West Conshohocken, PA, 1996; V. 5.01, pp 162−168.

(12) Standard Test Method for Density and Relative Density of Liquids by Digital Density Meter (D4052-92). In *Annual Book of ASTM Standards*; Storer, P. A., Ed.; ASTM Press: West Conshohocken, PA, 1996; V. 5.02, pp 688−691.

(13) Standard Test Method for Distillation of Petroleum Products (D86-95). In *Annual Book of ASTM Standards*; Storer, P. A., Ed.; ASTM Press: West Conshohocken, PA, 1996; V. 5.01, pp 10−22.

(14) Sikora, Z.; Salacki, W. Use of Near-Infrared (NIR) Spectroscopy to Predict Several Physical and Operating Properties of Oil Fractions and Diesel Fuels. *Pet. Coal* **1991**, *38*, 65−68.

(15) Zhaohui, C.; Xinlu. F. Using NIR Spectroscopy for Online Gasoline Analysis. *Hydrocarbon Processing* **1992**, *71*, 94−96.

(16) Zelter, M. S.; Politzer, B. A. Online Octane Control with NIR Analyzers. *Hydrocarbon Processing* **1993**, *72*, 103−106.

(17) Espinoza, A.; Martens, A. Online NIR Analysis and Advanced Control Improve Gasoline Blending. *Oil Gas J.* **1994**, October, 49−56.

(18) Hochhauser, A. M. Gasoline and other Motor Fuels. In *Encyclopedia of Chemical Technology*, 4th ed.; Kroschwitz, J. I., Howe-Grant, M., Eds.; Wiley: New York, 1994; V. 12, pp 341−388.

(19) Dorst, N. LabVIEW 5.1. The Newest Version of National Instruments' Graphical Programming Environment Offers an Array of Internet Tools and Integrates the Power of MATLAB. *Sensors* **1999**, February, 45−48.

(20) Cowe, I. A.; McNichol, J. W. The Use of Principal Components in the Analysis of Near Infrared Spectra. *Appl. Spectrosc.* **1985**, *39*, 257−266.

(21) Martens, H.; Næs, T. *Multivariate Calibration*; Wiley: New York, 1991.

(22) Seasholtz, M. B.; Kowalski, B. R. Recent Developments in Multivariate Calibration. *J. Chemometrics* **1991**, *5*, 129−145.

(23) Golub, G. H.; van Loan, C. F. *Matrix Computations*; John Hopkins University Press: Baltimore, 1990.

(24) Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics: A Practical Guide*; Wiley: New York, 1998.

CI000447R