absorbance detection along the axis perpendicular to the capillary, as demonstrated here, coupled with RI gradient detection along the axis parallel to the capillary as described by Pawliszyn (30). Making both of these measurements simultaneously, yet separately, would require a two-dimensional PSD. Two-dimensional PSD's are commercially available, and some interest in their use in laser beam deflection sensing has been reported (31). Future work is in this direction.

ACKNOWLEDGMENT

We thank R. Olund for his electronics expertise in preparing the circuitry associated with the PSD.

LITERATURE CITED

(1) Upor, E.; Mohoi, M.; Novak, G. In *Comprehensive Analytical Chemistry*; Svehla, G., Ed.; Elsevier: New York, 1985; Vol. 20.
(2) Williams, D. H.; Fleming, I. *Spectroscopic Methods in Organic Chemistry*; Urmo: Bilbao, Spain, 1984; 224 pp.
(3) Stevenson, R. L. *Chromatogr. Sci.* **1983**, *23*, 23-86.
(4) Dolan, J. W.; Berry, V. V. *LC Mag.* **1984**, *2*, 290.
(5) DiCesare, J. L.; Ettre, L. S. *J. Chromatogr.* **1982**, *251*, 1-16.
(6) Abbott, S. R.; Tusa, J. *J. Liq. Chromatogr.* **1983**, *6*, 77-104.
(7) Green, R. B. In *Chemical Analysis: Detectors for Liquid Chromatography*; Yeung, E. S., Ed.; Wiley: New York, 1986; Vol. 89.
(8) Leach, R. A.; Harris, J. M. *J. Chromatogr.* **1981**, *218*, 15-19.
(9) Buffet, C. E.; Morris, M. D. *Anal. Chem.* **1982**, *54*, 1824-1825.
(10) Buffet, C. E.; Morris, M. D. *Anal. Chem.* **1983**, *55*, 376-378.
(11) Pang, T.-K. J.; Morris, M. D. *Anal. Chem.* **1985**, *57*, 2153-2155.
(12) Yang, Y.; Hairrell, R. E. *Anal. Chem.* **1984**, *56*, 3002-3004.
(13) Collette, T. W.; Parekh, N. J.; Griffin, J. H.; Carreira, L. A.; Rogers, L. B. *Appl. Spectrosc.* **1986**, *40*, 164-169.
(14) Sepaniak, M. J.; Vargo, J. D.; Kettler, C. N.; Maskarinec, M. P. *Anal. Chem.* **1984**, *56*, 1252-1257.
(15) Leach, R. A.; Harris, J. M. *Anal. Chim. Acta* **1984**, *164*, 91-101.
(16) Harris, J. M.; Dovichi, N. J. *Anal. Chem.* **1980**, *52*, 695A-706A.
(17) Dovichi, N. J.; Nolan, T. G.; Weimer, W. A. *Anal. Chem.* **1984**, *56*, 1700-1704.
(18) Bornhop, D. J.; Dovichi, N. J. *Anal. Chem.* **1987**, *59*, 1632-1636.
(19) Woodruff, S. D.; Yeung, E. S. *Anal. Chem.* **1982**, *54*, 1174-1178.
(20) Bobbitt, D. R.; Yeung, E. S. *Anal. Chem.* **1985**, *57*, 271-274.
(21) Wilson, S. A.; Yeung, E. S. *Anal. Chem.* **1985**, *57*, 2611-2614.
(22) Skogerboe, K. J.; Yeung, E. S. *Anal. Chem.* **1986**, *58*, 1014-1018.
(23) Callis, J. B.; Illman, D. L.; Kowalski, B. R. *Anal. Chem.* **1987**, *59*, 624A-637A.
(24) Ruzicka, J.; Hansen, E. H. H. *Flow Injection Analysis*; Wiley: New York, 1981.
(25) Synovec, R. E. *Anal. Chem.* **1987**, *59*, 2877-2884.
(26) McKone, H. T.; Ivie, K. *J. Chem. Educ.* **1980**, *57*, 321-322.
(27) Hamamatsu, Technical Note TN-102, Jan 1982.
(28) Kaye, W. *Anal. Chem.* **1981**, *53*, 369-374.
(29) Pardue, H. L.; Rodriquez, P. A. *Anal. Chem.* **1967**, *39*, 901-907.
(30) Pawliszyn, J. *Anal. Chem.* **1986**, *58*, 3207-3215.
(31) Pawliszyn, J.; Weber, M. F.; Dignam, M. J. *Rev. Sci. Instrum.* **1985**, *56*, 1740-1743.

RECEIVED for review November 30, 1987. Accepted February 8, 1988.

Partial Least-Squares Methods for Spectral Analyses. 1. Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information

David M. Haaland* and Edward V. Thomas

Sandia National Laboratories, Albuquerque, New Mexico 87185

Partial least-squares (PLS) methods for spectral analyses are related to other multivariate calibration methods such as classical least-squares (CLS), inverse least-squares (ILS), and principal component regression (PCR) methods which have been used often in quantitative spectral analyses. The PLS method which analyzes one chemical component at a time is presented, and the basis for each step in the algorithm is explained. PLS calibration is shown to be composed of a series of simplified CLS and ILS steps. This detailed understanding of the PLS algorithm has helped to identify how chemically interpretable qualitative spectral information can be obtained from the intermediate steps of the PLS algorithm. These methods for extracting qualitative information are demonstrated by use of simulated spectral data. The qualitative information directly available from the PLS analysis is superior to that obtained from PCR but is not as complete as that which can be generated during CLS analyses. Methods are presented for selecting optimal numbers of loading vectors for both the PLS and PCR models in order to optimize the model while simultaneously reducing the potential for overfitting the calibration data. Outlier detection and methods to evaluate the statistical significance of results obtained from the different calibration methods applied to the same spectral data are also discussed.

Partial least-squares (PLS) modeling is a powerful new multivariate statistical tool that has been successfully applied to the quantitative analyses of ultraviolet (1, 2) near-infrared (3-5), chromatographic (6-8), and electrochemical (9) data. An excellent review of this multivariate statistical method has been presented by Martens (10), which also includes a number of published papers. Recently Lorber et al. (11) presented a theoretical basis for the PLS algorithm, and Geladi and Kowalski published a tutorial on the PLS algorithm (12). PLS software has also recently been made available by several Fourier transform infrared (FT-IR) instrument manufacturers for quantitative spectral analyses. Since software using PLS techniques is now available, it is important for infrared spectroscopists to understand the PLS method and its relation to methods more commonly used in quantitative IR spectroscopy. Therefore, PLS will be described along with the classical least-squares (CLS) (13-16), inverse least-squares (ILS) (17-19), and principal component regression (PCR) (20-23) multivariate statistical methods which have been applied to quantitative IR analyses in the past. A detailed description and understanding of the PLS algorithm is presented here which indicates that while it is similar to PCR, the PLS calibration can be broken down into steps that separately involve CLS calibration and prediction followed by ILS calibration. Thus PLS has properties which combine some of the separate advantages of CLS and ILS methods while making some potential improvements over PCR. In addition, it will be shown that this detailed understanding of the PLS algorithm helps us identify how qualitative information might be extracted from the intermediate steps of the PLS modeling. This chemically interpretable spectral information available during the PLS calibration and prediction

0003-2700/88/0360-1193$01.50/0

has not been fully utilized in near-infrared or UV analyses since these spectroscopies do not contain the wealth of chemical information that is available in a mid-infrared spectrum. It is, therefore, the purpose of the first paper in this series to present the theory behind the various multivariate statistical methods, improve the conceptual understanding of the PLS algorithm by identifying its relationship to other calibration methods, and show how this detailed knowledge can improve the extraction of qualitative information from this relatively new method. We will also present procedures for selecting the optimal number of PLS or PCR factors and identify methods to compare the statistical significance of differences in the results obtained from several multivariate calibration techniques applied to the same data. The following paper (*24*) applies the methods developed in this paper, compares PLS and PCR methods applied to simulated spectral data, and evaluates the three full-spectrum methods (CLS, PCR, and PLS) using the infrared spectra of bulk multicomponent glass samples.

THEORY

Relation of PLS to Other Multivariate Methods for Quantitative Spectral Analyses. PLS is capable of being a full-spectrum method and therefore enjoys the signal averaging advantages of other full-spectrum methods such as PCR and CLS (*14*). Because PLS is a full-spectrum method, efficient outlier detection methods are available from spectral residuals, and limited chemically interpretable spectral information can be obtained from PLS in some cases. (Outliers are samples that are not representative of the calibration samples, and therefore, their estimated concentrations must be treated with caution. Spectral residuals are the difference between the measured and estimated spectra.) PLS is one of several factor analysis methods that are available along with PCR and CLS (although CLS is not commonly presented as a factor analysis method). PLS also has characteristics and advantages of the ILS method which is limited in the number of spectral frequencies that can be included in the analysis. Therefore, to understand the advantages of applying PLS, it is useful to consider the more common multivariate calibration methods that have been used for quantitative spectral analyses. These methods have all generally presumed that there is a linear relationship between absorbance and component concentrations. In addition, each method has a calibration step where the relationship between the spectra and component concentrations is estimated from a set of reference samples. This step is followed by prediction in which the results of the calibration are used to predict or estimate the component concentrations from the "unknown" sample spectrum. Most of the recent quantitative infrared studies involving multivariate statistical methods have made use of two basic statistical approaches. These are the classical least-squares (CLS) (*13–16*) and inverse least-squares (ILS) (*17–19*) methods which have often been labeled the **K** and **P** matrix methods, respectively, by infrared spectroscopists. CLS has also been labeled direct (*10*) or total (*11*) calibration while ILS has been referred to as multiple linear regression (MLR) (*10*), indirect (*10*), or partial (*11*) calibration. Both CLS and ILS methods use multiple linear regression techniques, but they exhibit very different properties and each method has its own set of advantages and disadvantages. Several of the advantages and disadvantages of each method have recently been outlined (*25*).

In the following sections, we use notation prevalent in the infrared literature. However, the matrix representing the spectral data is transposed from its normal configuration in ref 13–19 to be consistent with literature describing the PLS and PCR algorithms. Boldface upper case letters are used for matrices, primes for transposed matrices and vectors, boldface lower case characters for vectors, and lower case italic characters for scalars. We also use the convention that all vectors are expressed as column vectors. Row vectors are expressed as transposed column vectors.

A. Classical Least-Squares Methods. The CLS method assumes the Beer's law model with the absorbance at each frequency being proportional to the component concentrations. Model error is presumed to be due to error in spectral absorbances. In matrix notation, the Beer's law model for m calibration standards containing l chemical components with spectra of n digitized absorbances is given by

$$\mathbf{A} = \mathbf{CK} + \mathbf{E}_A \tag{1}$$

where $\mathbf{A}$ is the $m \times n$ matrix of calibration spectra, $\mathbf{C}$ is the $m \times l$ matrix of component concentrations, $\mathbf{K}$ is the $l \times n$ matrix of absorptivity–path length products, and $\mathbf{E}_A$ is the $m \times n$ matrix of spectral errors or residuals not fit by the model. $\mathbf{K}$ then represents the matrix of pure-component spectra at unit concentration and unit path length. The classical least-squares solution to eq 1 during calibration is

$$\hat{\mathbf{K}} = (\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\mathbf{A} \tag{2}$$

where $\hat{\mathbf{K}}$ indicates the least-squares estimate of $\mathbf{K}$ with the sum of squared spectral errors being minimized. During prediction, the least-squares solution for the vector of unknown component concentrations, $\mathbf{c}$, is

$$\hat{\mathbf{c}} = (\hat{\mathbf{K}}\hat{\mathbf{K}}')^{-1}\hat{\mathbf{K}}\mathbf{a} \tag{3}$$

where $\mathbf{a}$ is the spectrum of the unknown sample and $\hat{\mathbf{K}}$ is from eq 2.

Equation 1 shows that CLS can be considered a factor analysis method since the spectral matrix $\mathbf{A}$ is represented as the product of two smaller matrices $\mathbf{C}$ and $\mathbf{K}$. The pure-component spectra (rows of $\mathbf{K}$) are the factor loadings (also called loading vectors) and the chemical concentrations (elements in $\mathbf{C}$) are the factors (or scores). This model changes the representation of the calibration spectra into a new coordinate system with the new coordinates being the l pure-component spectra rather than the n spectral frequencies. Although this coordinate system is not necessarily orthogonal, it has the advantage that the l spectral intensities for each mixture in this new coordinate system of pure-component spectra are the elements of $\mathbf{C}$; i.e., the intensities in the new coordinate system are the component concentrations. This is clear when one considers that the component concentrations represent the amount (or intensities) of the pure-component spectra which make up any given mixture spectrum.

Since CLS is a full-spectrum method, it can (1) provide significant improvements in precision (*14*) over methods that are restricted to a small number of frequencies, (2) allow simultaneous fitting of spectral base lines (*14–16*), and (3) make available for examination and interpretation least-squares estimated pure-component spectra and full-spectrum residuals (*16*, *26*, *27*). A major disadvantage of the CLS method is that all interfering chemical components in the spectral region of interest need to be known and included in the calibration. This restriction can be reduced significantly by performing the analysis one spectral band at a time followed by pooling the results in a statistically efficient manner (*14–16*, *28*). This allows a high degree of rejection of spectral bands which do not follow Beer's law or which include the presence of major interfering components. Nevertheless, we have found cases in which spectral overlap occurs throughout the spectral range, and a knowledge of all components in the sample is essential for accurate quantitative spectral analysis by the CLS method (*29*).

B. Inverse Least-Squares Method. The inverse least-squares method assumes that concentration is a function of absorbance. The inverse Beer's law model for m calibration

standards with spectra of n digitized absorbances is given by

$$\mathbf{C} = \mathbf{AP} + \mathbf{E}_C \quad (4)$$

where $\mathbf{C}$ and $\mathbf{A}$ are as before, $\mathbf{P}$ is the $n \times l$ matrix of the unknown calibration coefficients relating the l component concentrations to the spectral intensities, and $\mathbf{E}_C$ is the $m \times l$ vector of random concentration errors or residuals that are not fit by the model. Since model error is presumed to be error in the component concentrations, this method minimizes the squared errors in concentrations during calibration. The inverse representation of Beer's law has the significant advantage that the analysis based on this model is invariant with respect to the number of chemical components, l, included in the analysis. If it is assumed that the elements in different columns of $\mathbf{E}_c$ are independent, an identical analysis for each individual analyte can be obtained by considering the reduced model for one component

$$\mathbf{c} = \mathbf{Ap} + \mathbf{e}_c \quad (5)$$

Here $\mathbf{c}$ is the $m \times 1$ vector of concentrations of the analyte of interest in the m calibration samples, $\mathbf{p}$ is then $n \times 1$ vector of calibration coefficients, and $\mathbf{e}_c$ is the $m \times 1$ vector of concentration residuals not fit by the model.

During calibration, the least-squares solution for $\mathbf{p}$ in eq 5 is

$$\hat{\mathbf{p}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{c} \quad (6)$$

During prediction, the solution for the analyte concentration in the unknown sample is simply

$$\hat{c} = \mathbf{a}'\hat{\mathbf{p}} \quad (7)$$

This means a quantitative spectral analysis can be performed even if the concentration of only one component is known in the calibration mixtures. The components not included in the analysis must be present and implicitly modeled during calibration. The above capability of the ILS method has resulted in it being used for near-infrared analysis (NIRA) methods (*30*).

One disadvantage of the ILS method is that the analysis generally has to be restricted to a small number of frequencies. This is because the matrix which must be inverted in eq 6 has the dimension equal to the number of frequencies, and this number cannot exceed the number of calibration mixtures used in the analysis. In addition, collinearity problems (i.e., the near linear relationships between absorbances at multiple frequencies) can become significant when the number of frequencies becomes too large, and the precision of the results are actually degraded when too many frequencies are included in the analysis. Therefore, the large improvements in precision and the full-spectrum advantages of CLS methods are not possible with the inverse method. In addition, determining how many and which frequencies to include in the analysis is not a trivial problem for complex samples. Although statistical methods such as stepwise multiple linear regression have been used for the selection of frequencies in NIRA (see ref 31 for a discussion of frequency selection methods), infrared spectroscopists have not yet made use of these frequency selection methods. Suboptimal frequency selection can introduce problems such as poor base-line modeling, noise inflation from the collinearity problem, and overfitting; i.e., the noise and errors in the calibration data are modeled during calibration.

C. PLS and PCR Models. The PLS and PCR algorithms and their general goals are similar, so the common features of these methods will be discussed together before the PLS1 algorithm is described in detail. It should be noted that PCR is simply principal component analysis (PCA) followed by a regression step (*20*). Both PLS and PCR are factor analysis methods which have many of the full-spectrum advantages of the CLS method. However, they retain the ILS advantage of being able to perform the analysis one chemical component at a time while avoiding the ILS frequency selection problems. As in CLS, the full-spectrum capabilities are retained by forming a new coordinate system consisting of full-spectrum basis vectors which comprise one of the smaller matrices (see eq 8). The other matrix in the decomposition corresponds to the intensities in the new full-spectrum coordinate system. By use of the spectral decomposition notation of Lindberg et al. (*1*), the calibration spectra can be represented for either the PCA or PLS model as follows:

$$\mathbf{A} = \mathbf{TB} + \mathbf{E}_A \quad (8)$$

where $\mathbf{B}$ is a $h \times n$ matrix with the rows of $\mathbf{B}$ being the new PLS or PCA basis set of h full-spectrum vectors, often called loading vectors or loading spectra. $\mathbf{T}$ is an $m \times h$ matrix of intensities (or scores) in the new coordinate system of the h PLS or PCA loading vectors for the m sample spectra. In PCA the rows of $\mathbf{B}$ are eigenvectors of $\mathbf{A}'\mathbf{A}$, and the columns of $\mathbf{T}$ are proportional to the eigenvectors of $\mathbf{AA}'$. $\mathbf{E}_A$ is now the $m \times n$ matrix of spectral residuals not fit by the optimal PLS or PCR model. The analogy between eq 8 of PLS or PCA and eq 1 for CLS is quite clear since both equations involve the decomposition of $\mathbf{A}$ into the product of two smaller matrices. However, now rather than the basis vectors being the pure-component spectra, they are the loading vectors generated by the PLS or PCA algorithms. The intensities in the new coordinate system are no longer the concentrations as they were in CLS, but they can be modeled as linearly related to concentrations as shown later. The new basis set of full-spectrum loading vectors is composed of linear combinations of the original calibration spectra. The amounts (i.e., intensities) of each of the loading vectors which are required to reconstruct each calibration spectrum are the scores.

In general, in a noise-free system only a small number of the full-spectrum basis vectors are required to represent the calibration spectra ($\mathbf{A}$). When the rank of $\mathbf{A}$ which is important for concentration prediction is r, then the optimal PLS or PCR model in eq 8 will have the dimension h equal to r. A new method for the selection of the optimal number of loading vectors is described later. In general $r < m$ and $r < n$, in which case PLS and PCA will have reduced the number of intensities (n) of each spectrum in the spectral matrix $\mathbf{A}$ to a small number of intensities (r) in the new coordinate system of the loading vectors. This data compression step also reduces the noise (*32*) since noise is distributed throughout all loading vectors while the true spectral variation is generally concentrated in the early loading vectors.

The spectral intensities ($\mathbf{T}$) in the new coordinate system can be related to concentrations with a separate inverse least-squares analysis using a model similar to eq 5. However, rather than solving eq 5 by least-squares methods with the problem of calculating $(\mathbf{A}'\mathbf{A})^{-1}$, we solve the following set of equations by least squares:

$$\mathbf{c} = \mathbf{Tv} + \mathbf{e}_c \quad (9)$$

where $\mathbf{v}$ is the $h \times 1$ vector of coefficients relating the scores to the concentrations and $\mathbf{T}$ is the matrix of scores (intensities) from the PLS or PCA spectral decomposition in eq 8. The details of how $\mathbf{T}$ and $\mathbf{B}$ are calculated by PLS are given later in this paper. Both PCA and the PLS algorithm to be described here generate intensities (scores) that are orthogonal, and hence, the collinearity problem originally encountered in the inverse least-squares solution to eq 5 is eliminated. Therefore, all the useful spectral information contained in the reduced representation of $\mathbf{A}$ can be included, and the analysis can be performed one chemical component at a time without the hazards of collinearity. The least-squares solution for $\mathbf{v}$ in eq 9 has the same form as the ILS solution given in eq 6 with $\mathbf{T}$ substituted for $\mathbf{A}$ and $\hat{\mathbf{v}}$ in place of $\hat{\mathbf{p}}$. Because the

columns of **T** are orthogonal in both PLS and PCA, the least-squares solution to **v** in eq 9 involves the trivial inversion of the diagonal (**T'T**) matrix. Thus, PLS and PCR both involve an inverse least-squares step (PCR is simply PCA followed by the separate regression step for the model given in eq 9). Since PCA and PLS can both reduce spectral noise in the compressed representation of the spectral **A** matrix, the assumption in the ILS model that the dominant errors are in concentrations rather than spectra is more closely approximated. In PCA the reduction of errors in spectral space is optimal in a least-squares sense, but the resulting loading vectors are not specifically related to any single analyte. In PLS the overall reduction in spectral noise is not necessarily optimal, but the PLS basis vectors are generated to relate to the chemical component of interest. Therefore, PLS and PCR can be considered hybrid methods which combine some of the advantages of both CLS and ILS methods.

Although PCA and PLS are similar, the methods to accomplish the goals of spectral decomposition and concentration prediction are different. Both methods as implemented here involve stepwise algorithms which calculate the **B** and **T** matrices one vector at a time until the desired model has been obtained. In general, different **T** and **B** matrices are generated by the PLS and PCA methods. In PCA, the columns of **T** are orthogonal and the rows of **B** are orthonormal while in the version of PLS presented here only the columns of **T** are orthogonal. The PCA algorithm used here is the NIPALS (nonlinear iterative partial least-squares) algorithm developed by Wold (*33*). NIPALS is an efficient iterative algorithm which extracts the full-spectrum loading vectors (eigenvectors of **A'A**) from the spectra in the order of their contribution to the variance in the calibration spectra. After the first loading vector has been determined, it is removed from each calibration spectrum, and the process is repeated until the desired number of loading vectors has been calculated. The potential problem with PCR is that the loading vectors which best represent the spectral data may not be optimal for concentration prediction. Therefore, it would be desirable to derive loading vectors so that more predictive information is placed in the first factors. The PLS algorithm presented here (*1*) is a modification of the NIPALS algorithm, and it achieves the above goal by using concentration information to obtain the decomposition of the spectral matrix **A** in eq 8. Concentration-dependent loading vectors are generated (**B**) and the computed scores (**T**) are then related to the concentrations or concentration residuals after each loading vector is calculated. Therefore, in principle, greater predictive ability is forced into the early PLS loading vectors.

Basic PLS1 Algorithm. We only discuss in detail the PLS method in which the calibration and prediction analyses are performed one component at a time. That is, only the concentrations of the chemical component of interest are used in the calibration; other concentrations, even if known, are not included in the analysis. This is the PLS1 algorithm also called PLS regression (*10*). The PLS1 calibration and prediction algorithms are summarized in Tables I and II, respectively. Two or more components can be calibrated or analyzed simultaneously by using a global PLS algorithm called PLS2. In analyzing real samples, we have found empirically that PLS1, which is a subset of PLS2, more often exhibits better predictive properties than PLS2. When PLS2 is used, the component concentrations must be normalized such that their error variances are equal (i.e., corrected for the different precisions with which the component concentrations are known) and often these values are not known. In addition, the optimal number of PLS loading vectors is often different for each component, and usually the PLS2 algorithm has been restricted to finding a single optimal number of loading vectors for all chemical components. Finally, much of the chemical information contained in each chemical component or property is more difficult to interpret when applying PLS2 since information from the various chemical components gets intimately mixed. PLS2 is, therefore, probably better suited for classification or pattern recognition applications rather than for prediction of individual species. The PLS1 algorithm discussed below is a corrected form of the algorithm presented by Lindberg et al. (*1*). Other algorithms have been presented which yield the same prediction results but are different in the individual steps (*10*, *12*, *34*). Kowalski and co-workers have presented detailed explanations of PLS2 which include the geometric (*12*, *35*) or theoretical aspects (*11*) of the method. However, these discussions of PLS2 tend to obscure the simplicity of the PLS1 algorithm which can be viewed as a series of simplified CLS and ILS steps.

Table I. PLS1 Algorithm for Calibration

step 1. Pretreatment of data		
center **A** and **c**		
scale **A** (optional, see text)		
set index h to 1		
step 2. Forming the weight loading vector, $\hat{\mathbf{w}}_h$		
model	$\mathbf{A} = \mathbf{c}\mathbf{w}_h' + \mathbf{E}_A$	(10)
L.S. solution	$\hat{\mathbf{w}}_h = \mathbf{A}'\mathbf{c}/\mathbf{c}'\mathbf{c}$	(11)
	normalize $\hat{\mathbf{w}}_h$	
step 3. Formation of the score (latent variable) vector, $\hat{\mathbf{t}}_h$		
model	$\mathbf{A} = \mathbf{t}_h\hat{\mathbf{w}}_h' + \mathbf{E}_A$	(12)
L. S. solution	$\hat{\mathbf{t}}_h = \mathbf{A}\hat{\mathbf{w}}_h/\hat{\mathbf{w}}_h'\hat{\mathbf{w}}_h = \mathbf{A}\hat{\mathbf{w}}_h$	(13)
step 4. Relating score vector, $\hat{\mathbf{t}}_h$, to the concentrations		
model	$\mathbf{c} = \mathbf{v}_h\hat{\mathbf{t}}_h + \mathbf{e}_c$	(14)
L. S. solution	$\hat{\mathbf{v}}_h = \hat{\mathbf{t}}_h'\mathbf{c}/\hat{\mathbf{t}}_h'\hat{\mathbf{t}}_h$	(15)
step 5. Formation of $\hat{\mathbf{b}}_h$, the PLS loading vector for **A**		
model	$\mathbf{A} = \hat{\mathbf{t}}_h\mathbf{b}_h' + \mathbf{E}_A$	(16)
L. S. solution	$\hat{\mathbf{b}}_h = \mathbf{A}'\hat{\mathbf{t}}_h/\hat{\mathbf{t}}_h'\hat{\mathbf{t}}_h$	(17)
step 6. Calculation of the residuals in **A** and **c**		
spectral residuals	$\mathbf{E}_A = \mathbf{A} - \hat{\mathbf{t}}_h\hat{\mathbf{b}}_h'$	(18)
concentration residuals	$\mathbf{e}_c = \mathbf{c} - \hat{v}_h\hat{\mathbf{t}}_h$	(19)
step 7. Increment h, substitute $\mathbf{E}_A$ for **A** and $\mathbf{e}_c$ for **c** in step 2 and continue for desired numbers of loading vectors		

Table II. PLS1 Algorithms for Prediction

Method 1

step 1.	center **a** using calibration data scale **a** if calibration spectra were scaled set $h = 1$.	
step 2.	$t_h = \hat{\mathbf{w}}_h'\mathbf{a}$	(20)
step 3.	$c_h = c_{h-1} + \hat{v}_h t_h$	(21)
step 4.	$\mathbf{e}_h = \mathbf{e}_{h-1} - \hat{\mathbf{b}}_h t_h$	(22)
step 5.	increment h, substitute $\mathbf{e}_h$ for **a** and repeat with step 2 until $h = r$	
note:	$\hat{\mathbf{w}}_h$, $\hat{v}_h$, and $\hat{\mathbf{b}}_h$ are from the PLS1 calibration, c_0 is the average concentration of the analyte in the calibration samples and $\mathbf{e}_0 = \mathbf{a}$.	

Method 2

step 1.	after PLS1 calibration, calculate the final regression coefficients $\mathbf{b}_f = \hat{\mathbf{W}}'(\hat{\mathbf{B}}\hat{\mathbf{W}}')^{-1}\hat{\mathbf{v}}$	(23)
step 2.	center **a** using calibration data (scale **a** if calibration spectra were scaled)	
step 3.	$\hat{c} = \mathbf{a}'\mathbf{b}_f + c_0$	(24)
note:	$\hat{\mathbf{W}}$ and $\hat{\mathbf{B}}$ have r rows composed of vectors $\hat{\mathbf{w}}_h$ and $\hat{\mathbf{b}}_h$, and $\hat{\mathbf{v}}$ is formed from the individual $\hat{v}_h$ terms for the optimal PLS1 model ($h = r$); c_0 is the same as given in method 1.	

PLS1 Calibration (Table I). *Step 1.* Geladi and Kowalski (*12*) have discussed the pretreatment of the data in some detail, and this discussion will not be repeated here. We will point out that mean centering the data (i.e., the average

calibration spectrum is subtracted from each spectrum, and the average calibration concentration is subtracted from each concentration) eliminates the need to fit a nonzero intercept, and therefore, centering can often decrease the complexity of the model by reducing the number of PLS factors required to model the data by one (see ref 24).

After centering, the spectral data may also be scaled at each frequency. Scaling is performed to give greater weight to those frequencies with greater information content. Since the proper weighting is not often known a priori, the spectral data at each frequency either are not scaled or are autoscaled to unit variance (see ref 12). Centering and scaling each make the computations less prone to roundoff and overflow problems. However, it must be stressed that the results of the PLS analysis depend on the nature of the scaling of the spectral data, and different results are to be expected depending on how the spectral data are pretreated (this is true of PCR also). Therefore, it may not be appropriate to scale spectral data if, as is often the case, the errors are independent of the magnitude of spectral intensity changes. Autoscaling can also degrade the results if much of the data contain spectral regions with little or no spectral variation. In this case, data with minimal spectral variation will contain primarily noise but will be given the same importance in the analysis as data which experiences composition-related variations. On the other hand, autoscaling may also be useful for deemphasizing the effects of chemical components with large spectral features that may not be of interest in the analysis. In the following steps, it will be assumed that the data are always mean centered but scaled only if appropriate.

Step 2. This step in the PLS1 algorithm is actually a classical least-squares calibration in which the analysis is performed assuming that the concentrations of only one component are known in the calibration samples. The model used in this classical least-squares analysis is given by eq 10 in Table I. During the first pass through the algorithm (i.e., $h = 1$), the least-squares estimate in eq 11 for $\mathbf{w}_h$, $\hat{\mathbf{w}}_1$, is a $n \times 1$ vector which represents the least-squares estimate (i.e., a first-order approximation) of the pure-component spectrum for the component of interest. The first normalized weight vector, $\hat{\mathbf{w}}_1$, is therfore a vector that is proportional to a weighted average of the centered calibration spectra, the weights in the average being proportional to the centered concentrations of the component being analyzed. Subsequent $\hat{\mathbf{w}}_h$ vectors are constructed to be mutually orthogonal. This step is a departure from the NIPALS algorithm for PCA since concentration information is directly introduced into the calculation of the loading vectors.

Step 3. This step of the PLS1 algorithm is similar to a classical least-squares prediction step in which the assumption is continued that only one component is present in the calibration samples. Therefore, we use a model similar to that in eq 10, and our first-order, least-squares approximation to the pure-component spectrum for the component of interest, i.e., $\hat{\mathbf{w}}_1$, is used in a one-component CLS prediction step to estimate the amount (or concentration) of the $\hat{\mathbf{w}}_h$ spectral component in each of the calibration spectra. However, we substitute $\mathbf{t}_h$ (sometimes called the latent variable (*10*)) for $\mathbf{c}$ in the model in eq 10 since the least-squares solution will be only a first-order approximation to $\mathbf{c}$. Thus, the model used in this step is eq 12. The least-squares estimate of $\mathbf{t}_h$, $\hat{\mathbf{t}}_h$, is obtained by regressing $\mathbf{A}$ on $\hat{\mathbf{w}}_h'$ (Table I, eq 13). The individual elements of $\hat{\mathbf{t}}_h$, therefore, indicate how much of $\hat{\mathbf{w}}_h$ is contained in each calibration spectrum. The vector $\hat{\mathbf{t}}_1$ represents the intensities (or amounts) of the first weight loading vector in the calibration samples for the new PLS coordinate system. Since $\hat{\mathbf{w}}_1$ is a first-order attempt to represent the pure-component spectra from the calibration spectra, $\hat{\mathbf{t}}_1$ represents a first-order attempt to determine the amount of the pure component (i.e. its concentration) in each calibration spectrum. With PLS, each $\hat{\mathbf{t}}_h$ vector is related to both $\mathbf{A}$ and $\mathbf{c}$ rather than solely to $\mathbf{A}$ as in PCA.

At this point, it is interesting to note that if there had been only one spectrally active component in the spectral region being analyzed, then steps 2 and 3 of the PLS1 algorithm are identical with the basic CLS calibration and prediction algorithms. Thus $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{t}}_1$ are no longer just first-order approximations of the pure-component spectra and concentrations, respectively; they are the normalized, centered pure-component spectra and the centered concentration. Only in the presence of two or more independently varying spectral components will the two methods be different and make the next steps necessary.

Step 4. The score vector (latent spectral variable), $\hat{\mathbf{t}}_h$, representing the intensities in the new PLS coordinate system can be related to concentrations using a linear least-squares regression just as spectral intensities are related to concentrations in the ILS analysis or as the scores in PCR are related to concentrations. However, in PLS a separate relation between scores, $\hat{\mathbf{t}}_h$, and concentrations (or concentration residuals) is found after each weight vector is estimated. The relation between $\hat{\mathbf{t}}_h$ and $\mathbf{c}$ is modeled by eq 14. Equation 15 gives the estimate for v_h which is the scalar regression coefficient relating $\hat{\mathbf{t}}_h$ to the concentration of the component of interest. Equation 15 is similar to the ILS solution in eq 6 in that the sum of squared concentration errors is minimized. However, in this step of PLS1, the ILS-like solution is constructed one element at a time.

Step 5. Orthogonal $\hat{\mathbf{t}}_h$ vectors are desirable in order to remove the collinearity problem which was present in the original inverse least-squares regression (eq 6). Orthogonal $\hat{\mathbf{t}}_h$ vectors can be obtained by forming a new model for $\mathbf{A}$ based on the latent variable $\hat{\mathbf{t}}_h$. The new model for $\mathbf{A}$ is given in Table I, eq 16, where $\mathbf{b}_h$ is the $n \times 1$ PLS loading vector. This step in combination with the next step of the algorithm assures that the $\hat{\mathbf{t}}_h$ vectors will be mutually orthogonal. The least-squares regression is simultaneously performed at all frequencies and is given in Table I, eq 17. Unlike the first PCA loading vector, the first PLS loading vector, $\hat{\mathbf{b}}_1$ determined from eq 17 does not account for the maximum variance in the calibration spectra, $\mathbf{A}$. It represents an attempt to account for as much variation in $\mathbf{A}$ while simultaneously correlating with $\hat{\mathbf{t}}_h$ which approximates $\mathbf{c}$. Also unlike PCA, the $\hat{\mathbf{b}}_h$ vectors which comprise the matrix $\mathbf{B}$ in eq 8 are not mutually orthogonal. Since $\hat{\mathbf{t}}_1$ is a first-order approximation to the centered concentrations, frequencies associated with the largest positive elements in $\hat{\mathbf{b}}_1$ tend to indicate those frequencies which exhibit the greatest dependence on concentration for that particular loading vector. However, the $\hat{\mathbf{w}}_1$ vector which is directly related to $\mathbf{c}$ will exhibit this tendency better than $\hat{\mathbf{b}}_1$, and therefore $\hat{\mathbf{w}}_1$ will be more useful than $\hat{\mathbf{b}}_1$ for extracting qualitative information from the PLS1 analysis.

Step 6. The product of the scores ($\hat{\mathbf{t}}_h$) and loading vector ($\hat{\mathbf{b}}_h$) is the PLS approximation to the calibration spectra. Residuals, $\mathbf{E}_A$, in the calibration spectra $\mathbf{A}$ are computed by subtracting the PLS approximation to the calibration spectra from the measured calibration spectra as given in eq 18. Similarly, we remove the part of the concentrations that have been modeled by PLS to obtain the concentration residuals $\mathbf{e}_c$ as given in eq 19. The product, $\hat{v}_h\hat{\mathbf{t}}_h$, in eq 19 represents the PLS estimated concentration based on the spectrum.

PLS1 Prediction (Table II). PLS1 prediction can be obtained from an unknown sample spectrum, $\mathbf{a}$, after centering by applying one of the two methods outlined in Table II. The first method is more involved but allows spectral residuals to be calculated. This first method involves the calculation

of the spectral intensities (t_h in eq 20) of the sample spectrum in the new full-spectrum PLS coordinate system obtained during PLS calibration. These intensities are then related to analyte concentration by using a prediction equation analogous to ILS prediction (i.e., eq 21 summed for all values of h from 1 to r is similar to the ILS prediction given in eq 7). For the optimal number factors in the model (i.e., $h = r$), the prediction of concentration based on the unknown sample spectrum and eq 21 is then c_r. From $\mathbf{e}_r$ in eq 22, a measure of the ability of the calibration set to fit the sample spectrum can be obtained. This is presented in more detail in Appendix B which discusses outlier detection.

The second method for obtaining concentration predictions from PLS1 involves the calculation and use of the vector of final calibration coefficients, $\mathbf{b}_f$. The vector, $\mathbf{b}_f$, has dimensions of an individual spectrum, and it can be calculated in several ways. One direct method is given in ref 10 and is presented as method 2 in Table II. The $\mathbf{b}_f$ vector need only be calculated once from the calibration results. Although $\mathbf{b}_f$ provides an efficient method to estimate concentrations from any unknown sample spectrum (eq 24), it does not allow a determination of the residual spectrum. Therefore, no diagnostic information about the quality of the fit is available when predictions are obtained by using the $\mathbf{b}_f$ vector.

Selection of the Optimal Number of Factors for the PLS Model. We would like to select the number of loading vectors, r (or alternatively, the number of scores or factors), in the PLS algorithm which will allow us to model as much of the complexity of the system without overfitting the concentration data. To accomplish this goal, we use the cross-validation method leaving out one sample at a time (*36*). Given a set of m calibration spectra, we perform the PLS1 calibration on $m - 1$ calibration spectra, and using this calibration, we predict the concentration of the sample left out during calibration. This process is repeated a total of m times until each sample has been left out once. The concentration predicted for each sample is then compared with the known concentration of this reference sample. The sum of the squared concentration prediction errors (i.e., $\mathbf{e}_c'\mathbf{e}_c$) for all calibration samples (prediction error sum of squares or PRESS) is a measure of how well a particular PLS model fits the concentration data. PRESS is calculated in the same manner each time a new factor is added to the PLS1 model. One reasonable choice for the optimal order (number of factors) of a PLS1 model would be that order which yielded the minimum PRESS. Although PRESS is a reasonable measure to use to evaluate the "goodness" of a model, it is based on a finite number of samples, and therefore, it is subject to error. Thus, using the number of factors (h^*) which yields a minimum in PRESS can lead to some overfitting. A better criterion to select the optimal model involves the comparison of PRESS from models with fewer than h^* factors. The model selected is the one with the fewest number of factors such that PRESS for that model is not significantly greater than PRESS for the model with h^* factors (the **F** statistic is used to make the significance determination). Use of this criterion gives rise to more parsimonious PLS models involving fewer factors and mitigates the overfitting problem. The details of the selection procedure are given in Appendix A. Another possible criterion to select the optimal model which may not be as sensitive to the presence of outliers involves estimation of the error in PRESS (*37*). The model selected would then be that model which has the fewest number of loading vectors which yields a PRESS within one standard error of the PRESS obtained from the model yielding the minimum PRESS. We have used both methods and find empirically that only occasionally have the two methods yielded different numbers of factors. However, we prefer the former method since in this case statistical probabilities can be assigned during the model selection procedure or when comparing competing calibration methods (see Appendix A). Once the optimal number of PLS factors is determined, it is necessary to perform the final calibration, using all m calibration samples with the optimal number of PLS factors.

Obtaining Qualitative Information from PLS Calibration and Prediction. Since these PLS1 methods are capable of being full-spectrum methods, chemically interpretable qualitative information sometimes can be obtained. The discussion of the PLS1 algorithm in this paper suggests that the first weight loading vector, $\hat{\mathbf{w}}_1$, should contain useful qualitative spectral information since it is the first-order approximation to the pure-component spectrum of the analyte. Thus, $\mathbf{w}_1$, may be useful for making band assignments and determining which regions of the spectrum are most relevant to a particular analyte. The vector of final calibration regression coefficients, $\mathbf{b}_f$, from the calibration may also contain interpretable information. The $\mathbf{b}_f$ vector indicates which spectral regions are important for prediction and is related to the pure-component spectrum taking into account all the effects of interfering components, molecular interactions, and base-line variations. Finally, useful chemical information and outlier detection are available from the spectral residuals.

Because the CLS method yields direct Beer's law estimates of the pure-component spectra, it will yield a higher quality of chemical information than is possible from PLS. In the past, the CLS method has been used to determine the presence and source of deviations in Beer's law. The type of qualitative information that can be obtained with CLS methods includes (1) the presence of molecular interactions and which parts of the molecules present in the sample interact, (2) the presence of spectrometer nonlinearities (*16*), (3) the presence and identity of unexpected components in the unknown samples (*24*, *26*), (4) the presence of outliers (*38*), (5) which components in reacting mixtures react and what are the reaction products (*27*), and (6) information leading to rapid chemical or structural assignments of the spectral bands (*38*). The list of direct qualitative information obtained from the PLS1 method is not as extensive, but some information is, nevertheless, possible.

Direct information about the presence and source of nonlinearities is not readily available from PLS (or PCR) because the method is capable of modeling some types of nonlinearities. True nonlinearities cannot be fit with linear models used in PLS and PCR. However, interactions which cause new types of molecular bonding can often be described with an additive linear model, yet are considered nonlinear in Beer's law because we have no explicit information about the concentrations of the new species generated by the interaction. PLS or PCR can empirically model the number of new components necessary for prediction, and for this reason PLS and PCR methods are sometimes called soft modeling methods. Thus a system with interactions may be more accurately fit by the PLS or PCR model, and molecular interaction information in the residual spectra is lost. Nevertheless, spectral residuals still indicate which regions of the spectra do not follow the overall PLS model. Therefore, outlier detection and the identification of unexpected components are still possible with PLS (*24*, *38*). Spectral residuals can be determined directly from the calibration or prediction analyses (i.e., eq 18 and 22, respectively, when $h = r$). It is shown in the companion paper (*24*) that these full-spectrum residuals allow outlier detection and help indicate the presence and identity of unexpected components in the samples.

EXPERIMENTAL SECTION

Simulated spectral data were prepared to demonstrate the qualitative information available from the PLS algorithm. These

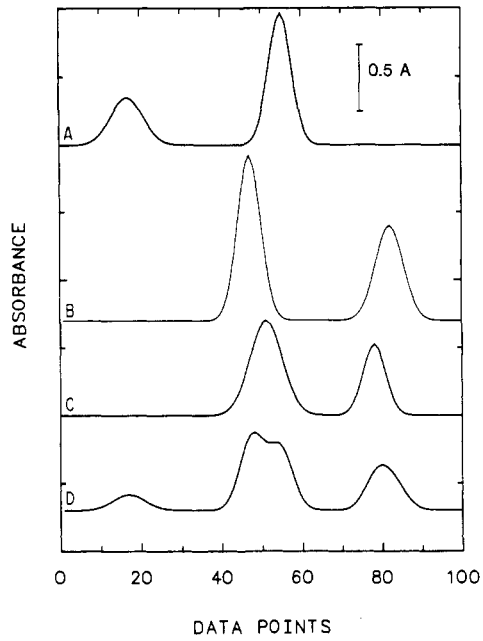

Figure 1. Pure-component spectra of simulated spectral data without noise: (A) component 1, (B) component 2, (C) component 3, and (D) equal molar mixture of the three components.

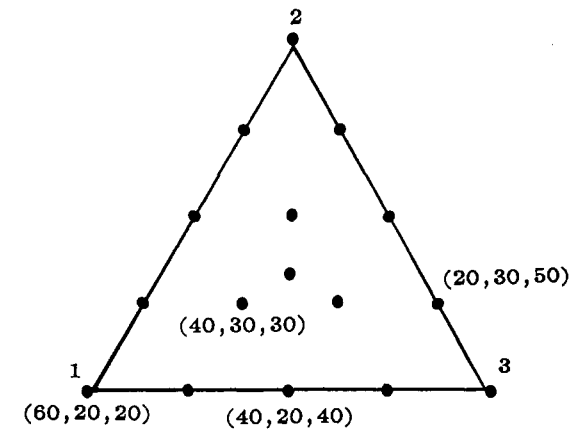

Figure 2. Mixture design for the three-component constrained mixtures used in the simulated data sets.

spectra simulated a three-component mixture system with the constraint that the concentrations of the components summed to 100 mol %. The pure-component spectra were each composed of two Gaussian bands of different intensities and bandwidths. The pure-component spectra and the spectrum of the equal molar mixture generated from them are shown in Figure 1. Spectral data sets of 16 samples were taken either from a mixture design (*39*) to statistically maximize the information content in the spectra or at random (uniform over the simplex). Figure 2 illustrates the mixture design in the calibration sample concentrations. Each component concentration was constrained to be ≥20 and ≤60 mol %.

DEMONSTRATION OF QUALITATIVE INFORMATION DERIVED FROM PLS

The first weight vector, $\hat{\mathbf{w}}_1$, in the PLS1 analysis is a first-order approximation to the pure-component spectrum of that component, so there is often interpretable information available in this vector which can be useful in making assignments of spectral bands. However, the quality of $\hat{\mathbf{w}}_1$ is dependent to some degree on relative intensities of spectral bands, and the information in $\hat{\mathbf{w}}_1$ will be less useful if the analyte of interest only has relatively small spectral features or relatively small spectral variation in the calibration set. The

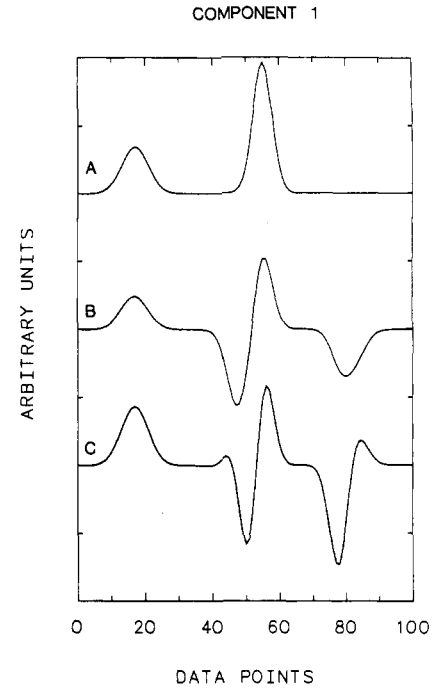

Figure 3. Analysis of simulated data without noise, component 1: (A) pure-component spectrum, (B) first PLS1 weight loading vector, $\hat{\mathbf{w}}_1$, and (C) PLS1 vector of regression coefficients, $\mathbf{b}_f$.

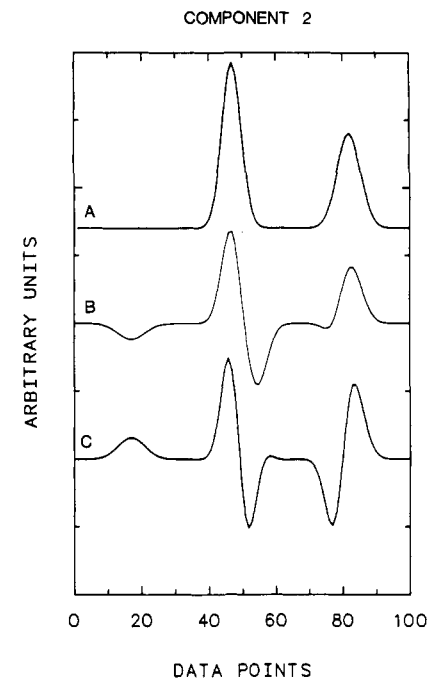

Figure 4. Analysis of simulated data without noise, component 2: (A) pure-component spectrum, (B) first PLS1 weight loading vector, $\hat{\mathbf{w}}_1$, and (C) PLS1 vector of regression coefficients, $\mathbf{b}_f$.

weight vector also depends on the calibration design (even in a noiseless system), and orthogonal factorial designs (*39*) will generate the most useful qualitative information by maximizing the pure-component information content in $\hat{\mathbf{w}}_1$. Therefore, statistically designed calibration sets should always be used when qualitative information is important during PLS analysis.

The vector of calibration regression coefficients, $\mathbf{b}_f$, can also contain useful information that is less dependent on the calibration design, but interferences of three or more components in a given spectral band can cause problems for interpretation due to the required compensation in other spectral regions, as will be discussed. Figures 3–5 show the

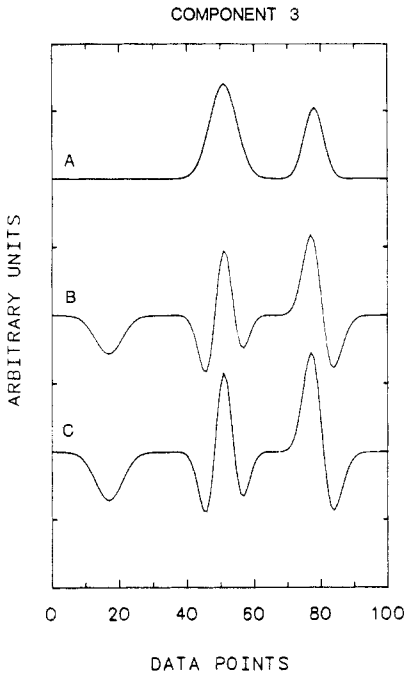

Figure 5. Analysis of simulated data without noise, component 3: (A) pure-component spectrum, (B) first PLS1 weight loading vector, $\hat{\mathbf{w}}_1$, and (C) PLS1 vector of regression coefficients, $\mathbf{b}_f$.

pure-component spectra, the first weight loading vector ($\hat{\mathbf{w}}_1$), and the regression coefficients ($\mathbf{b}_f$) for each of the three components in simulated data sets without noise. Of course, in this case of noiseless data, the pure-component spectrum is the same as that estimated by CLS methods from the calibration set of mixtures. It is clear that $\hat{\mathbf{w}}_1$ yields information that would be useful in making band assignments. In this case, all positive peaks are due to the component of interest while negative peaks correspond to interfering components. The $\mathbf{b}_f$ vector also has interpretive information, but as can be seen in Figure 4, the interpretation is not as straightforward in component 2 where $\mathbf{b}_f$ has a positive peak for the isolated band of component 1. This positive band is due to the presence of two interferences in the middle band which must be compensated for in the isolated band of component 1 and illustrates the potential problem of using $\mathbf{b}_f$ for obtaining qualitative information.

PCR has no vector comparable to $\hat{\mathbf{w}}_1$, but it does yield a vector of regression coefficients (*10*) which is similar to that in the PLS analysis (it is identical with the PLS vector of regression coefficients for a noiseless system that follows Beer's law), and individual PCA loading vectors could be examined as well. Although some limited qualitative information is present in the PCR loading vectors (*40*), they are the same for all chemical components. Unless only one of the major PCR loading vectors is strongly related to the component of interest, PCR in general will be less useful for obtaining qualitative information than either CLS or PLS.

CONCLUSIONS

The full-spectrum multivariate calibration methods (CLS, PCR, and PLS) are all shown to reduce the calibration spectral intensity data at many frequencies to a relatively small number of intensities in a transformed full-spectrum coordinate system. The PLS1 calibration algorithm has been shown to be composed of simplified CLS and ILS calibration and prediction steps. Therefore, PLS exhibits many of the advantages of ILS and CLS methods for spectral analyses without suffering the disadvantages of these more commonly used statistical methods. The detailed understanding of the PLS method helps to identify the intermediate steps in the algorithm which contain chemically interpretable spectral information. The qualitative information available from CLS is greater than that obtained from PLS. Yet, PLS can yield some chemically interpretable information that is useful for making band assignments. In addition, full-spectrum residuals obtained during PLS analyses, like those obtained from CLS, can be useful for determining the presence and possibly the identity of unexpected components (*24*). Statistically designed calibration sets maximize the qualitative information obtained during PLS analysis relative to calibration samples that are obtained from samples with random concentrations. These points are illustrated with simulated and real data in the companion paper (*24*).

New methods have been presented for the selection of the optimal number of factors for PLS and PCR. Rather than selection of the model which yields a minimum in prediction error variance or PRESS, the model selected is the one with the fewest number of factors such that PRESS for that model is not significantly greater than the minimum PRESS. This reduces the possibility of overfitting the data while providing sufficient numbers of PLS or PCR factors to adequately model the data. The same general method to obtain the optimal model can also be used to compare the prediction abilities of different calibration methods.

Finally, it should be noted that ILS, PCR, and PLS all have the potential to estimate not only component concentrations but also chemical and physical properties from their infrared spectra. This can be accomplished with ILS only if a proper selection of a small number of spectral frequencies is made. However, PLS and PCR can both be used as full-spectrum calibration methods to estimate properties from the spectra of materials. This potential to estimate material properties can be realized if (1) the properties are dependent on the molecular structure of the material, (2) changes in molecular structure associated with changes in the property are reflected in the infrared spectra, and (3) the properties are linearly related to spectral intensities. Thus, in the discussions presented in this paper, it should be assumed that property could be substituted for concentration whenever it appeared in the paper in conjunction with ILS, PCR, and PLS methods.

APPENDIX A. USE OF THE *F* STATISTIC FOR SELECTING THE OPTIMAL NUMBER OF FACTORS AND COMPARING THE RESULTS FROM COMPETING METHODS

As a guide to select the optimal model, we have computed PRESS for cross-validated models with various numbers of factors (or loading vectors). The model that yields the minimum PRESS is used as a benchmark, and the number of factors associated with this model will be denoted by h^*. All models with fewer factors ($h < h^*$) are compared against this benchmark. The purpose of this comparison is to find the smallest model (fewest number of factors) such that PRESS for this model is not significantly greater than PRESS for the model with h^* factors.

The selection of the optimal model proceeds as follows:

Step 1. Compute $F(h)$ = PRESS (model with h factors)/PRESS (model with h^* factors) for $h = 1, 2, \ldots, h^*$.

Step 2. Choose as the optimal number of factors the smallest h such that $F(h) < F_{\alpha;m,m}$ where $F_{\alpha;m,m}$ is the $(1 - \alpha)$ percentile of Snedecor's F distribution with m and m degrees of freedom (m is the number of calibration samples).

Assuming that the prediction errors have zero mean and are mutually independent (both within and between models) and normally distributed, Prob $\{F(h) > F_{\alpha;m,m} | \sigma_h^2 = \sigma_{h^*}^2\} = 2\alpha$. Here, σ_h^2 and $\sigma_{h^*}^2$ represent the prediction error variance of the model with h and h^* factors, respectively, and α is a probability value to be selected. Notice that the probability is 2α rather than α. This is because we have selected

PRESS(h^*) to be the denominator of the F statistic rather than randomly selecting between PRESS(h) and PRESS(h^*). Therefore, the computed F statistic is always greater than or equal to 1. Note also that, in practice, Prob $\{F > F_{\alpha;m,m}\} < 2\alpha$ because of the positive correlation of prediction errors between models with h and h^* factors. The potential effect of this positive correlation is underfitting (selecting a model with too few factors).

In general, the size of the selected model increases with α. If α is too small, then there will likely be underfitting problems, whereas if α is too large, then overfitting will result. We have found that the choice of $\alpha = 0.25$, although somewhat arbitrary, is a good compromise in practice.

A F statistic, based on PRESS, can also be used to aid in the comparison of the prediction abilities of two different calibration methods. Let F = PRESS(method 1)/PRESS(method 2). Assuming that the prediction errors have zero mean and are mutually independent (both within and between methods) and normally distributed, Prob $\{F > F_{\alpha;m,m} | \sigma_1^2 = \sigma_2^2\} = \alpha$. Here σ_1^2 and σ_2^2 represent the prediction error variances of each of the two methods. One would reject the hypothesis that $\sigma_1^2 = \sigma_2^2$ in favor of $\sigma_1^2 > \sigma_2^2$ if $F > F_{\alpha;m,m}$. Because of the need for a definite probability interpretation, α is not chosen arbitrarily. Common choices for α, which is the probability of falsely concluding that $\sigma_1^2 > \sigma_2^2$, when in fact $\sigma_1^2 = \sigma_2^2$, are 0.1 and 0.05.

In the likely event that the prediction errors are positively correlated between methods, then Prob $\{F > F_{\alpha;m,m}\} < \alpha$. In this case, the α-level hypothesis test based on F and $F_{\alpha;m,m}$ is conservative. This means that, although we lose some ability to discriminate between methods with different prediction error variances, the chance of falsely concluding the prediction error variances are different, when they are not, is less than α.

While we believe that the procedures outlined above for model selection and comparison of methods will work reasonably well in practice, we admit that they could be improved by taking into consideration the correlation of prediction errors. Further work to establish procedures that provide a more suitable solution to this problem is proceeding. This work will be reported on later.

APPENDIX B. OUTLIER DETECTION

During calibration, outliers can be detected both from concentration F ratios or spectral F ratios while outlier detection in the unknown samples must rely solely on the spectral F ratios. The concentration F ratio (not leverage corrected, see below), $F_{1,m-1}(c_j)$, for each sample during cross-validated calibration is calculated from

$$F_{1,m-1}(c_j) = (m-1)\left(e_{c_j}^2\right) \Big/ \left(\sum_{i \neq j} e_{c_i}^2\right) \qquad \text{(B-1)}$$

where the subscripts on $F(c_j)$ indicate the degrees of freedom and e_{c_i} is the difference between reference and estimated concentrations for the ith sample left out of the calibration during cross validation. The spectral F ratio for each sample during cross-validated calibration is calculated from

$$F_{x,y}(\mathbf{a}_j) = (m-1)\left(\sum_{k=1}^{n} e_{a_j k}^2\right) \Big/ \left(\sum_{i \neq j} \sum_{k=1}^{n} e_{a_i k}^2\right) \qquad \text{(B-2)}$$

where $\mathbf{a}_j$ represents the spectrum of the sample left out during cross validation, the subscripts on $F(\mathbf{a}_j)$ indicate the degrees of freedom and n is the number of frequencies included in the analysis. Because of the likelihood of nonconstant variance and dependence among the various error elements in eq B-2, it is not possible to give a simple and general formula for the effective degrees of freedom, x and y. Lindberg et al. (*1*) believe that $x = (n-r)/2$ and $y = (n-r)(m-r-1)/2$ are good estimates of x and y. However, for practical purposes, when there are hundreds of frequencies and no outliers, the F ratio will always be close to 1. For example, if the effective degrees of freedom are each 120, then the probability that F < 1.5 (given there is no outlier) is about 0.01. However, for simulated and real IR spectral data where the error is dependent on the absorbance levels, spectral F ratios less than 3 do not appear to indicate a highly significant outlier.

The appearance of F ratios close to 3 is in part due to the fact that the residuals are not leverage corrected. Weisberg (*41*) illustrates how residuals can be leverage corrected in the context of linear regression. However, it is not clear how to make leverage corrections here since PLS, as we have shown, is a hybrid method that uses both classical and inverse least-squares steps.

For unknown samples, the spectral F ratios are calculated by using the final calibration of all m calibration samples. Therefore, the spectral F ratios for an unknown sample are given by

$$F_{x,y}(\mathbf{a}_s) = m\left(\sum_{k=1}^{n} e_{a_s k}^2\right) \Big/ \left(\sum_{i=1}^{m} \sum_{k=1}^{n} e_{a_i k}^2\right) \qquad \text{(B-3)}$$

where $\mathbf{a}_s$ represents the spectrum of the unknown sample. The F ratios can be used as guide in the calibration and prediction steps to flag possible outlier samples (see ref 42). This is especially important when these methods are used on a large number of samples, such as quality control applications, where the individual spectral residuals may not be examined for each sample.

LITERATURE CITED

(1) Lindberg, W.; Persson, J.-A.; Wold, S. *Anal. Chem.* **1983**, *55*, 643.
(2) Otto, M.; Wegscheider, W. *Anal. Chem.* **1985**, *57*, 63.
(3) Martens, H.; Jensen, S. A. In *Proceedings, 7th World Cereal and Bread Congress, Prague, June, 1982*; Holas, J., Kratochvil, J., Eds.; Elsevier: Amsterdam, 1983; pp. 607–647.
(4) Geladi, P.; MacDougall, D.; Martens, H. *Appl. Spectrosc.* **1985**, *39*, 491.
(5) Martens, M.; Martens, H. *Appl. Spectrosc.* **1986**, *40*, 303.
(6) Dunn III, W. J.; Stalling, D. L.; Schwartz, T. R.; Hogan, J. W.; Petty, J. D.; Johansson, E.; Wold, S. *Anal. Chem.* **1984**, *56*, 1308.
(7) Lindberg, W.; Ohman, J.; Wold, S.; Martens, H. *Anal. Chim. Acta* **1985**, *171*, 1.
(8) Lindberg, W.; Ohman, J.; Wold, S.; Martens, H. *Anal. Chim. Acta* **1985**, *174*, 41.
(9) Otto, M.; Thomas, J. D. R. *Anal. Chem.* **1985**, *57*, 2647.
(10) Martens, H. A. "Multivariate Calibration: Quantitative Interpretation of Non-Selective Chemical Data, Parts I and II of the 1985 PhD Dissertation"; The Norwegian Institute of Technology, University of Trondheim: Trondheim, Norway, NR-Report No.: 786 and 787, 1986, ISBN 82-539-0273-5 and 82-539-0274-3.
(11) Lorber, A.; Wangen, L. E.; Kowalski, B. R. *J. Chemom.* **1987**, *1*, 19.
(12) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1 and 19.
(13) Antoon, M. K.; Koenig, J. H.; Koenig, J. L. *Appl. Spectrosc.* **1977**, *31*, 518.
(14) Haaland, D. M.; Easterling, R. G. *Appl. Spectrosc.* **1980**, *34*, 539.
(15) Haaland, D. M.; Easterling, R. G. *Appl. Spectrosc.* **1982**, *36*, 665.
(16) Haaland, D. M.; Easterling, R. G.; Vopicka, D. A. *Appl. Spectrosc.* **1985**, *39*, 73.
(17) Brown, C. W.; Lynch, P. F.; Obremski, R. J.; Lavery, D. S. *Anal. Chem.* **1982**, *54*, 1472.
(18) Kisner, H. J.; Brown, C. W.; Kavarnos, G. J. *Anal. Chem.* **1983**, *55*, 1703.
(19) Maris, M. A.; Brown, C. W.; Lavery, D. S. *Anal. Chem.* **1983**, *55*, 1694.
(20) Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.
(21) Fredericks, P. M.; Lee, J. B.; Osborn, P. R.; Swinkels, D. A. J. *Appl. Spectrosc.* **1985**, *39*, 303.
(22) Fredericks, P. M.; Lee, J. B.; Osborn, P. R.; Swinkels, D. A. J. *Appl. Spectrosc.* **1985**, *39*, 311.
(23) Brown, C. W.; Obremski, R. J.; Anderson, P. *Appl. Spectrosc.* **1986**, *40*, 734.
(24) Haaland, D. M.; Thomas, E. V. *Anal. Chem.*, following paper in this issue.
(25) Haaland, D. M. *Proc. SPIE* **1985**, *553*, 241.
(26) Haaland, D. M.; Barbour, R. L. *Am. Lab.* (*Fairfield, Conn.*) **1985**, *17*(7), 14.
(27) Ward, K. J.; Brinker, C. J.; Haaland, D. M., submitted to *Appl. Spectrosc.*
(28) Haaland, D. M. *Spectroscopy* **1987**, *2*(6), 56.
(29) Haaland, D. M.; Thomas, E. V. Presented at the 1986 Eastern Analytical Symposium, New York, Paper 86.
(30) Wetzel, D. L. *Anal. Chem.* **1985**, *55*, 1165A.

(31) Honigs, D. E.; Freelin, J. M.; Hieftje, G. M.; Hirschfeld, T. B. *Appl. Spectrosc.* **1983**, *37*, 491.
(32) Malinoswki, E. R.; Howery, D. G. *Factor Analysis in Chemistry*; Wiley: New York, 1980.
(33) Wold, H. *Multivariate Analysis*; Krishnaiah, P. R., Ed.; Academic: New York, 1966; p 391.
(34) Manne, R. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 187.
(35) Beebe, K. R.; Kowalski, B. R. *Anal. Chem.* **1987**, *59*, 1007A.
(36) Stone, M. *J. R. Statistical Soc., B* **1974**, *36*, 111.
(37) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: Belmont, CA, 1984.
(38) Haaland, D. M. *Anal. Chem.* **1988**, *60*, 1208-1217.
(39) Cornell, J. A. *Experiments with Mixtures*; Wiley: New York, 1981.
(40) Cowe, I. A.; McNicol, J. W. *Appl. Spectrosc.* **1985**, *39*, 257.
(41) Weisberg, S. *Applied Linear Regression*; Wiley: New York, 1985.
(42) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; Wiley: New York, 1978.

RECEIVED for review September 30, 1987. Accepted February 8, 1988. This work was performed at Sandia National Laboratories and supported by the U.S. Department of Energy under Contract DE-AC04-DP00789. Portions of this paper were presented at the 1986 Eastern Analytical Symposium, New York, October 20-24, 1986, Paper 86.

Partial Least-Squares Methods for Spectral Analyses. 2. Application to Simulated and Glass Spectral Data

David M. Haaland* and Edward V. Thomas

Sandia National Laboratories, Albuquerque, New Mexico 87185

Partial least-squares (PLS) methods for quantitative spectral analyses are compared with classical least-squares (CLS) and principal component regression (PCR) methods by using simulated data and infrared spectra from bulk seven-component, silicate-based glasses. Analyses of the simulated data sets show the effect of data pretreatment, base-line variations, calibration design, and constrained mixtures on PLS and PCR prediction errors and model complexity. Analyses of the simulated data sets also illustrate some qualitative differences between PLS and PCR. For example, the PLS model approaches the optimal prediction model more rapidly than the PCR model and is computationally more efficient. PLS and PCR predicted concentration errors from the simulated data sets and a set of the Fourier transform infrared spectra of silicate-based glasses (S-glass) show that prediction errors are not statistically different between these two methods for these individual data sets with limited numbers of samples. However, PLS and PCR are both superior to CLS methods in the case of the analysis of S-glass where only one analyte is known in the calibration samples and the components of unknown concentration overlap all the spectral features of the analyte components. CLS analysis precision significantly improves when the three known analyte concentrations (B_2O_3, P_2O_5, and OH) are used in the calibration. In this latter case, PLS and PCR concentration predictions are unchanged, and although they each still yield a lower standard error of prediction than the CLS method, there is no longer strong statistical evidence that these differences between PLS or PCR and CLS are outside experimental error for the B_2O_3 component. The ability of CLS and PLS methods to provide chemically useful estimates of the pure-component spectra is also demonstrated.

Multivariate statistical methods coupled with computerized spectrometers are making improvements possible in the precision and accuracy of the quantitative spectral analyses of chemical samples. In addition, the range and complexity of problems that can be solved with quantitative spectroscopy have been increased by the use of these methods. In a companion paper (*1*), the partial least-squares (PLS) method for quantitative and qualitative spectral analyses was reviewed, and its relation to classical least-squares (CLS), inverse least-squares (ILS), and principal component regression (PCR) methods was discussed. (CLS and ILS are sometimes referred to as **K** and **P** matrix methods, respectively, by infrared spectroscopists.) PLS has been shown to be highly effective in the quantitative analysis of chemical samples when using near-infrared (*2*) and UV spectra data (*3*). Analyses of simulated spectral data are used in this paper to compare PLS and PCR methods and to illustrate a few of the factors such as base-line variations and data pretreatment which affect model complexity and errors in concentrations estimated from the spectral data. In addition, the quantitative analysis of infrared spectra from a small set of silicate-based glasses is used to illustrate and compare PLS, PCR, and CLS methods for this seven-component glass for which the concentrations of only a few of the chemical components are known. The qualitative information available in the PLS and CLS analyses of the glass is compared. The ILS method was not applied to these data since ILS is not a full-spectrum method, and its performance will be dependent on the frequency selection method employed.

On comparison of available quantitative techniques, it must be determined if the differences in concentration estimation errors for any given data set are statistically significant. Methods to permit these determinations were presented in Appendix A of the companion paper (*1*), and their use is illustrated in this paper. However, care must always be taken that conclusions based on the observed analysis differences from a given data set not be overly generalized since a variety of parameters such as spectral and concentration noise, numbers of frequencies and samples used in the calibration set, relative intensities, degree of spectral overlap, the presence of base-line variations, backgrounds, and model error can all affect the relative performances of these multivariate statistical analysis methods.

EXPERIMENTAL SECTION

Preparation of Simulated Spectral Data Sets. Simulated spectral data were prepared by using three-component mixtures with the spectrum of each pure component consisting of two Gaussian bands with different intensities and bandwidths. These spectral data are described more completely and illustrated in ref 1. The mixtures were constrained so that the concentrations of the components summed to 100 mol %. A variety of spectral data sets were generated. Spectral data sets of 16 samples were

0003-2700/88/0360-1202$01.50/0