Comparison of Multivariate Calibration Methods for Quantitative Spectral Analysis

Edward V. Thomas and David M. Haaland*

Sandia National Laboratories, Albuquerque, New Mexico 87185

The quantitative prediction abilities of four multivariate calibration methods for spectral analyses are compared by using extensive Monte Carlo simulations. The calibration methods compared include inverse least-squares (ILS), classical least-squares (CLS), partial least-squares (PLS), and principal component regression (PCR) methods. ILS is a frequency-limited method while the latter three are capable of full-spectrum calibration. The simulations were performed assuming Beer's law holds and that spectral measurement errors and concentration errors associated with the reference method are normally distributed. Eight different factors that could affect the relative performance of the calibration methods were varied in a two-level, eight-factor experimental design in order to evaluate their effect on the prediction abilities of the four methods. It is found that each of the three full-spectrum methods has its range of superior performance. The frequency-limited ILS method was never the best method, although in the presence of relatively large concentration errors it sometimes yields comparable analysis precision to the full-spectrum methods for the major spectral component. The importance of each factor in the absolute and relative performances of the four methods is compared. A relatively simple model involving the mean squared prediction errors is developed for estimating the prediction errors for each calibration method over the range of variation of the factors considered. These results offer the analyst guidelines to be used in evaluating which multivariate calibration method will provide the best predictions when applied to a given spectral data set. In the absence of specific information about the data set, we would recommend the use of PLS since it is usually optimal or close to optimal.

INTRODUCTION

Recently, quantitative spectroscopy has been greatly improved by the use of a variety of multivariate statistical methods (*1–16*). Multivariate calibrations are useful in spectral analyses because the simultaneous inclusion of multiple spectral intensities can greatly improve the precision and applicability of quantitative spectral analyses. With multivariate calibrations, empirical models are developed that relate the multiple spectral intensities from many calibration samples to the known analyte concentrations of the samples. These empirical relationships can then be used in multivariate prediction analyses of spectra of unknown samples to rapidly predict analyte concentrations. The diversity of multivariate methods available for application to quantitative spectroscopic problems can leave the spectroscopist uncertain as to which method to use for a given set of data. Often, only limited and subjective comparisons between the various multivariate calibration methods have been made in the literature. When detailed comparisons have been made, the comparisons generally were based on a single data set or a small number of data sets (*5–16*). Observed differences between methods applied to a given data set often have not been assessed for statistical significance. While it is certainly possible to demonstrate the superiority of one method over another for a single data set, it is difficult to generalize this superiority to situations involving other data sets. This difficulty arises from the fact that data sets are generally very complex, and each set of data can be characterized by a multitude of underlying factors (i.e., spectral noise, spectral overlap, numbers of samples or intensities included in the calibration, etc.). In principle, the performances of each calibration method are differentially affected by each of these underlying factors. This often means that the relative performances of the methods are dependent on the particular data set being analyzed, and it is found that no single method always yields the best performance. It is important, from an analyst's point of view, that the optimal calibration method be used for a given data set. Currently there is very little guidance available with respect to method selection.

Therefore, it is the intent of this paper to give some insight into determining which method will be optimal for a given set of data. This goal can be achieved by systematically studying the performances of several competing multivariate calibration methods. The particular procedure used to investigate the relative performance of the various methods involved a series of structured Monte Carlo computer simulations. The simulations required the generation of many thousands of data sets each involving many spectra.

The multivariate calibration methods investigated in this paper include the four most common methods supplied by the manufacturers of infrared spectrometers. They are classical least squares (CLS) (*4, 5*), inverse least squares (ILS) (*12, 13*), partial least squares (PLS) (*1, 3, 6–10, 15, 16*), and principal component regression (PCR) (*2, 6–8, 14–16*). Haaland and Thomas (*6*) have presented a recent discussion of these four multivariate calibration methods. Briefly, CLS is a multivariate least-squares procedure based directly on Beer's law. Infrared spectroscopists have sometimes referred to this method as the **K**-matrix method (*12, 13*). The CLS model accounts for errors in the spectral measurements. CLS can accommodate spectral intensities at all frequencies for all calibration samples. In general, all overlapping spectral components should be known for optimal performance of CLS. By being a full-spectrum method, CLS has the ability to achieve improved precision since there is a signal averaging effect when many or all the spectral intensities are included in the analysis. ILS is a least-squares method [sometimes called the **P**-matrix method by infrared spectroscopists (*12, 13*) or multiple linear regression (MLR) when applied to near-infrared data] that uses the inverse of Beer's law as its model. That is, concentration is modeled as a linear combination of absorbances. The ILS model accounts for errors in the reference concentrations. ILS is a frequency-limited method and, therefore, is not capable of the precision improvements of CLS from signal averaging of multiple intensities. However, ILS can often be a useful method even if only one component is known for the calibration samples.

PLS and PCR are both factor-based methods that are capable of being full-spectrum methods. These methods have been explained and contrasted recently by Haaland and

0003-2700/90/0362-1091$02.50/0

Table I. Factor Levels

factor	label	low level	high level
concentration noise	(X_1)	none	σ_c = 0.02 mol fraction
spectral noise	(X_2)	σ_a = 0.001 AU	σ_a = 0.005 AU
separation of spectral features	(X_3)	25 cm^{-1}	12.5 cm^{-1}
spectral base-line variation	(X_4)	none	random linear
number of intensities per spectrum	(X_5)	200	25
calibration set configuration	(X_6)	designed	random
number of calibration samples	(X_7)	15	7
pure-component intensities for components A:B:C	(X_8)	1:0.1:0.1	1:1:1

Thomas (*6*). Like ILS, PLS and PCR can be employed even when only one component is known in the calibration samples. Both PLS and PCR methods factor the spectral data calibration matrix into the product of two smaller matrices. This amounts to a data compression step where the intensities at all frequencies used in the analysis are compressed to a small number of intensities in a new full-spectrum coordinate system. This new coordinate system is composed of loading vectors that can be used to represent the original spectral data. The intensities in the new full-spectrum coordinate system (called scores) are then used in a model where concentration is presumed to be a linear function of these intensities. Thus, PLS and PCR are methods that are concerned with modeling both spectra and concentrations during calibration. PCR performs the factoring of the spectral data matrix without using information about the concentrations. Therefore, there is no guarantee that the full-spectrum basis vectors that are associated with PCR are relevant for concentration prediction. PLS, on the other hand, performs the spectral factoring trying to account for the spectral variation while assuring that the new basis vectors relate to the calibration concentrations. Thus, PLS sacrifices some fit of the spectral data relative to PCR in order to achieve better correlations to concentrations during prediction.

This work introduces statistical experimental design based on computer simulations as a tractable and powerful approach to analyzing the performances of multivariate calibration methods over some user-specified range of conditions. By using the results presented here and applying them to cases where Beer's law is followed, it is possible to predict the effect on the performances of these methods induced by variations in the factors that were studied. It is also possible to identify those factors in the data that have different relative effects on the methods.

EXPERIMENTAL STRATEGY

Simulations and Experimental Design. The rationale underlying this work was to study a group of competing calibration methods under a variety of experimental conditions using simulated data for which the linear relationships between spectral absorbances and concentrations were maintained in the absence of introduced random errors. First, we identified a group of eight data set characteristics (or factors) that define a wealth of calibration situations and that might have some differential effects on the performances of the calibration methods considered. Two levels of each of the eight characteristics were studied. The particular experimental conditions studied were obtained by using a 2^8 factorial design (see ref 17). In this design, each level of each characteristic was studied with all possible combinations of the levels of the other characteristics. Thus, 2^8 = 256 separate experimental conditions were studied. The factors investigated and the two levels used for each factor in the simulations are presented in Table I. The two levels of each factor were selected somewhat arbitrarily but were adjusted after obtaining preliminary simulation results and after considering constraints on computation times.

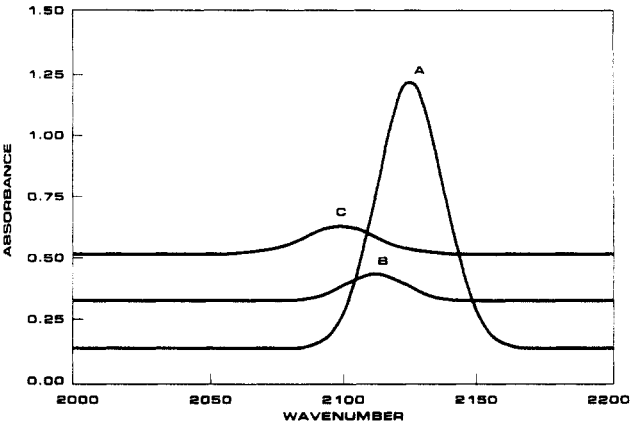

Figure 1. Pure-component spectra when intensities for components A:B:C are in the ratio 1:0.1:0.1 and band separation is at a minimum.

The particular system studied consisted of three chemical components whose mole fractions were constrained to sum to one. The pure-component spectra were each generated by using a single Gaussian spectral band for each chemical component with the full width of each band at half maximum being 30 cm^{-1}. The relative intensities of the pure-component spectra were either 1:1:1 or 1:0.1:0.1 with the major band(s) having an intensity of one absorbance unit (AU). Figure 1 illustrates the pure-component spectra of each of the three components for the case when components B and C have intensities that are one-tenth that of component A and when the peak separation is at a minimum. When this difference in relative spectral intensities exists, the effect of relative spectral intensity on performance could be investigated. In this case, we consider component A to be the major spectral component and components B and C to be the minor spectral components.

At each of the 256 experimental conditions, 25 calibration sets were generated by using different noise at the specified level. A calibration set consists of the reference concentrations and the artificially generated spectrum (based on the concentration information, the pure-component spectra, adherence to Beer's law, spectral noise levels, and the presence or absence of a random linear base line) for each of the calibration samples. Given the calibration data, each multivariate method is used to construct a separate calibration model for each of the three components based on concentrations and artificially generated spectra. These models are then used to estimate the composition of 50 new samples whose spectra have been artificially generated according to the same experimental conditions. The performance of each method for each component and for a particular experimental condition is based on how close (on average) the estimated concentrations are to the known concentrations for the 50 new prediction samples presented to each of the 25 independent calibration sets. The flow chart in Table II illustrates the complete simulation procedure.

The simulated data were such that when concentration noise (X_1) (i.e., concentration errors in the reference method) was at the "low level", no noise was added. When concentration noise was at the "high level", Gaussian noise was added with zero mean and a standard deviation (σ_c) of 0.02 mole fraction. Note that concentration noise was added without regard to the constraint that the sum of the mole fractions add to unity. When spectral noise (X_2) was at the low level, the noise added was Gaussian with zero mean and a standard deviation (σ_a) of 0.001 AU, while at the high level the spectral noise had σ_a = 0.005 AU. Noise was introduced by the IMSL library routine GGNPM (*18*). The overlap between spectral features (X_3) was decreased by a factor of 2 (peak-to-peak difference changed from 25 to 12.5 cm^{-1}) at the high level. This range of overlap was chosen to distinguish differences in the performance of the calibration methods when

Table II. Simulation Details

for L = 1 to 256 (256 = 2^8 experimental conditions)
  for K = 1 to 25 (25 is the number of calibration sets per condition)
  1. obtain concentrations for calibration set (configuration and size appropriate for the Lth condition)
  2. generate noise-free mixture spectra based on (1)
  3. add noise to mixture spectra
  4. add noise to concentrations (if appropriate)
  5. form calibration model for each method and component based on spectra from (3) and concentrations from (4)
  6. generate test set (50 samples)
    randomly select concentrations (see Appendix A)
    generate mixture spectra as in (2) and (3)
  7. estimate test set composition by using models from (5) and spectra from (6)
  END compute average performance measure (mean squared prediction error, MSPE) for each method for each component; total number of concentration estimates per component per method is 25 × 50 = 1250
END total number of concentration estimates per component per method is 256 × 25 × 50 = 320 000; total number of concentration estimates is 3 840 000

presented with difficult spectra. When a random linear base line was present (X_4), the spectral end points of the base line were randomly and independently selected from a uniform distribution between 0.1 and 0.3 AU [IMSL routine GGUBS (*18*)]. The base-line values at frequencies other than the end points were found by interpolating between the end-point values. In the other cases, the base line was constant at 0.1 AU. The number of uniformly spaced spectral intensities (X_5) was either 200 or 25. Rather than plotting spectra using numbers of data points as the abscissa, the spectra are displayed in all figures in constant energy units from 2000 to 2200 wavenumbers. The levels of the other two factors involve the number (X_6) and location (X_7) of calibration design points, in the compositional simplex, that make up the calibration set.

The distinction between low and high levels is that the high level is generally a more degrading condition (with respect to predicting concentrations) than the low level (although, in some cases, a factor did not affect the performance of a method). The case where all factors are at their low (less degrading) level will be considered our benchmark conditions with the minimum prediction errors. Our primary interest is to understand how much each of the calibration methods is affected by changing from a low to a high level for each of the factors. The identification of factors that affect methods differently (degrade performance more or less) and the absolute prediction abilities are key areas of concern. Knowledge of these would allow for reasonable assessment of how each competing method would perform in some particular situation of interest. From this knowledge, it should be possible to identify a method (or methods) that is best in that particular situation. The spectroscopist's experience and knowledge of each particular system should also aid in relating the results from these studies to practical problems.

Because the effects of the relative spectral intensity (X_8) increased the complexity of the simulation results, this factor is treated somewhat differently than the other factors. For example, when the relative intensities were at 1:0.1:0.1 for the three components, the abilities to predict the major and minor spectral components were quite different. In addition, when prediction errors are modeled (described later), the inclusion of factor X_8 causes complicated four-way interactions to become important. Therefore, to ease the interpretation of the results, most analyses and interpretations of the X_8 factor are restricted to the case where the relative intensities are 1:0.1:0.1. In this case, major and minor spectral component results are separated. However, some important and interesting results from the case where the relative intensities were 1:1:1 will be presented separately.

Figure 2 illustrates the spectra formed by using various combinations of the factors X_2, X_3, and X_4) that relate to the spectrum. These spectra were generated based on an assumed (1/3, 1/3, 1/3 composition) mixture with the relative intensities being 1:0.1:0.1 for the three components. Figure 2a represents the case where each of these three factors (X_2, X_3, X_4) is at the low level. Figure 2b–d represents cases where two of these three factors are at their low level and the third is at a high level. Figure 2e is a spectrum formed by setting all three of the spectral factors at high levels.

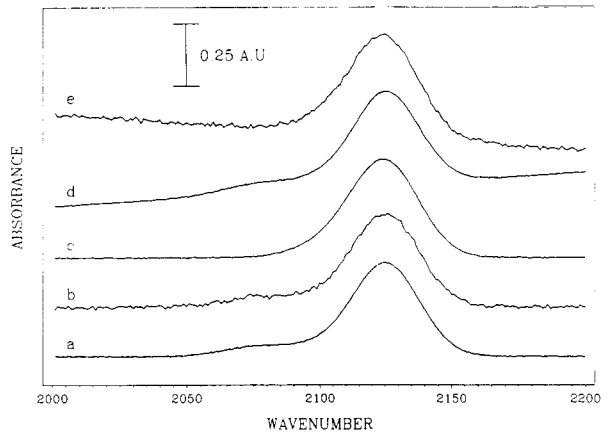

Figure 2. Representative simulated spectra when spectral intensities are 1:0.1:0.1 and concentrations are 1/3:1/3:1/3. (a) Spectral noise, spectral overlap, and random linear base line factors are at their least degrading conditions. (b) Spectral noise is at the high level. (c) Spectral overlap is at its more degrading condition. (d) Linear base line has been added. (e) Spectral noise, spectral overlap, and random linear base line factors are all at their more degrading condition.

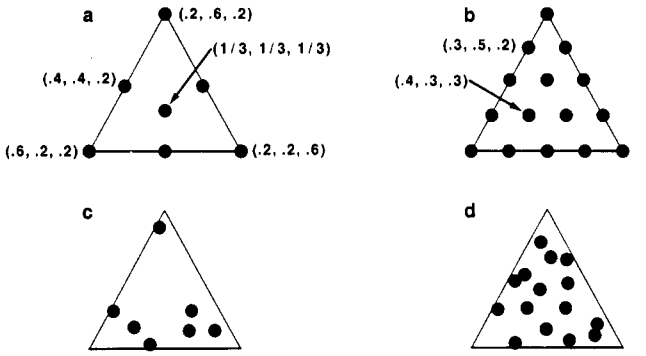

Figure 3. Calibration designs and configurations: (a) efficient mixture calibration design with seven calibration samples; (b) efficient mixture calibration design with 15 calibration samples; (c) random calibration with seven calibration samples; (d) random calibration with 15 calibration samples.

Figure 3 illustrates how the two factors, X_6 and X_7, affect the number and placement of calibration samples in the three-dimensional compositional space. Parts a and b of Figure 3 represent the two cases when the placement of the samples is governed by a fixed design. (For more information on experimental designs involving mixtures see ref 19.) Parts c and d of Figure 3 represent two calibration sets where the placement of the calibration samples was selected at random (see Appendix A). The concentrations of the 50 new prediction samples in each test set were determined by the same random selection procedure with a new set of 50 samples generated for each calibration set.

Multivariate Calibration Methods. The competing multivariate methods considered were two variants of CLS (denoted CLS-1 and CLS-3), PCR, PLS, and ILS. CLS-1 is the variation of CLS in which only the concentration of a single analyte of interest is assumed known during calibration. CLS-3 is the variation of CLS in which it is assumed that the concentrations of all three analytes are known. The CLS algorithm implemented here is from Haaland et al. (fit 2) where a linear base line is simultaneously fit over the spectral range being analyzed (*4*). As is traditional with CLS methods in spectroscopy, the spectral and concentration data were not centered. In general, the methods of Brown et al. (*12*) were implemented for ILS, and there was no centering of the data. PCR and PLS along with the two variants of CLS are full-spectrum methods. As such, they use all spectral information. The PLS and PCR algorithms used here were described by Haaland and Thomas (*6*). The PLS1 algorithm that performs the PLS analysis one component at a time was used in

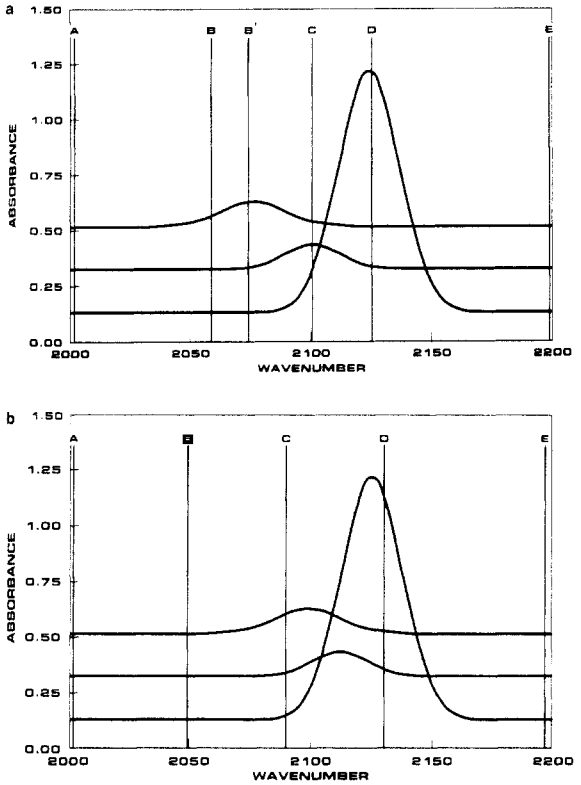

Figure 4. Frequencies selected for ILS when pure-component intensities for components A:B:C are 1:0.1:0.1. (a) Case where spectral overlap is low. Frequencies A, B, C, D, and E were selected when the spectral noise has σ_a = 0.001 AU. Frequencies A, B′, C, D, and E were selected when the spectral noise has σ_a = 0.005 AU. (b) Case where spectral overlap is high. Selected frequencies were independent of spectral noise level in this case.

analyzing these simulated data. The spectral and concentration data were centered but not scaled when PLS and PCR were used in these simulations.

Selection of Frequencies for ILS. Unlike the other methods, ILS uses only a subset of the available spectral frequencies for analysis. The reasons why ILS can only use a small number of frequencies for analysis are well documented (*5*). The number and selection of frequencies used for ILS depended on experimental conditions such as the overlap of spectral features and presence of a spectral base line. A separate series of simulations was performed to determine the optimal set of frequencies to use for ILS. Optimal here means the set of frequencies that results in the smallest mean squared prediction error (MSPE). The number of sources of spectral variation depend on the presence or absence of a random linear spectral base line. Using a limited number of simulations, we have found that the optimal number of frequencies used for ILS is three (no linear base line present) or five (linear base line present). The optimal set of frequencies depends also on the degree of spectral overlap of the three components and the magnitude of the spectral noise. Our assessment of the members of the optimal set of frequencies was accomplished by performing a limited all-sets search. It was limited in the sense that the only spectral frequencies considered were in the regions where signals were not negligible (i.e., frequencies at either end of the spectral range that involved only base-line information were not included in the search). These frequencies were selected by using data containing no base-line variations or concentration noise. While the optimal set of frequencies (for a specific set of experimental conditions) depended slightly on the particular component considered, very little precision was lost by assigning a single set of frequencies for each experimental condition, regardless of the component. Figure 4 indicates the optimal set of frequencies for each combination of factors that affect the frequency selections. Frequencies A and E are included when there is a linear base line present in the spectral data since these two frequencies clearly have the greatest sensitivity to the base-line variations.

In the case of low spectral overlap (Figure 4a), the set of optimal frequencies depends somewhat on the level of the spectral noise. When the spectral noise level is low, frequency B is selected. When spectral noise is high, the selected frequency is shifted to the B′ position in Figure 4a. Obviously the signal-to-noise ratio associated with a frequency and specificity of the signal for a given component at a candidate frequency are each important. The specificity is determined by the relative size of the signal associated with the component to be analyzed as compared to the total signal associated with all sources. In this case, there is trade-off between these two important factors.

In the case of relatively high spectral overlap, the noise level does not appear to greatly affect the frequency selection. Note, however, that frequency B is very far to the left in Figure 4b. This indicates that the lack of specificity is the limiting determinant with respect to prediction in this case of high overlap. As seen from this discussion and elsewhere (*20*), frequency selection is a difficult task. The intent here was to give ILS its best possible prediction performance by using the optimal set of frequencies for a set of experimental conditions. In practice, the inability to select the optimal set of frequencies for a set of real sample spectra will hamper the ILS predictions. Therefore, the results given here for ILS are generally better than those one would obtain in practice by using typical frequency selection methods. In fact, initial efforts using stepwise multiple linear regression (*21*) to select frequencies yielded significantly poorer prediction performance than that demonstrated in this paper using the limited all-sets search selection.

Optimal Number of Loading Vectors for PLS and PCR. Peculiar to both PLS and PCR is the selection of the optimal-sized model or number of loading vectors (full-spectrum basis vectors) that are required to obtain minimum prediction errors in unknown samples (*6*, *22*). The optimal number of loading vectors depends on the number of independently varying chemical species present as well as the presence of other sources of systematic spectral variation such as the presence of randomly varying base lines (*6*). The optimal number of loading vectors could also be weakly dependent on spectral and concentration noise levels, but these factors were not found to be important for selecting the number of loading vectors in these simulations. In this experiment, the number of independently varying components is two (not three because of the constraint that the mole fractions must sum to one and because the concentration and spectral data were centered when using the PLS and PCR algorithms employed here). The optimal number of loading vectors for PLS and PCR is equal to the two required for modeling spectral variations due to the chemical components and two more (slope and intercept of the base line) when a linear, frequency-dependent base line is present.

Performance Measure. The mean squared prediction error (MSPE) is used as the basis for the measure of performance. Formally

$$\mathrm{MSPE}_{ilm} = \frac{1}{25 \cdot 50} \sum_{k=1}^{25} \sum_{j=1}^{50} (\hat{c}_{ijklm} - c_{ijkl})^2$$

where $\hat{c}_{ijklm}$ (c_{ijkl}) is the predicted (actual) concentration obtained by using the ith component of the jth test sample within the kth calibration set with the lth experimental condition using the mth calibration method.

Note that the MSPE is the squared prediction error averaged over the bounded compositional region illustrated in Figure 3. The MSPE for a given set of conditions can be decomposed into two elements by

$$\mathrm{MSPE} = \mathrm{biâs}^2 + \mathrm{variâncе}$$

The estimated bias represents that part of the MSPE that is due to consistent under- or overestimation. The estimated variance represents the average squared variation of the predicted concentration about the sum composed of the true concentration and the estimated bias. Over the range of this simulation, the bias was not significant. Typically, the contribution of the squared bias was well under 1%. The squared bias accounted for at most about 10% of the MSPE. The MSPE's varied from 10^{-8} to 10^{-1} depending on the calibration method, component, and simulation condition. In addition to the MSPE, the standard error of the MSPE was estimated for each calibration method, component,

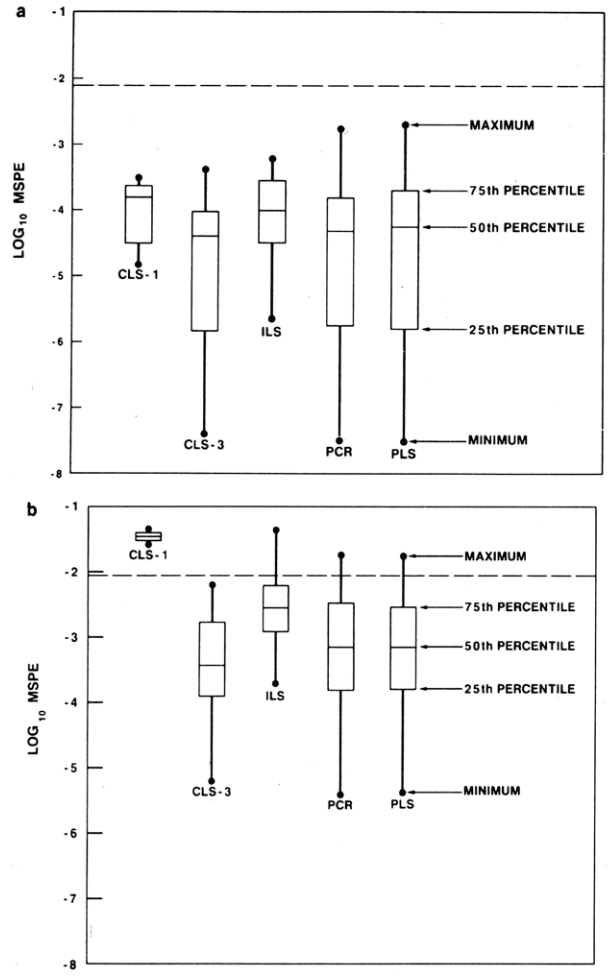

Figure 5. Distribution of log (MSPE) results for each of the various multivariate calibration methods for (a) the major spectral component A and (b) the minor spectral component C.

and simulation condition in order to determine the statistical significance of the results.

To assess the effects of the experimental factors on each method as well as compare methods, it was necessary to transform the MSPE's such that the standard errors of the transformed MSPE's are of comparable size. Therefore, Y = log (MSPE) was used as the performance measure for analytical purposes.

RESULTS AND DISCUSSION

Overall Performance Comparisons. The focus of this analysis is to describe how each of the experimental factors affects each of the calibration methods. First, however, it is informative to consider the range of performance exhibited by each method over all experimental conditions. For this purpose, we present two cases for study. Each involves the condition where the spectral intensities of the pure components are in the 1:0.1:0.1 ratio for the three components, i.e., X_8 is fixed with all other factors allowed to vary. The first case involves the major spectral component A, while the second involves the minor spectral component C.

Figure 5 indicates the relative performances and the range and distribution of performances exhibited by the methods under consideration over all 128 conditions (i.e., the condition where all three pure components have equal intensity has been excluded from consideration). Note that better prediction precision is indicated when log (MSPE) is more negative. Prominent in these figures is a horizontal line at log (MSPE) $\simeq -2.05$ which represents the value of log (MSPE) that would result if the average component concentration (1/3) was used for all predictions. This is used as a benchmark to assess the predictive relevance of a method at any particular condition

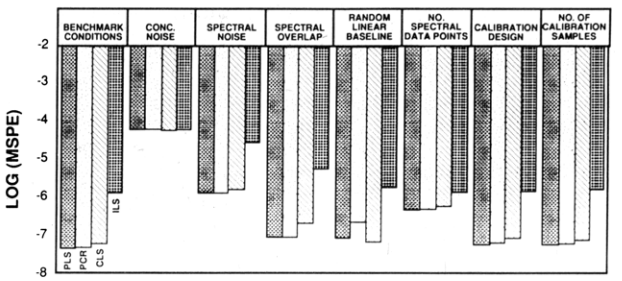

Figure 6. log (MSPE) for each of the multivariate calibration methods for the major spectral component (A) when all factors are at their benchmark (least degrading) conditions and when each factor is separately set to its more degrading condition.

(see Appendix A). If for a given method, some of the distribution of log (MSPE) is more positive than −2.05, then the method has no predictive relevance at least for some conditions. Therefore, as illustrated in Figure 5b, CLS-1 has no predictive relevance for the minor spectral component. In fact, no further presentation of CLS-1 performance will be made, as it is clearly not a competitive method under the conditions presented here. Note that in the case of real spectra with multiple spectral features CLS can be a viable option even when all components are not included in the calibration if the procedures developed by Haaland et al. (*4*) are adopted.

It is clear from Figure 5 that, for each of the given methods, there is a considerable range in performance over the conditions studied. Comparison between parts a and b of Figure 5 also illustrates the important effect that spectral intensity has on each of the methods (i.e., prediction of the major component is superior to that of the minor component). The results for the minor spectral component B are similar to those found for minor spectral component C with component B generally being predicted more poorly than component C because of its greater degree of overlap with component A. Because the results for component B do not add additional insights relative to the results of component C, the remainder of the paper will involve only discussions of the results for components A and C.

Effects of Specific Factors on Prediction Performance. Figure 6 indicates the relative magnitude of the log (MSPE) values for the major spectral component for each calibration method when all factors are at their least degrading (benchmark) condition and when each of the main factors has separately been switched to its more degrading condition. Again, the more negative the bar is in Figure 6, the better the predictive ability of the multivariate method.

An initial observation that can be made from Figure 6 is that, for these cases, the full-spectrum methods are never worse and usually significantly better than the frequency-limited ILS method. This observation also holds true for all cases investigated involving either the major or minor spectral components. Superior performance for full-spectrum methods might be expected since they allow signal averaging over many spectral frequencies, whereas ILS is not capable of signal averaging in spectral space. Given the range of changes in the factors selected for this study, concentration noise (i.e., errors in the reference concentrations) degrades performance more than any of the other factors in the case of the major spectral component. At the noise levels used in these simulations, the effect of adding concentration noise dominates the effect of spectral noise to such a degree that even the spectral intensity averaging of the full-spectrum methods is not sufficient to overcome the degrading effects of high concentration noise. This is an important observation since it implies that no method is able to completely overcome the presence of poor precision in the reference methods used for concentration determinations in real data sets. This clearly

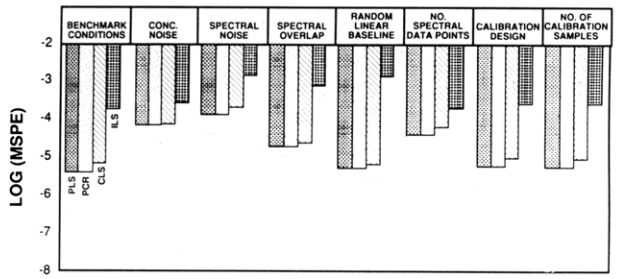

Figure 7. log (MSPE) for each of the multivariate calibration methods for the minor spectral component (C) when all factors are at their benchmark (least degrading) conditions and when each factor is separately set to its more degrading condition.

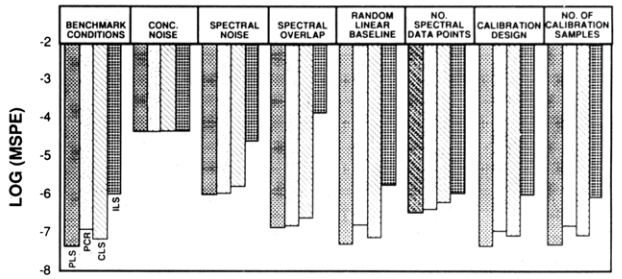

Figure 8. log (MSPE) for each of the multivariate calibration methods for component (A) when all pure-component spectra have equal intensities. Presented is the case when all factors are at their benchmark (least degrading) conditions and when each factor is separately set to its more degrading condition.

puts high priority on the need to optimize the accuracy and precision of our reference methods in order to take full advantage of the better performance possible with the full-spectrum multivariate methods relative to frequency-limited methods. As demonstrated in Figure 6, large errors in the reference method have the effect of leveling the predictive performance of all multivariate methods for the major varying spectral components.

Increasing spectral noise yields the second most important main effect among the factors given the ranges of the factors studied. All four multivariate methods are degraded by nearly the same amount when only spectral noise is increased. When spectral overlap is separately increased, CLS and ILS are found to be degraded more than PLS or PCR, which are degraded by comparable amounts. When random linear base lines are present, CLS is not degraded at all since the CLS method used in the simulations included a simultaneous and explicit fit of a linear base line to the spectral data. This superior result for CLS illustrates the advantage of using an explicit model that is correct. It is also interesting to note that the performance of PCR is degraded more than PLS by the presence of random linear base lines. This might be expected since PLS was designed to put more predictive concentration information in the early loading vectors. Therefore, the first four PLS loading vectors are less influenced by the presence of linear base lines, and better predictions are found for PLS relative to PCR under these conditions. This is the case in Figure 6 where the difference between PLS and PCR is largest and has the highest level of statistical significance.

Decreasing the number of spectral data points has no effect on the ILS results in these simulations since the selected frequencies are limited to either three or five for ILS. The full-spectrum methods are degraded since there is less signal averaging across intensities when one-eighth the number of spectral intensities are included in the analysis. The predictive abilities of the full-spectrum methods decrease and converge to that of ILS as the number of frequencies used in the full-spectrum analyses approach the number of factors required for optimal predictions.

Finally, the change to a random calibration design and a decrease in the number of calibration samples each have the smallest effect on the predictive results for these simulations. However, they often have a greater effect when both these degrading conditions are present or when they are present in combination with other factors. A discussion of these and other interaction effects between factors is presented as supplementary material. See end of paper for ordering information.

The individual effects of each factor relative to the minor spectral component C are shown in Figure 7. The order of magnitude decrease in spectral intensity of this component relative to component A causes the MSPE to be degraded by nearly 2 orders of magnitude for the benchmark conditions for each method relative to that of component A. ILS is degraded more than the full-spectrum methods in going from the major to minor spectral component. The effect of varying the level of spectral noise from low to high causes the greatest degradation in predictive ability for the minor component. With spectral signal-to-noise ratios degraded by a factor of 10 for minor spectral component C relative to that of component A, spectral noise now has a greater effect on predictive ability than the level of concentration noise chosen.

In contrast to the results found for the major spectral component, the concentration noise level is such that the signal averaging over intensities obtained by the full-spectrum methods is sufficient to make them outperform ILS for the minor spectral components. In this case of the minor spectral components, larger reference concentration errors would be required to equalize the prediction abilities of all calibration methods as had been found with the major spectral component. The other conclusions for the major component are also found to be valid for the minor component except that PLS and PCR are comparable even in the presence of random linear base lines. In these simulations, the range of spectral variation for the minor spectral components only amounted to 0.04 AU, while the base-line variations were as much as 0.2 AU. Apparently, when the magnitude of the random base-line variations is significantly greater than that of the spectral variations caused by the analyte, PLS loses its advantage over PCR.

Figure 8 illustrates the results for component A when all three pure-component spectra have identical peak intensities. Comparisons between predictive ability of methods within Figure 8 and between Figures 6 and 8 provide for some interesting conclusions. First, the performance of PLS improves while that of PCR is degraded under the benchmark conditions relative to those for Figure 6. This is qualitatively comparable to the relative performance of PLS and PCR in the presence of a random linear base line as shown in Figure 6. The linear base line is comparable to the presence of two additional random spectral components (offset and slope) while the presence of three spectral components of comparable magnitude means that there is a major source of systematic spectral variation unrelated to the component of interest. PLS again performs better than PCR because it gives less attention to the unrelated sources of systematic spectral variations. Thus, PLS achieves better predictions in the presence of overlapping but independently varying spectral features. This paper presents the first clear evidence of cases where PLS achieves better predictions than PCR that are statistically significant. In Figure 8, statistically significant improvements in the performance of PLS over PCR remain in the presence of the degrading effects of several of the other factors as well.

A final observation with respect to Figure 8 is the large relative degradation of ILS when spectral overlap of the three bands is increased. This may be due to an inability of ILS

to handle severe overlap of major spectral features or may simply be a result of the fact that the spectral frequencies of ILS were not reoptimized for the case where all three pure components had the same intensity.

Modeling of Prediction Errors. In order to understand the degree to which the experimental factors jointly affect each calibration method, Y = log (MSPE) was modeled for each calibration method and spectral component combination. It was found that for each method–component combination, a third-order interaction model represented the observed data to within the estimated error of the MSPE

$$Y = \mu + \sum_{i=1}^{7} \alpha_i I_i + \sum_{i \neq j} \sum \beta_{ij} I_{ij} + \sum \sum_{\substack{i \neq j \\ i \neq k \\ j \neq k}} \sum \gamma_{ijk} I_{ijk} + \epsilon \quad (1)$$

where μ = log (MSPE) with all factors at their respective low levels

$$I_i = \begin{cases} 1 \text{ if the } i\text{th factors are at the high level;} \\ 0 \text{ otherwise} \end{cases}$$

$$I_{ij} = \begin{cases} 1 \text{ if both the } i\text{th and } j\text{th factor are at their high levels;} \\ 0 \text{ otherwise} \end{cases}$$

$$I_{ijk} = \begin{cases} 1 \text{ if the } i\text{th, } j\text{th, and } k\text{th factors are all at their high levels;} \\ 0 \text{ otherwise} \end{cases}$$

ϵ = model error

The change in performance by changing the ith factor from low to high while all of the other factors are at the low level is represented by α_i. These effects can be inferred from Figures 6, 7, and 8. The change in performance by changing both the ith and jth factor from low to high while all of the other factors are low is β_{ij}. Finally, the change in performance by changing the ith, jth, and kth factor from low to high while all of the other factors are low is γ_{ijk}.

Greater detail concerning the above modeling is available as supplementary material. Tables III and IV in the supplementary material provide estimates of μ and the α_i as well as the significant second- and third-order parameters (β_{ij} and γ_{ijk}) for each model. From the parameter estimates given in these tables, it is possible to estimate the performance of a particular method (for a given component) at a fixed set of conditions. The estimated performance is simply a linear combination of the appropriate parameters. The advantage of performing this modeling is that the reader can reasonably estimate the MSPE for any particular set of factor levels using the results of Tables III and IV. Thus, it is not necessary to present tables with results from all 128 conditions listed separately for the major and minor spectral components.

One can compare the parameter estimates between methods in Tables III and IV to assess if (and how much) the methods are differentially affected by the experimental factors. However, a more sensitive way to make comparisons between calibration methods is by using the set of differences

$$D_{12} = Y(\text{method 1}) - Y(\text{method 2})$$

obtained by applying the different methods to identical calibration and prediction data for the range of conditions. These differences between the two methods can be modeled with respect to each of the factors as in eq 1

$$D_{12} = \mu + \sum_{i=1}^{7} \alpha_i I_i + \sum_{i \neq j} \sum \beta_{ij} I_{ij} + \sum \sum_{\substack{i \neq j \\ i \neq k \\ j \neq k}} \sum \gamma_{ijk} I_{ijk} + \epsilon \quad (2)$$

Tables V and VI in the supplementary material provide the estimates of the first-order parameters as well as the significant second- and third-order parameters for three combinations of methods ({PLS, PCR}, {PLS, CLS}, {PCR, CLS}) for the case of the major spectral component A (Table V) and the case of the minor spectral component C (Table VI). Further discussions of the results presented in these tables are also presented in the supplementary material. From the results given in Tables V and VI, it is possible to make direct comparisons of the performance of the three full-spectrum methods (CLS, PLS, and PCR). Since ILS was such a poor performer and therefore less affected by degrading conditions, it was not included in these pairwise comparisons.

An important point that can be made with regard to the information contained in Tables V and VI in the supplementary material is that sets of conditions can be found where a given full-spectrum method has a better prediction ability relative to another. On one extreme for the major component, the MSPE for PCR is predicted and observed to be 6.2 times larger than that found for CLS and 1.8 times larger than that for PLS (X_4 and X_7 high with all others low). Conversely, for another set of conditions (X_3, X_6, and X_7 high), the MSPE for PLS for the major component was predicted from the model in eq 1 to be 3.5 times smaller than that found for CLS. (In this latter case, the improvement in MSPE for PLS relative to CLS obtained from the actual simulation was a factor of 5 rather than the model estimate of 3.5. This case demonstrates an extreme difference between the observed values of MSPE and those estimated by using the parameter estimates in the tables.) The parameter estimates in Tables V and VI could be used by the reader to estimate, within the precision of the model, the difference in performances between any two of the full-spectrum methods for any specific set of conditions that were tested here.

Additional Comparisons. The performances of the calibration methods were evaluated at conditions specified by somewhat arbitrary levels of the experimental factors (i.e., spectral noise, σ_a, of 0.001 or 0.005 AU). In some cases, it is possible to interpolate performance at factor levels not considered in the experiment. For instance, for a given method, additional simulations confirmed that in the absence of both concentration noise and a random linear base line, log (MSPE) $\propto$ log (σ_a^2), when all other factors are held constant. The performances of the full-spectrum methods improved as the number of analytical frequencies, n, increased. For either CLS, PCR, or PLS, when concentration noise and a linear base line are absent and when all other factors are held constant, additional simulations confirmed that log (MSPE) $\propto$ log ($1/n$) as would be expected. Note, however, that one underlying assumption is that the spectral errors are independent at adjoining frequencies which may not always hold for spectra obtained from real samples.

It is important to note that some of the differences between PLS (PCR) and CLS are a result of differences in the implementation of the algorithms used by each method. The algorithms used here are the same as they are generally implemented in quantitative infrared applications. Thus, the spectral and concentration data were centered for PLS and PCR but not centered for CLS and ILS. Note that CLS requires that all components in the calibration samples be known. Therefore, when the data are centered, only two components can be directly analyzed in this constrained system (the matrix to be inverted is singular if all three are included in the analysis). The third component must be determined by difference. Additional simulations confirmed that centering could improve the performance of CLS for the two directly analyzed components while the third component was predicted with poorer precision. The improved performance for the centered CLS algorithm with the directly analyzed components makes it comparable to PLS and PCR

in predictive ability for the cases where concentration noise is absent.

Finally, additional comparisons between calibration methods can be made that are not available from the data presented in either the figures or the tables. For example, in the 128 cases examined for the major spectral component, PLS is significantly better than PCR 50 times (using a 2 standard deviation criterion) while PCR is significantly better than PLS 56 times. In these cases where PCR outperforms PLS, most involve the presence of concentration noise. The analysis of variance (results in Tables III and V) was not sufficiently sensitive to detect this small but real difference. Thus, there is an indication that the presence of concentration noise can cause PCR to slightly outperform PLS, but the magnitude of this better performance is quite small at the 5% relative concentration error level (i.e., $\sigma_c = 0.02$ mole fraction). The presence of greater amounts of concentration noise might make this effect more pronounced. Relatively large concentration errors might be expected to affect PLS more than PCR since concentrations enter the PLS algorithm in both the spectral decomposition and regression steps while concentration is only used in the regression step for PCR. For the minor spectral component, PLS performs significantly better than PCR 42 times while PCR is significantly better than PLS only twice. Therefore, although the difference in performances is relatively small between PLS and PCR (Table VI), PLS appears to have a slight advantage for minor spectral components.

CONCLUSIONS

A conceptually simple procedure for studying the performances of various competing multivariate calibration methods has been introduced. This procedure, which links classical statistical experimental design concepts with computer simulation, has proven to be a useful tool for understanding the performance attributes of competing multivariate calibration methods for the quantitative analysis of spectral data.

In this particular study, it was found that the full-spectrum calibration methods (CLS, PCR, and PLS) outperformed the ILS method over a wide range of experimental conditions. Therefore, we infer that when Beer's law holds, ILS can be competitive only when the concentrations of the calibration set are contaminated with reasonably large errors. Among the full-spectrum methods, PCR and PLS are very similar. The major difference between these two methods is that PLS seems to predict better than PCR in the cases when there are random linear base lines or independently varying major spectral components which overlap with the spectral features of the analyte. This is not surprising when one considers the different ways in which the two methods decompose (or factor) the spectral matrix. The PCR decomposition is based entirely on spectral variation without regard to the component concentrations while the PLS decomposition is dependent on the component concentrations. Because the spectral variations caused by the presence of a random linear base line or major spectral components can be reasonably large, the PCR decomposition is significantly influenced by variations which have no relevance to the analyte concentrations. Therefore, PCR is not able to predict as well as PLS in these situations.

The differences between (PLS, PCR) and CLS are more numerous and complex but are also affected by the algorithms used in the analysis. Over the set of conditions studied, the relative performances of (PLS, PCR) and CLS varied considerably. For example, in two extreme cases, MSPE of PLS was about 3.5 times larger or about 5 times smaller than that of CLS.

As we have demonstrated, the optimal choice of calibration method depends on the particular experimental conditions. However, PLS seems to be a reasonable choice over a wide range of conditions. PLS is the optimal performer or is close to optimal over the wide range of conditions considered. Unlike CLS, all overlapping spectral components do not have to be known, nor does the spectral base line have to be explicitly modeled. It seems that the only inherent dangers of using PLS (or PCR) result from over- or underfitting by an inappropriate number of factors. This has not been a problem here because the number of sources of spectral variation was known for these simulations.

It is important to note that factors other than prediction ability often need to be considered when choosing a calibration method. For example, the estimated pure-component spectra generated by CLS contain significant qualitative information (*4*) which is often useful to the spectroscopist. PLS can also yield qualitative information of better quality than generally possible with PCR but of poorer quality than CLS (*6*). The full-spectrum methods are greatly superior to ILS in their ability to detect spectral outlier samples among calibration and unknown samples since full-spectrum spectral residuals yield far more detailed information about the quality of the spectral fits. These spectral residuals can also often be interpreted by the spectroscopist.

The conclusions about relative predictive performance based on this study should be relevant to a larger number of spectral applications. However, we are sure that there are factors, which were not studied here, that can affect performance. Similar studies, incorporating additional factors, could be undertaken to assess the relative effects of new factors on the performance of the various methods. For example, if the Beer's law model were valid for only a few spectral intensities, it is conceivable that ILS could outperform the full-spectrum methods since the full-spectrum methods would then include significant additional spectral intensities that do not follow Beer's law. We are in the process of extending these simulation studies to cases where there are significant deviations from the Beer's law model. These results should yield additional insight to the capabilities and performance of the various multivariate calibration methods applied to real spectral data.

APPENDIX A. RANDOM PLACEMENT OF CALIBRATION SAMPLES

The compositional space that is considered throughout this paper is a three-dimensional simplex. This space can be represented by a modified ternary diagram (see e.g. Figure 3). The modification is that the minimum and maximum values associated with each component are 0.2 and 0.6, rather than the normal extremes which are 0 and 1. Note that the sum of coordinates ($c_1 + c_2 + c_3$), representing any point in this space, is one. This reflects the assumed condition whereby the sum of the mole fractions of the three components is one. Random sampling (selection of c_1, c_2, and c_3) within the compositional space was done by assuming that the likelihood of sampling within any region of this simplex was proportional to the area in the modified ternary diagram associated with that region.

More formally, this sampling was performed in accordance with a special version of the Dirichlet distribution (*23*). The density function of the Dirichlet distribution is usually denoted by

$$f(c_1, c_2, c_3) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} c_1^{\alpha_1 - 1} c_2^{\alpha_2 - 1} (1 - c_1 - c_2)^{\alpha_3 - 1}$$

where $0 \leq c_1, c_2, c_3 < 1$, $c_1 + c_2 + c_3 = 1$, and Γ is the gamma function. In the special case considered here, $\alpha_1 = \alpha_2 = \alpha_3 = 1$. This special case of the Dirichlet distribution is relative to the unmodified ternary diagram. With the appropriate change of coordinates, it is easy to change to the modified

ternary diagram that we consider. In this case

$$E[c_i] = 1/3 \quad i = 1, 2, 3$$

$$\mathrm{Var}[c_i] = \frac{2}{36}\cdot 0.4^2 \quad i = 1, 2, 3$$

$$\mathrm{Cov}[c_i, c_j] = -\frac{1}{36}\cdot 0.4^2 \quad i \neq j, \quad i, j = 1, 2, 3$$

In the analysis of performances obtained through simulations, $\mathrm{Var}[c_i]$ is an important benchmark that is used to assess the degree of predictive relevance for a given method and analyte in a particular situation. If the observed MSPE is larger than $\mathrm{Var}[c_i]$, then this is numerical evidence that the method has no predictive relevance for that particular situation. This is true because, in this situation, an analyst who always predicted c_i with the value $1/3$ instead of using the calibration would, on average, have a smaller MSPE than the observed MSPE obtained by using the calibration method.

The following procedure was used to generate a random compositional point (c_1, c_2, c_3).

Step 1. Generate two independent variables, U_1 and U_2, from the uniform distribution on the interval from 0 to 1. The uniform random numbers were generated by the linear congruential random number generator, GGUBS from the IMSL library (*18*).

Step 2. Form $U_{(1)}$ and $U_{(2)}$, where

$$U_{(1)} = \min\{U_1, U_2\}, \text{ and}$$

$$U_{(2)} = \max\{U_1, U_2\}$$

Step 3. Form the triplet (D_1, D_2, D_3), where

$$D_1 = U_{(1)},$$

$$D_2 = U_{(2)} - U_{(1)}, \text{ and}$$

$$D_3 = 1 - U_{(2)}.$$

Step 4. Form the triplet (c_1, c_2, c_3), where

$$c_1 = 0.2 + 0.4D_1,$$

$$c_2 = 0.2 + 0.4D_2,$$

$$c_3 = 0.2 + 0.4D_3$$

Supplementary Material Available: Detail concerning modeling of prediction errors and Tables III–VI, estimates of μ and method comparisons of first-, second-, and third-order parameters (9 pages). Photocopies of the supplementary material from this paper or microfiche (105 × 148 mm, 24× reduction, negatives) may be obtained from Microforms & Back Issues Office, American Chemical Society, 1155 16th Street, NW, Washington, DC 20036. Orders must state whether for photocopy or microfiche and give complete title of article, names of authors, journal issue date, and page numbers. Prepayment, check or money order for $19.00 for photocopy ($21.00 foreign) or $10.00 for microfiche ($11.00 foreign), is required and prices are subject to change.

LITERATURE CITED

(1) Lindberg, W.; Persson, J.-A.; Wold, S. *Anal. Chem.* **1983**, *55*, 643.
(2) Fredericks, P. M.; Lee, J. B.; Osborn, P. R.; Swinkels, D. A. *J. Appl. Spectrosc.* **1985**, *39*, 303 and 311.
(3) Fuller, M. P.; Ritter, G. L.; Draper, C. S. *Appl. Spectrosc.* **1988**, *42*, 217 and 228.
(4) Haaland, D. M.; Easterling, R. G.; Vopicka, D. A. *Appl. Spectrosc.* **1985**, *39*, 73.
(5) Haaland, D. M. *Spectroscopy* **1987**, *2* (6), 56.
(6) Haaland, D. M.; Thomas, E. V. *Anal. Chem.* **1988**, *60*, 1193 and 1202.
(7) Haaland, D. M. *Anal. Chem.* **1988**, *60*, 1208.
(8) Cahn, F.; Compton, S. *Appl. Spectrosc.* **1988**, *42*, 865.
(9) Naes, T.; Martens, H. *Commun. Statist.-Simula. Computa.* **1985**, *14*, 545.
(10) Naes, T.; Martens, H. *Appl. Statist.* **1986**, *35*, 195.
(11) Nyden, M. R.; Forney, G. P.; Chittur, K. *Appl. Spectrosc.* **1988**, *42*, 588.
(12) Brown, C. W.; Lynch, P. F.; Obremski, R. J.; Lavery, D. S. *Anal. Chem.* **1982**, *54*, 1472.
(13) Brown, C. W. *Spectroscopy* **1986**, *1* (4), 32.
(14) Donahue, S. M.; Brown, C. W.; Obremski, R. J. *Appl. Spectrosc.* **1988**, *42*, 353.
(15) Donahue, S. M.; Brown, C. W.; Caputo, B.; Modell, M. D. *Anal. Chem.* **1988**, *60*, 1873.
(16) Seasholtz, M. B.; Archibald, D. D.; Lorber, A.; Kowalski, B. R. *Appl. Spectrosc.* **1989**, *43*, 1067.
(17) Cochran, W. G.; Cox, G. M. *Experimental Designs*, 2nd ed.; Wiley: New York, 1957.
(18) International Mathematical and Statistical Libraries, Inc. (IMSL), Users Manual, Edition 9.2; IMSL: Houston, TX, 1984.
(19) Cornell, J. A. *Experiments with Mixtures*; Wiley: New York, 1981.
(20) Mark, H. *Appl. Spectrosc.* **1988**, *42*, 1427.
(21) Draper, N. R.; Smith, H. *Applied Regression Analysis*, 2nd ed.; Wiley: New York, 1981.
(22) Osten, D. W. *J. Chemom.* **1987**, *2*, 39.
(23) Johnson, N. L.; Kotz, S. *Continuous Multivariate Distributions*; Wiley: New York, 1972.

RECEIVED for review September 11, 1989. Revised manuscript received December 27, 1989. Accepted February 20, 1990. This work performed at Sandia National Laboratories supported by the U.S. Department of Energy under Contract No. DE-AC04-76DP00789. Portions of this work were presented at the 1987 Eastern Analytical Symposium, New York, September 12–16, 1987, Paper 132, and at the 1988 meeting of the Federation of Analytical Chemistry and Spectroscopy Societies, Boston, October 30–November 4, 1988, Paper K05.

CORRESPONDENCE

Determination of Bonded Phase Thickness in Liquid Chromatography by Small Angle Neutron Scattering

Sir: The characterization of physical and chemical properties of chemically modified surfaces is of considerable importance for an improved understanding of interfacial phenomenon in such fields as catalysis, electrochemistry, and chromatography. Alkylated substrates prepared by the reaction of chlorosilanes with porous silica are widely used in reversed-phase liquid chromatography for the separation of polar and nonpolar analytes. Such bonded phase layers have unique properties that differ from solid and liquid states of matter. By nature of the covalent bond, the degrees of freedom of immobilized chains are reduced, and chain motion is intermediate to that in corresponding alkyl liquids and solids. Understanding the physical nature of the bonded phase is requisite to completely describing solute retention mechanisms. The direct determination of bonded phase morphology represents a difficult analytical problem. Properties such as