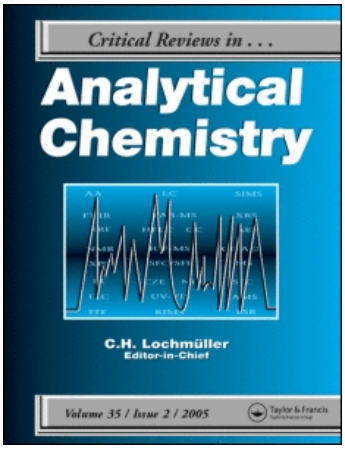

Critical Reviews in Analytical Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713400837

Advances and Perspectives in Near-Infrared Spectrophotometry

James K. Drennen[a]; Elizabeth G. Kraemer[b]; Robert A. Lodder[b]
[a] Department of Pharmaceutical Chemistry and Pharmaceutics, School of Pharmacy, Duquesne University, Pittsburgh, PA [b] Division of Medicinal Chemistry and Pharmaceutics, College of Pharmacy, University of Kentucky, Lexington, KY

To cite this Article Drennen, James K. , Kraemer, Elizabeth G. and Lodder, Robert A.(1991) 'Advances and Perspectives in Near-Infrared Spectrophotometry', Critical Reviews in Analytical Chemistry, 22: 6, 443 — 475
To link to this Article: DOI: 10.1080/10408349108051642
URL: http://dx.doi.org/10.1080/10408349108051642

Critical Reviews in Analytical Chemistry, 22(6):443–475 (1991)

Advances and Perspectives in Near-Infrared Spectrophotometry

James K. Drennen

Department of Pharmaceutical Chemistry and Pharmaceutics, School of Pharmacy, Duquesne University, Pittsburgh, PA 15282

Elizabeth G. Kraemer and Robert A. Lodder

Division of Medicinal Chemistry and Pharmaceutics, College of Pharmacy, University of Kentucky, Lexington, KY 40536-0082

ABSTRACT: Near-infrared spectrophotometric analysis is a rapid technique that typically uses the reflectance of a solid sample at several wavelengths to determine the sample's composition. A computerized modeling process is generally used to correct for background and sample-matrix interferences. The modeling process employs a training set of samples to, in effect, "teach" the computer to recognize relationships between minute spectral features and sample composition. The contents of the training-set samples must be determined initially by some other reference method before applying the near-IR technique. The model developed from near-IR spectra and reference values gives the sample composition using a number of linear equations. Each of these equations expresses a particular component concentration as a weighted sum of the signals observed at a number of near-IR wavelengths. Instruments used for near-IR spectrophotometry can be as simple as a filter photometer or a grating monochromator. The broad spectral peaks and highly correlated wavelength vectors generally limit the number of wavelengths used in the model. Little or no sample preparation is required by near-IR methods, and many solid samples can be directly analyzed. Near-IR spectrophotometry has found application in agriculture, industry, biology, medicine, and even satellite remote sensing.

KEY WORDS: chemometrics, pharmaceutics, in-line, on-line, at-line.

I. INTRODUCTION

The field of near-infrared (IR) spectroscopy is growing rapidly and the literature regularly reveals new discoveries from industrial and academic laboratories. Modern near-IR instrumentation is proving to be both more accurate and easier to use than more conventional analytical methods in many applications.

Near-IR spectroscopy has numerous advantages over traditional methods of analysis. These advantages include:

1. Virtually all organic compounds absorb in the near-IR region, but none absorb strongly. (This feature gives near-IR light good sample penetration even in the aqueous samples that thwart conventional IR analyses, and makes the Beer-Lambert Law useful in the near-IR region.)
2. The high-throughput optics, efficient detectors, bright sources, and wide slits of near-IR monochromators provide the superior signal-to-noise (S/N) ratio needed for effective use of near-IR spectra; S/N ratios in

1040-8347/91/$.50

excess of 100,000:1 are achievable with modern near-IR instrumentation, and analytes have been determined in intact samples at levels in the 10 to 100 ppm region.[1]

3. Near-IR instrumentation is simple and relatively inexpensive. (Glass optics, grating spectrometers, and even filter photometers generally suffice for analyses.)
4. Efficient operation in transmission, transflectance, interactance, and diffuse-reflectance modes make sample preparation steps generally unnecessary. (The absence of sample preparation requirements, in combination with the low energy of near-IR photons [approximately 1 eV], makes most near-IR analyses noninvasive and nondestructive.)
5. Very rapid analyses (less than 10 s) are possible with computerized near-IR techniques.

Although the sensitivity of near-IR instruments is increasing and several references report the analysis of extremely low levels of analytes, the method is most often used for samples whose analytes are present at concentrations of 1% or greater. The spectroscopist can determine constituents in concentrations of 0.01 to 1% when ideal conditions exist.

Some of the earliest applications of near-IR spectrometry involved the determination of protein and moisture content of grain samples. Applications for near-IR spectroscopy have now been found in the pharmaceutical, cosmetic, tobacco, food, chemical, polymer, textile, paint, coal, and petroleum industries. Examples of many of these applications are discussed in the following pages.

The low molar absorptivities of most materials in the near-IR region, when combined with the fact that nearly all organic compounds show near-IR absorption bands, make the technique useful for clinical analysis. Light of other wavelengths does not penetrate the body sufficiently to permit *in vivo* testing; however, noninvasive near-IR measurements of blood oxygen, hematocrit, body fat, and tissue hydration can be made. Near-IR methods of imaging tumors and organs are being developed. *In vitro* measurements of blood cholesterol and lipoproteins are being made, and noninvasive methods of analysis may soon become available.

Another new application of near-IR spectrometry is a method for reconstructing decorative images concealed beneath layers of paint in historic buildings. Near-IR spectroscopy can detect the presence of hidden decorative patterns and provide chemical information about the paint layers, allowing decisions to be made on the best method of restoration and/or removal of the overlaying layers.

This article discusses a number of articles that have been published on near-IR spectroscopy and covers selected topics in near-IR instrumentation. The many topics in chemometrics and qualitative and quantitative analysis are discussed as well.

II. OVERVIEW OF NEAR-INFRARED SPECTROMETRY

Buchanan and Honigs[2] published a general article on near-IR spectrophotometry containing a number of important references. The article describes the near-IR region as ranging from 750 to 2650 nm and containing signals arising from overtone and combination vibrations. Near-IR spectra often arise from molecules whose bonds contain hydrogen. The near-IR region is particularly useful for organic compounds because of the large number of carbon-hydrogen bonds as well as nitrogen-hydrogen and oxygen-hydrogen bonds found in organic compounds. Agricultural applications of near-IR spectrophotometry, including near-IR determination of sugar content in breakfast cereals, protein and moisture in wheat, fat content in meats, carbohydrates in barley, starch and lipids in flour, and the analysis of simple sugars in aqueous solutions have all been published. Industrial applications of near-IR analysis include the determination of functional groups in the spectra of coal, the percentage of carbon in alkylated materials, and the analysis of water and solutions of alkali halides under pressure. Determination of the number of hydroxyl groups in polyols has also been accomplished using near-IR analysis. Biological applications of near-IR spectrometry and terrestrial remote-sensing have been undertaken using near-IR fiberoptics.

A discussion of mathematical techniques us-

ing near-IR analysis was also provided by Honigs and Buchanan. Principal component techniques and discriminant analysis using Mahalanobis distances were mentioned, and the number of samples required to determine a particular component was discussed in connection with several papers.

New developments in near-IR hardware, including new sources of near-IR light such as lasers and light-emitting diodes, and new wavelength selection devices such as a tilted-mirror interferometer, are now available. New detectors, including photodiode arrays, germanium detectors, charge-coupled devices, and silicon detectors are being increasingly employed in near-IR analyses.

As Honigs and Buchanan pointed out, in the 1980s a large increase occurred in the number of scientists working in the near-IR region. Engineers, chemists, pharmaceutical scientists, physicists, and physicians have been publishing papers in which near-IR spectrophotometry has been the principal method of analysis.

Watson[3] published an article on near-IR reflectance analysis of agricultural products. The use of near-IR spectrophotometry to determine fat concentration in milk is discussed in the article, and the rationale behind the use of multiple wavelengths in multicomponent analysis is described. Watson points out that one can determine as many components in a sample as there are spectrally independent wavelengths. In the near-IR region this number is commonly around ten.

IR instrumentation using interference filters in both a perpendicular and tilted configuration (with reference to the incident beam) are described as well as some elementary mathematical transformations based on Beer's Law.

Sample preparation has been described as the great weakness of near-IR spectrophotometry. Analysis of solid samples requires the samples to be uniform in terms of constituent distribution and particle size. Grinding is commonly used to assure both homogeneity of component concentrations and particle sizes. Studies indicating no consistent relationship between particle size distribution and near-IR prediction equations are described by Watson; however, Watson's paper indicated that correlations did exist between the type of grinder or mill used and the prediction equations developed by multiple regression procedures. The wide acceptance of near-IR reflectance methods is attributed to the rapidity with which analysis can be conducted as well as the cleanliness of near-IR methods (including the fact that no chemicals are required to perform an analysis) and the low cost per sample when large numbers of samples are analyzed.

The 1986 biannual "Fundamentals Review in Analytical Chemistry for Infrared Spectrometry" also discussed near-IR spectrophotometry.[4] Here, the near-IR region is described as ranging from 800 to 2500 nm, and the combination and overtone bands appearing in this region are described as 10 to 1000 times weaker than the corresponding bands in the mid-IR range. The fact that the signals in the near-IR are composed predominantly of overtones and combination bands is also used as an explanation for the broad spectral features observed in the near-IR region. It is indicated, however, that even these broad spectral bands can be attributed to functional groups in the different chemicals of a mixture. High S/N ratios (on the order of 100,000) can be obtained with diffuse reflectance instruments in the near-IR region. The reproducibility of intensity measurements is on the order of microabsorbance units. Because the intensity of near-IR absorption bands is much less than mid-IR bands, the penetration of near-IR radiation into a sample is greater in near-IR diffuse-reflectance measurements than in the corresponding mid-IR diffuse-reflectance measurements. Longer cell pathlengths in liquid analyses are also possible in near-IR transmission measurements. However, the *false-sample problem* can be a major stumbling block to the routine application of near-IR techniques. The false-sample problem occurs whenever a near-IR spectrometer is used to analyze a sample unlike any sample that appeared in the near-IR training set. In such an instance, the near-IR algorithm can only indicate what may be the composition of the sample. Near-IR techniques have also been used in the determination of glucose, fructose, and sucrose in water solutions. This application is ordinarily considered to be difficult in the near-IR region because of the strong hydrogen bonding of sugars to water, which induces considerable temperature dependence in the near-IR spectra.

In his 1983 article, "Near-Infrared Reflec-

tance Analysis: Sleeper Among Spectroscopic Techniques", Wetzel[5] discusses the current state of near-IR reflectance analysis. The use of near-IR reflectance analysis as a quantitative tool for major components present in concentrations of 1% or more is described. Quantitative analysis is made possible by the weak absorbances typically observed in the near-IR region and through the use of multivariate statistical procedures. Industrial applications, including the use of near-IR spectrometry to determine the molecular weight of propylene and ethylene glycol polymers, the moisture content of coal, textile blends of cotton and polyester and rayon and polyester, hydrocarbon mixtures, cross-linkage in chemically modified starch, moisture in pharmaceutical excipients and detergent powder, and weight loss on drying of cosmetics are all described as successful industrial applications of the near-IR spectrophotometric technique.

Wetzel describes the broad nature of near-IR bands, particularly in the analysis of agricultural commodities, and gives wavelength regions that correspond to certain types of sample constituents, including the 2100 nm region common to starch and cellulose, and the 2310 and 2348 nm regions that indicate the presence of lipids. Peaks present at 2055 and 2180 nm are attributed to the amide structure present in proteins.

The effect of particle size on light scattering and effective light pathlength (depth of light penetration into the sample) are apparent in near-IR spectra. Smaller particle sizes generally yield more scattering events per unit volume, lowering the near-IR reflectance spectral baseline and reducing the height of the spectral absorption peaks.

A number of the assumptions must be made by someone contemplating the use of analytical diffuse reflectance techniques. Some of the assumptions that are more important to analytical near-IR diffuse reflectance spectrophotometry include the one that weakly absorbing samples allow a simple specular reflectance correction to be handled by the multiple linear regression process, and that a constant specular background is observed at any wavelength where the refractive index is constant and the effective absorption is negligible. Weak overtone and combination bands shift in wavelength more easily as a result of the environment of a sample than the corresponding fundamental signals. Scattering is greater at shorter wavelengths in near-IR spectrophotometry than at longer wavelengths. A common feature of near-IR spectra is a low baseline absorbance at shorter wavelengths that gradually increases to a higher baseline absorbance at the longest wavelengths in the spectrum.

The importance of maintaining uniform particle size among samples is emphasized by Wetzel. Because near-IR wavelengths are generally on the order of 1 to 2 μm, the lower limit on particle size for reflectance measurements is 5 to 10 μm while the upper limit on particle size is determined by the need to get a statistically significant number of particles in the incident near-IR light beam. In practical terms, particle sizes on the order of 100 μm often seem to give the best results. Particle size and shape affect the amount of specular reflectance (the reflectance of incident light that does not strongly interact with the sample constituents) contributing to the observed spectrum.

The specular contribution to overall reflected energy is a function of both the absorbance of the sample and the refractive index of the sample. In the near-IR region the refractive index of most samples is essentially constant over the wavelength range of interest.

Wetzel also correctly emphasizes the need to use spectrometric instrumentation with very high S/N ratios (noise levels in the microabsorbance unit range) for quantification. Such instrumentation requires a high intensity, stable light source, and low noise detectors with high sensitivity. In the most commonly used near-IR instruments, a tungsten-halogen lamp serves as this light source and lead sulfide, lead selenide, or germanium semiconductor detectors are employed. These detectors are often used at room temperature, although lead sulfide in particular may be cooled by thermoelectric means. Liquid nitrogen-cooled lead sulfide and indium antimonide detectors are also occasionally employed in custom-designed instruments. Instrumentation can be based on interference filters, interferometers, or monochromators with f/2 optics (or better) to maximize light throughput.

Additional features of modern near-IR instruments include integrating spheres to collect a maximum amount of the diffusely scattered

radiation from the sample, and rotating sample cups to permit time averaging of spectra and to reduce the effects of repack variations on calibrations. When integrating spheres are not employed, the use of multiple detectors at a 45° angle from both the incident beam and the sample plane permit fairly efficient collection of diffusely scattered radiation while rejecting a significant portion of the specularly reflected radiation. Periodic referencing of the sample-reflected radiation to the incident radiation also helps to eliminate the effects of source intensity changes on the final calibration. This referencing is typically accomplished several times at each wavelength as the spectrum is recorded.

Near-IR spectrophotometry is usually described as a rapid method capable of performing an analysis in <1 min. The time-consuming portion of developing the near-IR method arises during the collection of the training spectra from which qualitative and quantitative prediction equations for sample constituents are calculated. Near-IR spectra are normally recorded at a number of wavelengths, ranging from a half dozen to several hundred. At least one of these wavelengths must include some response to the analyte of interest in the samples; the intensity of the signal observed at this wavelength is correlated with the concentration of the analyte of interest or sample property being studied. Typically, a large number of wavelengths are either positively or negatively correlated to the analyte of interest. The use of these additional information vectors can improve the overall correlation between the spectra and the analyte of interest. It is generally best, however, to minimize the use of additional information vectors to avoid spurious correlations resulting from noise (which may increase the correlation between the spectra and the analyte for the training set but reduce the accuracy of the relationship between the spectra and the analyte concentration for new samples being tested).

Various wavelengths or groups of wavelengths can be correlated to different analyte concentrations or sample properties, thus permitting a number of prediction equations to be developed for several analytes or sample properties in a single sample. Simultaneous multicomponent analyses can be performed by applying multiple prediction equations to a single near-IR scan collected from an individual sample.

The training process begins with the collection of a set of samples that have been previously analyzed for the sample property of interest by a reference method (a minimum of 30 such samples is often suggested). The analyte concentration in these training samples should cover the entire range of interest (i.e., the entire range that may be encountered during analysis of the samples). Furthermore, the training set should encompass the entire range of the expected matrix variability. In other words, if one wished to determine protein in wheat, where protein concentrations range around 10 to 20%, one would create a training set containing 10 to 20% protein in the samples. Furthermore, if one desires to determine protein in different species of wheat, a representative number of each of these species is included in the training set (with each species again covering the entire range of expected protein concentrations). The analyte values from the previous reference analysis of the training set are correlated using a computer to the absorbance values at all wavelengths in the recorded spectrum for each of the samples. Wetzel reports relative standard deviations of 1 to 2% for major constituents of agricultural products analyzed in such a manner with near-IR spectrophotometry.

The training process typically employs a spectral matrix whose rows represent sample numbers and whose columns represent wavelengths. The elements of this spectral matrix are absorbance values recorded at a particular wavelength for a particular sample number. Collinearity in this matrix generally prevents the use of multiple linear regression on the full spectral matrix. Matrix inversion requires fewer wavelengths in the spectral matrix than samples, a requirement often met by eliminating wavelengths from the analysis that do not correlate strongly to the analyte of interest. Three methods are still used commonly to select wavelengths for inclusion in the prediction equation: (1) all-possible-combinations multiple linear regression, (2) reverse stepwise multiple linear regression, and (3) forward stepwise multiple linear regression. All three techniques are intended to reduce the number of wavelengths included in a calibration to the point at which the calibration becomes

robust and produces prediction errors that are approximately equal to the estimation error calculated in the correlation process.

The steps involved in developing the near-IR prediction equation are (1) selection of discrete wavelengths or wavelength regions to be incorporated into the prediction model, (2) selection of a fitting procedure that assigns empirical coefficients to variables in prediction equations, and (3) testing the applicability of the prediction model to a range of samples (cross-validation). In most cases, additional chemical or spectroscopic knowledge provided by the spectroscopist is incorporated into step 1. This knowledge is not absolutely necessary, however, and near-IR spectrophotometric techniques can rely entirely upon correlation/transformation statistics. Other fitting procedures in use at the time of Wetzel's publication include least-squares curve fitting, principal component regression, and row reduction.

A historical look at the near-IR spectral region published by Wheeler[6] discusses problems in instrumentation that are relevant today. The origin of near-IR spectra, the characteristic absorption bands that appear in the near-IR region, the effects of hydrogen-bonding on spectra, and a brief description of analytical applications of near-IR spectrometry also appear in the paper.

IR detectors such as the bolometer are not effective in the near-IR region. However, a number of solid-state photoconductive detectors exist that have been in use in the near-IR for a long time. Photoconductive detectors have been constructed from lead sulfide (PbS), antimony sulfide (Sb_2S_3), and bismuth sulfide (Bi_2S_3). Incandescent lamps give continuous radiation to 3000 nm and are used commonly in the near-IR region. Optics are generally constructed from quartz, which is fairly transparent to 2500 or even 3000 nm.

The most common near-IR chromophores are bonds containing hydrogen (e.g., carbon-hydrogen, oxygen-hydrogen, and nitrogen-hydrogen bonds). The carbon-hydrogen bond has a fundamental absorption at 3500 nm and overtones at about 1800 nm (for the first overtone), 1200 nm (for the second overtone), 850 nm (for the third overtone), and a fourth overtone at the edge of the visible region, around 700 nm. Oxygen-hydrogen and nitrogen-hydrogen bonds both exhibit fundamental bands around 2800 nm and a first overtone around 1400 nm, a second overtone around 950 nm, and a third overtone at the edge of the visible region, around 700 nm.

Near-IR spectra have been of great utility in the study of both inter- and intramolecular hydrogen bonding. Of course, the amount of intermolecular hydrogen bonding varies with the concentration of the analyte. In consequence, the relative peak intensities of the two -OH peak maxima near 950 nm (given by free and bonded hydroxyl groups) have been used to determine the amount of hydrogen bonding between molecules. Methanol and isobutyl alcohol have been shown to exist in solutions as groups of three or four molecules in this manner.

McClure[7] has described future prospects for the near-IR technique. According to McClure, near-IR research has focused on demonstrating its potential for a rapid measurement of chemical constituents in ground samples or chemically complex matrices. The basic approach has been to use a training set, frequently containing 100 or more samples, to develop a prediction model on a computer that is applied to new spectra obtained from unknown samples to achieve simultaneous multicomponent analysis. Most of the analyses are done in the wavelength domain. However, McClure has shown successful near-IR analysis in the Fourier domain, using as few as 50 Fourier coefficients to reconstruct an entire near-IR spectrum. Moreover, McClure has demonstrated that the first 11 Fourier coefficients are generally adequate to estimate chemical compositions. Some other benefits of working in the Fourier domain include: (1) smoothing of spectra can be achieved without distortion or loss of end points; (2) deletion of the first Fourier coefficient from the calibration procedure generally corrects for particle size variations in solid samples; and (3) collinearity between wavelengths is broken up when transformation is made to the Fourier domain. McClure expects the "uncertainty surrounding data treatment" caused by wavelength selection procedures to disappear as full spectral compression techniques such as Fourier transformation and principal-axis transformation become more widely employed.

Another article on quantitative and qualita-

tive near-IR analysis was written by Tunnell,[8] in which the need for thermostatic control of samples (particularly liquid samples, whose spectra are highly dependent on temperature) is emphasized.

When the absorbances of a number of samples are measured at two wavelengths, the spectral data can be plotted as points in a two-dimensional space. With this method, samples of the same material appear as clusters of points together on a graph, while different materials form different groups that may or may not overlap. For each group, a standard deviation in the multidimensional space can be calculated. The group means and standard deviations are ruled by univariate statistics, and an ellipse representing three standard deviations can be drawn to enclose 99% of the samples belonging to a single group. Tunnell uses a multidimensional standard deviation metric known as the Mahalanobis distance to measure the distances between spectral clusters. Using this distance, only 1% of the samples in a group will fall outside the three standard deviation distance from the mean "center" of a cluster of spectral points.

The identity of an unknown sample is determined by measuring the absorbances of the sample at the relevant wavelengths and calculating the Mahalanobis distance to each group of samples in a spectral library. In many cases the spectrum of an unknown will be within three Mahalanobis units of only one spectral cluster, giving near-IR spectrometry the power to identify unknowns from a spectral library. In theory, if the unknown is not in the library, then the Mahalanobis distance from the unknown to each cluster will always be greater than three standard deviations. Of course, two wavelengths are not usually enough to separate a large number of compounds. Most near-IR spectrometers, however, are capable of recording a large number of wavelengths, and analyses can be accomplished in a multidimensional hyperspace.

A detailed description of the procedure by which a sample constituent is matched to a sample property measured by near-IR reflectance analysis has been published by Montalvo et al.[9] In this paper, the concentration of sucrose in water solutions and in cotton is measured using near-IR spectrophotometry and forward stepwise multiple linear regression. The cotton samples and solutions are "spiked" with sucrose to prepare standards for the demonstration. The use of hypothesis testing for a significant correlation between near-IR spectra and sucrose concentration is demonstrated in the paper. The hypothesis test is designed to work with a small number of samples (<10), and employs a model intended to determine whether the observed correlation is due to a true relationship between near-IR spectra and analyte concentration, or merely due to noise effects. The model uses one wavelength term in the regression equation and a test statistic based in part on the correlation coefficient, and is subjected to a randomization test to determine its significance.

A 1987 article by Beebe and Kowalski provides an overview of the multivariate calibration procedures used in near-IR spectrophotometry.[10] The overview concentrates on multiple linear regression, principal component regression, and partial least-squares fitting procedures. Standard linear algebraic notation is used throughout the article, and an introductory section provides definitions for all notations typically encountered in multivariate calibration literature. Matrix representations, row space, column space, hyperspace, projections, and factors are all defined carefully. The concepts of eigenvectors and eigenvalues are also explained in the introduction.

Multiple linear regression is demonstrated to fit a hyperplane to spectral data in hyperspace without attempting to model any underlying structure or factors that may exist in either the sample near-IR spectral points or the analyte concentration matrices. Multiple linear regression is described as an adequate procedure in ideal situations (a situation in which spectral matrices do not show strong colinearity and in which noise is random). In practical situations, however, multiple linear regression can be difficult to use when collinear matrices must be inverted. Multiple linear regression can also incorporate significant amounts of irrelevant information into the final prediction model.

Principal component regression (PCR) is shown to be a two-step model-building procedure. The first step of this procedure is the determination of the eigenvectors (or factors) of the near-IR spectral matrix (a matrix whose rows

represent sample numbers and whose columns represent spectral wavelengths). The near-IR spectral matrix is projected onto a reduced space to produce a matrix called the score matrix. The columns of the score matrix represent new axes in hyperspace. These new axes, or factors, are the axes that best describe the variations in the original wavelength variables. The first principal component, or axis, is formed by the linear combination of the original wavelength axes that point in a direction most correlated to all of the columns of the near-IR spectral matrix in row space. This is the direction in column space that best describes the largest source of variation in the samples. Principal component regression is an effective means of condensing near-IR spectral information to enable analysis in a reduced wavelength space. When all of the eigenvectors of the original spectral data are used, principal component regression produces results that are equivalent to those produced by multiple linear regression. An important advantage of principal-axis transformation as a precursor to regression is that the principal component axes are mathematically orthogonal. In other words, the PC axes are independent and can be added to or removed from the prediction model without changing the coefficients associated with the remaining axes in the model. Another useful feature of the principal-axis transformation process is that the major principal components generally contain the valuable spectral information, while the later principal components (or axes) tend to contain noise. In near-IR analysis it is not unusual to find that only the first ten or so principal axes contain useful spectral information. The next hundred (or even several hundred) principal axes contain only noise and can be eliminated from the prediction model.

Another calibration method, partial least-squares (PLS), is gaining increasing acceptance as a quantitative procedure in near-IR spectrophotometry. PLS is a method of constructing the prediction model that estimates simultaneously the underlying factors or axes in both the near-IR spectral information matrix and the analyte concentration matrix. These factors are then used to define a subspace in the near-IR spectral matrix that is best able to model the analyte concentrations. The PLS method is conceptually similar to principal component regression and produces a transformation matrix (often called the loadings matrix) that serves as a map connecting the near-IR spectral wavelength space to the reduced spectral wavelength space represented by the scores matrix.

Factor-based methods of near-IR spectrophotometric analysis, such as PCR and PLS, have two advantages over multiple linear regression. Multiple linear regression cannot be used when the wavelengths or columns in the near-IR spectral data matrix are linear combinations of each other (colinear). This situation is quite common in near-IR analysis and is a major reason for the origination of the wavelength-searching schemes. In addition, the inversion algorithms used commonly in quantification become unstable when some columns or wavelengths are very nearly linear combinations of the others. Because most near-IR analyses are conducted with a large number of wavelengths (closely spaced in some cases), and because near-IR spectral bands tend to be broad and superimposed on a variable baseline, unstable matrix inversions tend to be the rule in near-IR analysis rather than the exception. Wavelength search schemes are a less acceptable alternative to factor-based methods of analysis because the process of deleting wavelengths from a prediction model can inadvertently delete those with useful spectral information. Factor-based methods retain more of the important qualities of the full near-IR spectra while eliminating the problems caused by collinearity in the near-IR spectral data matrix.

While using all of the principal components in the calibration set will increase the correlation to the analyte concentration values, on cross-validation it becomes apparent that the use of all of the components merely permits the model building procedure to fit noise into the final prediction model. Cross-validation shows noise incorporation clearly by providing a large correlation between the analyte concentration values of the training set and their spectra, while producing a small correlation between "unknown" samples and their analyte concentration values predicted using the model developed on the training set.

A 1987 paper on process analytical chemistry by Callis et al.[11] describes off-line, at-line, and in-line process analytical chemistry. Off-line process analysis means that the sample is analyzed

in a centralized facility with sophisticated analytical instrumentation. The off-line scheme provides for economical use of expensive instrumentation; however, it introduces a delay between the submission of the sample and the reporting of the analysis, with additional administrative costs, and leads to competition among users for resources. In at-line analysis, an instrument is dedicated to a particular analytical purpose and is installed in close proximity to the process being monitored. This approach allows for faster sample processing and the employment of a simpler instrument that is more rugged and easier to use than those normally encountered in centralized off-line facilities. Off- and at-line analyses have been commonly used for years.

The area of on-line process analytical chemistry is newer, and process analytical chemistry is sometimes considered a distinct subdiscipline of analysis. A further variation on the on-line scheme, in-line process analytical chemistry, is used to describe the situation in which the actual process stream is monitored without having to split off a small portion of the stream to run through analytical instrumentation. The ultimate in process analytical chemistry is noninvasive analysis, a further subset of in-line analytical chemistry. In noninvasive analysis the probe does not physically contact the sample, making special sample handling considerations unnecessary. Callis and co-authors present near-IR spectrophotometry in the context of noninvasive process analytical chemistry.

The low absorptivities characteristic of organic compounds in the near-IR region means that a thin layer of adsorbed material on optical windows in process streams does not ordinarily degrade the analytical results. Moreover, quantitative measurements of highly scattering compounds can be made because the scattering coefficients are much greater than the absorption coefficients in the near-IR region. In these circumstances, Beer's Law provides an effective means of correlating analyte response to analyte concentration. Near-IR diffuse reflectance spectrophotometry is particularly useful with noninvasive process analysis because only a single window is needed in the process stream, rather than the two windows required by transmission measurement. Relatively simple fiberoptics are sufficient for most near-IR process measurements.

Callis presents a case study involving a study of 43 samples of unleaded gasoline of known octane number. On-line measurements are made using near-IR spectrophotometry and fiberoptics. A forward stepwise multiple linear regression procedure using only three wavelengths is found that correlates to reference octane numbers with a standard error of 0.3 octane, which is the known precision of the octane engine. Additional sample properties are obtainable from the near-IR spectrum. In fact, Callis and co-workers are able to perform eight tests on a gasoline spectrum in 20 s.

III. NEAR-INFRARED SPECTROPHOTOMETRIC INSTRUMENTATION

Near-IR spectrophotometric reflectance instrumentation was described by Rotolo.[12] Virtually all near-IR reflectance instrumentation uses multiple wavelength readings of a sample to enable effective quantitative analysis. Rotolo's instrumentation used multiple least squares regression of 6 to 19 reflectance readings to permit quantitative analysis. Samples were powdered and placed in either a glass-covered or an open sample cup. The glass-covered cup minimized problems of reproducible sample loading because the glass and holder were constructed to create a reproducible pressure on the powder in the sample cup. The open sample cup requires more expertise to pack reproducibly, but remains more effective for some types of samples such as those with a high oil content. Certain wavelength bands are found to correspond to agricultural components of interest: 2310 nm is used for oil detection, 2180 nm for protein detection, 2100 nm for starch detection, 1940 nm for moisture detection, and 2230 and 1680 nm are used as reference wavelengths. All of the instrumentation discussed has a light source, lens system, wavelength selector, and detector. The light source in each system is a tungsten lamp. Lens systems are used to collimate light beams onto the powdered samples. Wavelength selection is achieved through diffraction grating monochromators and narrow bandpass interference filters. A major advantage of interference filters is that they allow more energy to reach the sample. Furthermore,

interference filters are less expensive and can withstand more abuse than the average monochromator with sine-bar drive. Tilting interference filters allows them to be used at more than one wavelength, but also increases the bandpass of the interference filter. Lead sulfide solid-state detectors are used in all of the instrumentation presented by Rotolo.

Different schemes are available to separate specular from diffuse reflectance. The specular component of the reflectance can be produced by the glare of the glass cover over the sample or by surface glare of the sample, and contains little information about the composition of the sample itself. The specular reflectance is essentially the reflected image of the tungsten lamp. Diffuse reflectance, having penetrated the sample window and the surface of the sample powder, is scattered back out to the detector after interacting with absorbers in the sample. Diffuse reflectance therefore contains useful information about the contents of samples. Both direct-looking optics and integrating-sphere optics have been employed to enable illumination of the sample and collection of diffusely scattered radiation while minimizing pickup of specular reflection.

Direct-looking optics position the detectors at a 45° angle between the incident sample beam and the plane of the powdered sample surface. Light reaching the detectors is reflected directly from the sample surface in this design. Specular reflection is directed predominantly back into the light source. In the integrating sphere, more diffusely scattered light ultimately reaches the detector. Light in the direct-looking optics that is scattered by the sample and does not reach the detector is lost. The use of an integrating sphere to augment the direct-looking optics permits specular reflection to continue to be directed back into the light source. However, by reorienting the detectors from the 45° angle to a lower angle that "looks up" at the sphere wall, more diffuse reflectance can be picked up and effects caused by the orientation of particles on the sample surface can be averaged. One important characteristic of the integrating sphere design is that the detectors themselves do not directly view the sample, but instead view only diffusely scattered light that has undergone multiple reflections from the diffusely scattering surface of the integrating sphere wall. The integrating sphere wall is coated with a gold powder that has a high reflectance throughout the near-IR region. The use of a scattering, but highly reflective, surface in the integrating sphere permits the effects of superficial sample inhomogeneity (caused by directional dependence of the reflection from the sample) to be reduced.

Calibration of filter near-IR instrumentation is usually accomplished through multiple linear regression. A standard error of estimate (SEE) is calculated for the set of calibration samples and a standard error of prediction (SEP) is calculated from validation samples (samples that have not been included in the calibration set). Ideally, both the SEE and the SEP are equal to zero. Error is defined as the difference between the value predicted by the near-IR instrumentation and the value ascertained by reference laboratory procedures. In the case of perfect spectrophotometric analysis, both the near-IR instrument and the laboratory reference method gave the same sample composition. In practice, however, there is usually a small difference between the instrumental prediction and the laboratory reference value. In this case, both the SEE and SEP have some value slightly greater than zero. The major contributions of near-IR reflectance methods to prediction error are from sampling, sample preparation, and errors in the reference laboratory method.

It is not uncommon for certain constant factors to be added or subtracted to prediction results in near-IR reflectance spectrophotometry to account for biases in predictions. Sources of bias can be long-term instrumental drift, a change from one instrument to another, a change in instrumentation used to produce the laboratory reference values, or a change in the sample matrix. In many cases, such biases can be corrected by adding or subtracting a constant term to the prediction. The proper constant term to be added or subtracted from the prediction is ascertained by analyzing additional reference samples. The difference (or bias) revealed by this method is applied to the result obtained by near-IR spectrophotometry using the previously determined prediction model. In most process applications, once an instrument is calibrated properly the calibration need not be repeated unless there is a drastic change in the product being analyzed or the process involved.

A 1981 issue of *Crop Science* describes a

near-IR spectrophotometer driven by a digital PDP-11 minicomputer.[13] The monochromator-based instrument is used to analyze forage and grain samples. A concave holographic grating collects two full spectra per second by wobbling the grating back and forth. A special cam produces a scan that is linear with wavelength. The wavelength accuracy is ± 1 nm in the near-IR region. The cam drive mechanism also rotates a filter wheel that is synchronized with the motion of the grating. The filter wheel is used to reduce stray light and to select grating orders to obtain either near-IR or visible spectra of samples. During the forward swing of the grating, the filter wheel blocks all but the first order of the grating (near-IR); during the backward swing, the filter blocks all but the second order of the grating (visible). Each revolution of the filter wheel also creates a dark period that is used to cancel drifts in the detection electronics. The light source is a 100-W tungsten-halogen lamp.

The energy throughput of the monochromator at 2000 nm is approximately 350 μW. The wavelength repeatability of the system is ± 0.01 nm. The stray light measured at 2305 nm is 0.08%. The average root-mean-square (RMS) noise on a single scan is 550 microabsorbance units (this noise level is determined on a ceramic standard). Averaging ten scans produces a noise level of 110 microabsorbance units. In most cases, smoothed second derivative spectra are found to produce superior results in multiple linear regression with this instrument. The standard error of performance is approximately one third larger than the standard error of estimation for all sample constituents and methods of data processing. One hundred forage samples are used in the SEP calculations.

Wavelength selection is always an issue in calibration. It is possible to select wavelengths to fit just a few specific samples in the calibration set. The resulting equations in the prediction model are effective for the calibration set but ineffective for all similar samples that were not used to develop the prediction model. The "false sample problem" (i.e., the problem that arises when a sample is presented to the instrument that is not represented in the calibration set) can also contribute to inaccurate near-IR analyses. An H statistic is calculated for each sample analyzed by the instrument to delete false samples. A high value on the diagonal of the H matrix is used as an indicator that the sample under examination is a false sample (i.e., a sample that is different from the samples of the calibration set at the wavelengths being used in the prediction model). The effectiveness of the H statistic method in the detection of false samples is not discussed.

Near-IR instrumentation for process control was a topic for Cooper[14] who described the manufacturing process as a dynamic system with many variables and factors that must be controlled. These variables include cost efficiency, the availability of the process line (if the line is used for the manufacture of other products), the availability of raw materials, and delays and malfunctions during the production process. Cooper depicts a hypothetical manufacturing process involving four ingredients in one final product. The manufacture of the final product involves the processing of raw materials, further mixing and processing, conditioning, further mixing and processing, final mixing of the fourth ingredient, final conditioning, and packaging. In this hypothetical process, analysis is performed on the incoming raw materials as well as at five points in the process stream. In addition, an analysis of the finished product is made. The need to establish control limits on the process rapidly to prevent production of large amounts of unacceptable product is emphasized. Near-IR methods are shown to be ideal for timely analysis of both materials and product.

The spectrometer described by Cooper permits scanning from 750 to 2500 nm. Processes can be analyzed using diffuse reflectance spectra, transmission spectra, or a combination of transmission and reflectance techniques. Transmission spectra of strongly absorbing liquids, such as aqueous solutions and high-fat products, are usually scanned from 750 to 1100 nm. Transmission spectra can be recorded using pathlengths from 1 to 10 cm. In most cases it is necessary to scan a complete spectrum, at least initially, because it is difficult to predict which regions of the spectrum produce the best analytical results without prior knowledge.

The spectrometer uses a cam-driven concave holographic grating with visible and near-IR filter wheels. The tungsten-halogen lamp and lead sul-

fide detector commonly employed in near-IR instrumentation are also employed in this spectrometer. Another spectrometer using tilting interference filters was also described. Normalization of photometric responses between the two spectrometers permits calibrations on one spectrometer to be transferred to another. Generally, a representative subset of the original calibration samples (approximately 10 to 20 samples) must be scanned on both instruments in order to calculate the normalization.

Tilting interference filter instruments with three, six, and seven filters that cover 400 to 900 nm wide spectral regions are described by Cooper. The seven-filter instrument is designed particularly for on-line process control applications.

A 1987 report by Davies and Grant implicates fluctuations in atmospheric water vapor as a major source of noise in near-IR spectra.[15] Fluctuations in atmospheric water vapor are often caused by the cycling of the laboratory air-conditioning system. Their laboratory air-conditioning system maintains a temperature of 23°C, ± 1°C. Precise temperature control is required in order to collect near-IR spectra with minimum noise, and this temperature control is insufficient to keep the noise level of the scanning spectrometer within the manufacturer's specification. Noise spectra are recorded in the publication with the laboratory air-conditioning system on and off. The spectra are obtained over a range of 1100 to 2500 nm in 2 nm intervals, and averages of 50 scans are employed to calculate a single reported spectrum. Noise spectra are recorded by calculating the standard deviation at each wavelength of the 50 scans normally averaged to produce a single spectrum. In this manner, 20 noise spectra are recorded with the air conditioner on and 20 more with the air conditioner off. The difference between the two sets of spectra is dramatic and closely resembles the spectrum of water vapor recorded by Norris, who introduced steam into his spectrometer to produce a reference spectrum for water vapor. In order to achieve the lowest noise levels in near-IR spectrometers, it is necessary to not only control the temperature of the spectrometer and the sample but the relative humidity of the environment of the spectrometer as well.

Hirschfeld has published a highly optimized optical design for diffuse reflectance work.[16] He pointed out that in the IR and near-IR regions, performance of the diffuse reflectance technique is limited by the low efficiency of the sampling accessories. Hirschfeld quotes average efficiencies of 2 to 6% for integrating spheres and 10 to 12% for various ellipsoidal mirrors. He describes a concentric confocal ellipsoidal mirror arrangement with an efficiency of up to 37%, achieved by geometrically optimizing the accessory for diffuse reflectance. Hirschfeld points out that the throughput requirements of an optical beam are strongly increased when the diffuse reflectance mode is employed. The increased requirement is often met through the use of a larger detector. Unfortunately, in the near-IR region larger detectors emit greater noise than smaller detectors. In the near-IR detectors, noise increases as the square root of detector size. Diffuse reflectance accessories must optimize collection efficiency, transmission, and throughput-matching efficiency in order to make up for the large increase in the solid angle of the incident beam that occurs following diffuse reflectance from a sample.

Another recently developed accessory for near-IR reflectance analysis is described by Lodder and Hieftje.[17] Near-IR spectrometry began as a method of analyzing powdered solid samples. In recent years, however, many spectrometers have been sold for liquid sample analysis. Unfortunately, the liquid analysis accessories sold for these diffuse reflectance instruments have often been cumbersome and expensive, and generally require a relatively large volume of sample (in milliliters). Furthermore, complex purge/fill and wash cycles are necessary to prevent clogging. The large volume of the cell adds to the problem of thermostatically controlling sample temperature. The use of a conical aluminum reflector and single cavity microscope slide beneath an integrating sphere with a lead sulfide detector permits the analysis of small volumes of liquids in diffuse reflectance instruments. Furthermore, this liquid sample accessory can be used easily with hazardous samples when the microscope slide is discarded following use. The optics are based on an integrating sphere, a 90° primary conical reflector, and a 136° inverted secondary conical reflector positioned directly beneath the cavity of the microscope slide. The use of different cavity

slides provides for different pathlengths and the coverslip over the cavity lowers the liquid sample evaporation rate. Single-cavity microscope slides are available to contain 70 to 110 μl of a liquid sample.

The disposable liquid microcell is tested in the analysis of aqueous sodium chloride solutions. The determination of sodium chloride in water can be difficult in the near-IR for several reasons — sodium chloride has no absorption bands in the near-IR, water has very strong absorption bands in the near-IR, and the water absorption bands are temperature dependent. Eighty training samples and 80 validation samples are employed in testing the liquid microcell. The correlation coefficient for the training set (using five principal components) is 0.97 and the correlation coefficient (r^2) for the 80 validation spectra is also 0.97. An absolute detection limit of approximately 100 μg of sodium chloride is obtained.

The disposable liquid microcell has a number of practical advantages including: (1) the cell is faster and easier to use than conventional liquid analysis accessory; (2) 100 μl of liquid rapidly reaches thermal equilibrium; and (3) no purging/filling or wash cycles are required. Cells can be rapidly filled with a precision pipette and easily cleaned or discarded following analysis. The configuration of the disposable liquid microcell permits sensitive detection by enhancing transmission through the sample in a reflectance instrument.

A modification of the conical microsample cup for near-IR analysis has been published by Yeo and Honigs.[18] Conventional powdered sample cups require 5 to 8 g of solid sample to obtain chemical information from near-IR spectra. The Yeo and Honigs microsample cup requires only about 350 mg of sample powder. The sample cup is also used to obtain the spectrum of individual soybeans and liquid samples. The sample cup is able to detect as little as 100 μg of powdered soybean without any loss of chemical information when compared to the conventional 5 to 8 g sample cup. Furthermore, the microsample cup is able to analyze as little as 80 μl of liquid sample. The microsample cup is constructed of aluminum and brass, and it provides for an adjustable height. The height adjustment is accomplished by raising or lowering the aluminum inner cone, which is fixed atop a brass insert. The brass insert is threaded into an aluminum outer holder. The entire brass insert is raised or lowered by turning it in the aluminum outer holder. The microsample cup provides space for a lens holder which is placed over the sample. A 2.5-cm diameter f/1 lens made from fused silica is used in some experiments to increase collection efficiency.

When the microsample cup is used to collect spectra of liquid samples, strong absorbance peaks are observed at wavelengths of lesser absorptivity (the shorter wavelength regions). In the conventional liquid drawer, strong absorbance peaks are observed only at wavelengths of 1500 to 1600 nm. The microsample cup has a longer effective pathlength in the more weakly absorbing regions than the conventional liquid drawer. The microsample cup is therefore ideal for analysis of small amounts of liquid samples and for analysis in shorter wavelength regions. Cleaning of the cup is simple, reducing the danger of sample contamination.

A near-IR detector for high performance liquid chromatography (HPLC) has also been constructed.[19] In this preliminary publication describing a near-IR detector for HPLC, the detector's performance, including drift and noise levels, sensitivity, response time, linear dynamic range, dead volume, selectivity, and nondestructive character are evaluated. The near-IR HPLC detector is constructed with two specific problems in mind: (1) the detection of solutes without UV-visible chromophores and (2) preparative HPLC analysis (where concentrations are high enough that dilutions would be necessary for traditional detection methods). The selectivity of multiwavelength analysis and the ability of near-IR radiation to detect virtually all organic materials also provide strong reasons for considering the use of near-IR spectroscopy in HPLC detection. The cell constructed by Ciurczak has a nominal volume of 150 μl. When longer pathlengths are required, the volume of the cell can be increased to up to 400 μl by changing the thickness of an O-ring. Sucrose, glycine, and valine are analyzed successfully in concentrations from 10 mg/dl to the highest concentrations attainable. These compounds are selected for testing because they do not contain good UV-visible chromophores. Solutions of methanol, ethanol, propanol, and tetrahydrofuran are also examined by

HPLC with near-IR detection to demonstrate quantification of these normally undetectable materials. The near-IR detector is based on a commercial near-IR spectrometer using a 1000-W quartz-halogen lamp, grating monochromator, and low-noise lead sulfide detector. The S/N ratio for major peaks is approximately 50,000:1 for this detector. Detector drift is <0.5%/d.

The detector cell as designed has a very large dead volume. In analytical runs, this large dead volume would be a problem. In preparative scale separations, however, the volume would not adversely affect results. It should be noted that different detector designs now enable analysis to take place in a volume considerably under 100 μl. The current cell is resistant to flow variations from 0.2 to 5.0 ml/min. However, major variations are noted in spectra due to temperature changes. In general, raising the temperature of the sample lowers peak energies. This is demonstrated for the water band near 1940 nm. According to the author, temperature control should be maintained to ± 0.5°C when using a near-IR detector. Because carbon-hydrogen, oxygen-hydrogen, and nitrogen-hydrogen bonds have absorbances in the near-IR region, a near-IR detector is virtually a universal detector for HPLC.

The detection of Rift Valley Fever viral activity in Kenya[20] provides an extraterrestrial example of near-IR spectrometry. The fever survey is accomplished using the advanced theory high-resolution radiometer sensor on polar orbiting meteorological satellites operated by the National Oceanic and Atmospheric Administration (NOAA). The radiometric sensor records visible, near-IR, and IR spectra. The remotely obtained spectra are used to infer ecological parameters associated with Rift Valley Fever viral activity in Kenya. Outbreaks of Rift Valley Fever in domestic animals in sub-Saharan Africa are correlated with widespread and heavy rainfall that floods mosquito breeding habitats. Satellite remote sensing may be the only method available to conduct surveillance activities for Rift Valley Fever over an area as large and diverse as sub-Saharan Africa. Early detection of viral activity makes possible preparations for specific control measures before the infection grows out of control. The investigators in this research felt that the ground studies in remote sensing technology developed for Rift Valley Fever virus would soon be applied to other diseases that are ecologically linked.

Higher resolution near-IR remote-sensing devices are also under construction. While each pixel covers 15 km^2 in the Rift Valley Fever detection experiments, the NASA Jet Propulsion Laboratory uses an airborne visible and IR imaging spectrometer that records high-resolution reflectance spectra with 550 pixels, each pixel covering 20 m of surface. The airborne visible and IR imaging spectrometer covers the spectrum from 0.4 to 2.5 μm with a 10-nm sampling interval. The satellite and airborne near-IR spectrometric detectors are constructed from silicon and indium antimonide.

One of the newest developments in Raman spectrometry is based on excitation of the near-IR spectral region.[21] Raman spectroscopy in the near-IR avoids the fluorescence problems that commonly plague visible Raman spectrometry. Furthermore, Raman spectrometry in the near-IR can be performed on polymers and biological molecules that degrade when irradiated at visible wavelengths. Raman spectroscopy in the near-IR is usually conducted using a Fourier transform IR spectrometer. The throughput and multiplex advantages of Michelson interferometers, in combination with high sensitivity detectors, neodymium-YAG lasers, and high rejection laser filters has permitted FT-Raman spectrometry to develop.

The Nd-YAG laser, with a fundamental absorption band at 1064 nm (9398 cm^{-1}) is the light source for FT-Raman measurements. The Raman spectrum must be collected by an interferometer with an appropriate beamsplitter and detector for the near-IR region. Near-IR excitation is useful in Raman spectrometry for obtaining nonresonant Raman spectra of materials that absorb in the visible region, particularly (1) when information about the nonchromophoric parts of the molecule is desired, (2) when photochemistry can cause sample degradation, or (3) when background fluorescence is high.

Waters has suggested that near-IR Raman spectroscopy be accomplished with a Hadamard transform spectrometer.[22] In order to achieve a multiplex advantage, the Rayleigh line must be eliminated from an FT-Raman spectrum prior to

the entry of the beam into the interferometer. The rejection of Rayleigh scattering in stray light is often achieved by long-pass absorption filters or a system of interference filters. These filters require a very narrow rejection band with steep absorption edges and high transmission outside the absorption region. Current filters are less than ideal in these respects. Waters suggests that a properly configured Hadamard transform spectrometer could achieve the multiplex advantage while meeting the requirement for efficient rejection of Rayleigh scatter by operating in a manner similar to that of the double monochromator knife-edge rejector.

In a Hadamard transform spectrometer, a spectrum produced by a grating is sampled by a slotted mask placed in the focal plane. The sampled spectrum is then recombined by a dispersion stage to produce an image of the entrance slit at the detector. Only the mask moves or changes in a Hadamard transform spectrometer, making the spectrometer simpler mechanically than either a scanning double monochromator or a Michelson interferometer. Liquid crystal masks have been investigated for use in Hadamard transform spectrometry. The use of an electrooptic mask would eliminate all moving parts from the Hadamard transform spectrometer. The mask slots are designed to change into orthogonal sampling patterns that permit reconstruction of the original spectrum by a process similar to that used to solve simultaneous equations. The multiplex advantage arises from the simultaneous observation of n/2 spectral elements in each measurement interval. Waters calculates that a 511 or 1023 slot and coding mask would be required for a Hadamard transform Raman spectrometer.

A relatively new method of achieving wavelength selectivity in the near-IR region employs an acoustooptic tunable filter.[23] Acoustooptic tunable filters (AOTFs) are usually manufactured from crystals of thallium arsenic selenide or tellurium oxide, depending upon the wavelength range desired (tellurium oxide crystals are probably more common for near-IR spectrophotometric use). The AOTF system is operated by a transducer that propagates an acoustic wave across the optical path of the filter. The acoustic wave modulates the refractive index of the crystal, producing wavelength selectivity. Tellurium oxide crystals can be made to operate from 350 to about 4600 nm. High-speed data acquisition is possible with these wavelength selection devices: a 1000-point scan takes approximately 50 ms to collect. Light throughput is good because the light acceptance angle of the crystal is large. Resolution is currently better than 20 cm^{-1} and is expected to improve in the future. Diffraction in an acoustooptic tunable filter produces a spatial separation of the tuned wavelength as well as a 90° rotation of polarization from the incident radiation. Both phenomena can be used to separate the signal of interest. The device can be constructed with no moving parts, making it more rugged than typical grating monochromators. Complete spectrophotometric systems based on acoustooptic tunable filters are now available. These spectrometers are driven by microcomputers to make the systems easy to use.

Fiberoptics can be coupled to these spectrometers through a multiplexer. A number of light sources are available for use with acoustooptic tunable filter instruments, and thermoelectrically cooled detectors as well as liquid nitrogen cooled detectors can be employed. Currently, most acoustooptic tunable filter devices are intended for use in process environments. Research grade AOTF instruments may be available in the future.

IV. SAMPLE IDENTIFICATION AND QUALITATIVE THEORY

An early use of near-IR spectrophotometry in qualitative analysis is described by Honigs and co-authors.[24] The paper describes the use of cross-correlation to extract component spectra from the near-IR spectra of complex mixtures. During the training process, the computer implicitly generates the spectrum of each of the constituents the concentrations of which are being determined. The reconstruction method for spectra requires only a training set of near-IR spectra of samples in which the concentration of the analyte is known. The correlation between known concentration values and the absorbance or reflectance of each sample at individual wavelengths is used to calculate the spectrum of an individual constituent. The spectral reconstruction method can also be

used to determine the nature of matrix interactions between the analyte and the rest of the sample.

Cross-correlation is used to evaluate the similarity between two waveforms. In the case of spectral reconstruction, the sequence of absorbance values obtained for a series of mixtures at a particular wavelength comprises a sample waveform. Furthermore, the sequence of concentrations of a desired component in that same series of mixtures constitutes a second waveform. The cross-correlation of these two waveforms at each wavelength produces the spectrum of the analyte of interest.

The cross-correlation spectral-reconstruction algorithm is used to reconstruct the spectra of mixtures of cyclohexane, benzene, isooctane, and *n*-heptane. In addition, the spectra of protein and moisture in 50 wheat flour samples are calculated. The cross-correlation spectral-reconstruction algorithm was compared with spectral stripping or curve-fitting procedures, noting that spectral curve-fitting requires the spectrum of all pure components to be known and that this spectrum is often not available in near-IR spectrometry.

Another method for decomposing spectral data uses factor analysis. The eigenvectors generated in factor analysis provide spectral information on the individual contributors to the near-IR spectrum. The eigenvectors usually represent only orthogonally varying entities that do not necessarily depend directly on any specific analytes. Calculating eigenvectors can be time consuming on small computers. Spectral reconstruction by cross-correlation is a relatively simple procedure that is easily calculated on small computers.

Cowe and McNicol describe the use of principal component analysis to extract qualitative information from near-IR spectra.[25] They examined wheat flour and barley samples by near-IR spectrometry. Principal component analysis describes the variation in a full spectrum with a small number of independent variables. Principal component analysis constructs a transformation matrix that maps spectra represented as points in hyperspace into a new and smaller hyperspace. The original spectra described in an axis system of d-dimensions are recorded at d-wavelengths. The transformation matrix gives weights that are applied to each of the original wavelength dimensions to produce a new principal dimension (or axis). Thus, points in a wavelength space of 100 or more dimensions can be represented as points in a principal axis space of 10 or fewer dimensions.

Principal component analysis can be thought of as a principal axis transformation followed by a multiple linear regression. Principal axis transformation can in turn be thought of as comprising a translation of axes followed by a progressive rotation of axes to describe all of the variations present in a spectral data cluster. In transformation to principal axes, the original wavelength coordinate system is translated to the center of a spectral cluster in hyperspace. The first principal axis is then calculated by rotating the coordinate system so that the first axis is coincident with the major axis of the spectral cluster in hyperspace. The coordinate system is then rotated orthogonally along this first axis to calculate the second principal axis. The second principal axis becomes the axis that describes the largest remaining variation orthogonal to the original spectral variation. The process of rotating the coordinate system orthogonally continues until the process describes only random noise.

The coordinates of spectral points in the new coordinate system are often referred to as principal axis scores. Using the transformation matrix, these scores can be recalculated into original sample spectra. The columns of the transformation matrix provide a type of spectrum that describes the material(s) that give rise to signals on that particular principal axis.

Cowe and McNicol found that the vast majority of spectral variation is on the first principal component (which commonly contains 95% or more of the total spectral variation). When the first principal component has approximately equal weights across all wavelengths, it describes an additive "baseline bounce". This baseline bounce is often interpreted as reflecting particle size and orientation variations in the sample. The first six principal components almost always account for 100% of the total raw spectral variation. In most cases, subsequent principal components can be ignored in the formulation of quantitative models. One should note, however, that Cowe and McNicol are concerned with major constituents

of the flours, such as moisture and protein. Analysis of minor constituents requires the use of lower principal components (those that account for less of the total spectral variation).

Sometimes the shape of loadings spectra (columns of the transformation matrix) corresponds very closely to that of a known sample constituent, such as moisture. Areas of maximum weighting often correspond to known absorbance bands. In practice, however, the shape of loading spectra is often quite complex, with many areas of the spectrum and many analytes interacting. The absolute value of the weighting is important, so weightings near -1 and $+1$ can each be considered highly weighted. Sometimes changing the sign of the weights helps to determine the identity of the component. Cowe and McNicol determined that the second and third principal components of wheat flour corresponded to water, while the fourth principal component corresponded to protein. Some of the peaks in the loadings spectrum of the fourth principal component corresponded to reported peaks for wheat gluten. The fifth and sixth principal components showed little correlation with either protein or moisture, and together expressed only 0.06% of the total spectral variation.

Cowe and McNicol determined how principal components relate to a variable that is not a single constituent but is, in fact, a function of several constituents by assessing a series of 54 barley samples that had been analyzed for "hot water extract". Hot water extract is a function of the amount of sugar and low molecular weight dextrans produced during the mashing process as a result of breakdown of starch by enzymes, and is an important variable in the determination of the malting quality of barley. The first three principal components of barley spectra did not correlate well with hot water extract. The first principal component of the barley samples appeared to reflect particle size variations. The first principal component accounted for over 91% of the total spectral variation. The second principal component accounted for 7.65% of the total spectral variation, and its origin is not explained by the authors. The third principal component appeared to correlate with moisture. Together the first three principal components accounted for 99.6% of the total spectral variation. A small correlation is noted in the fourth principal component of barley flour to hot water extract ($R = 0.60$). The fifth principal component also had a slight correlation to hot water extract ($R = 0.48$). It is felt that the correlation to hot water extract occurred through a subsidiary correlation to protein rather than a positive correlation to carbohydrate.

Multiple linear regression modeling on principal component scores predicts percent protein and percent moisture in wheat flour samples. As long as a component does not have a correlation of close to zero to the analyte, the addition of extra terms to the regression model does not overfit the data (in contrast to multiple linear regression). One of the main advantages of PCR is that terms can be added to or subtracted from the calibration equation without changing the coefficients of the remaining terms. This advantage arises because of the orthogonal nature of the principal axes and spectral scores calculated from wavelength data. In most cases, Cowe and McNicol found that the first ten principal components are all that is needed to produce a usable calibration for an analyte in wheat flour. No validation samples are run for any of the principal component analyses. Cowe and McNichol show the utility of principal components analysis in the interpretation of near-IR spectra, particularly in the correlation of loadings spectra (from the transformation matrix) to analytes of interest in real samples.

Another paper describing the use of a factor analysis method in the interpretation of near-IR spectra was published in 1988.[26] In this paper, factor analysis is used to transform spectra from a space of high dimension to a space of low dimension. A section of factor space is assigned to the training set. Discriminant analysis using Mahalanobis distances is used to determine whether certain samples are similar to other samples. The main purpose of the procedure is to eliminate redundant samples from calibration sets. The procedure is applied to the analysis of protein, moisture, and oil in corn and rapeseed samples. The advantage of this procedure is that laboratory analyses can be substantially reduced for the calibration set without decreasing near-IR prediction accuracy. In many cases the laboratory analyses required to obtain analyte values for the

training set are considerably more time consuming than the recording of the near-IR spectra themselves. Thus, any method that enables a small number of calibration samples to be selected based only on near-IR spectra will greatly speed the development of a calibration set. This principle can also be used to decide if a new sample can be used with a particular calibration equation.

Data are collected using a fixed filter instrument with 19 near-IR wavelengths. Ten principal components account for 99.99% of the total variation of the spectra. In general, errors in prediction of protein range from 0.1 to 0.2%, while errors in the prediction of moisture range between 0.3 and 0.4%. The factor analysis/discriminant analysis procedure enables calibration sets to be reduced to 10 to 15 samples while maintaining approximately constant prediction errors.

Principal component and discriminant analyses are also used to extract pure component spectra from wheat semolina conditioned at three levels of water concentration. The water spectrum is extracted from near-IR spectra taken of the wheat semolina. The commonly known near-IR bands at 1940 and 1450 nm are observed by this technique. The band at 1450 nm is further resolved into two bands at 1410 and 1460 nm, which are assigned to free and bound water, respectively. The spectrum of water appears on the third eigenvector. Water concentration ranges from about 0.4% to more than 15% in the samples. At such levels it is not surprising that the spectrum of water can be extracted from the near-IR spectra of the wheat semolina. The first principal axis tends to reflect baseline variations consistent with particle size variation. The second and third principal components correspond to water content. Only the third principal component loading spectrum, however, strongly resembles water. Devaux and co-workers feel that the second principal component might model the different proportions of free and bound water.[27] It should also be noted that while the authors had no knowledge of exact water concentrations in the samples, they did have prior knowledge of the conditioning levels of the groups of samples.

A number of studies in qualitative near-IR spectrometry using Mahalanobis distances have been conducted by Howard Mark.[28-30] Mark identifies the testing of raw materials as a critical need in the pharmaceutical industry and proposes the use of near-IR spectroscopy as a rapid method of verifying the identity of every container of raw material at every point in the manufacturing process.

Mark points out that if absorbance can be measured with sufficient accuracy, then any wavelength at which absorbance differences exist between substances can be used to classify substances. As a practical matter, the choice of a small number of wavelengths to discriminate between materials speeds the analysis. By representing spectra recorded at d-wavelengths as single points in a d-dimensional space, the distance between clusters of spectral points in space can be used as a method of discriminating between different materials. This is precisely the intention of measurement by Mahalanobis distances. In most cases, two dimensions are not enough to describe the large number of raw materials that must be tested. Up to 15 different near-IR wavelengths are used, corresponding to analysis in a space of up to 15 dimensions. The Mahalanobis distance has been referred to as a "rubber yardstick" whose length depends upon the direction in hyperspace in which it is oriented. The unit distance vector for Mahalanobis distance is largest when the distance is measured along the major axis of an ellipse, while the unit distance vector is smaller in all other spatial directions. When spectral clusters in space are not all spherical, as is often the case in the near-IR spectrometric analysis of solid materials, the rubber yardstick must be used to measure the distance between groups to take into account the fact that the probability of a point appearing at a specific direction and distance from a spectral cluster in hyperspace is not constant. The Mahalanobis unit vector is described effectively by an ellipse or ellipsoid in wavelength space. The Euclidean unit vector, on the other hand, forms a circle or sphere in wavelength space. Because the Mahalanobis distance shares more closely the shape of near-IR spectral data clusters, the use of Mahalanobis distances in near-IR spectral discriminant analysis generally affords superior results to analysis using a simple Euclidean metric. Mahalanobis distances are based upon the calculation of the inverse covariance matrix of a matrix of near-IR spectra (whose rows represent samples and whose

columns represent wavelengths). Mahalanobis distances are calculated in effect from the Euclidean distance between the center of a spectral data cluster and a new spectral point. This Euclidean distance is then scaled by the Mahalanobis distance in the direction of the new spectral point to produce a distance in standard deviations.

Mahalanobis distances are used to (1) allow the detection of samples that are beyond the domain of a training set, (2) produce warnings of misclassification of samples in qualitative analysis, and (3) detect outliers during the calibration process. When an unknown sample is classified, the Mahalanobis distance from the unknown to each of the materials in a training set library is calculated, and the sample is assigned to the library group to which it is closest. In the usual case (when the unknown sample has actually been previously analyzed and entered into the training set library), the unknown will be within three Mahalanobis distances of one of the groups in the library. The sample will be more than three Mahalanobis distances away from every other group in the library as long as the wavelengths used in the analysis have been chosen properly. If the instrument is not functioning correctly (for example, the spectrometer may have an abnormally high noise level) and wavelengths are not chosen properly, groups of spectra may be closer than six Mahalanobis distances. Such groups actually overlap and may allow misclassification of samples by the Mahalanobis metric. This overlap problem should be detected during the process of building the training set library and dealt with at that time. In the absence of prior knowledge about which wavelengths should be used in the Mahalanobis distance calculation, Mark suggests calculating the distances between all pairs of spectral data clusters and then forming the sum of the inverse squared Mahalanobis distances. The spectral clusters that are closest together contribute most heavily to this sum, and selecting the wavelengths that cause this sum to be smallest results in a selection that best separate the closest groups in space.

The size and shape of spectral data clusters in multidimensional space is influenced mainly by differences in the reflectivity of the sample, i.e., differences in particle size and packing. While different spectral clusters tend to have the same shape and orientation in hyperspace, samples differ in the ease with which they can be packed reproducibly in a sample cell. Mark suggests a normalizing of group sizes in which a large number of spectral data clusters are pooled in a training library to prevent false rejection of samples that are actually part of a large group and false assignment of unknown samples to a small group when the sample in fact lies beyond the statistical boundary of that group.

Mark notes that it is possible to find several wavelength combinations that yield approximately equivalent classification capabilities. The original results are obtained in proprietary samples that are not identified in his paper. The almost flawless performance of the Mahalanobis algorithm must be tempered by the lack of knowledge about the nature of the actual samples being studied. Mark also notes that small variations in the size, shape, or orientation between different spectral groups, as well as small changes in the physical nature of the samples and drift in the instrument, could all cause unknowns to appear to lie beyond the three standard deviation limit of their corresponding library entries. Due to these factors, Mark suggests that a more reasonable cutoff point in the Mahalanobis distances lies between 10 and 15 multidimensional standard deviations, rather than the 3 that theory would propose.

As in all heuristic methods, a calibration set that covers the range of variability of samples that will actually be analyzed must be employed to train the algorithm. Mark points out that even in the few cases where the Mahalanobis metric failed to properly classify a sample, the misclassified samples are not classified at all rather than classified incorrectly. Mahalanobis distances vary as a function of the dimension of the space in which they are calculated and as a function of the number of training samples used to calculate the distance. Thus, cutoffs for determining when a given sample belongs to a particular training set cluster should be allowed to "float" a small amount (the t-distribution, for example, can be used to adjust the three standard deviations limit on a training set upward a small amount to account for small sample sizes). Nevertheless, the use of Mahalanobis distances for classification permits computer programs to be written to drive

near-IR spectrometers that contain considerable protection for nonskilled operators.

A related publication discusses the use of normalized Mahalanobis distances in qualitative near-IR spectrophotometry.[29] Normalized Mahalanobis distances are calculated using the RMS group-size computed from the Mahalanobis distances of all spectra of a given material from the group mean of that material. The RMS group size normalization allows one to avoid problems created by different spectral cluster sizes in hyperspace. However, the assumptions of shape and orientation of spectral clusters made during discriminant analysis using Mahalanobis distances still must not be violated in order for qualitative analysis to be truly effective.

In the description of normalized Mahalanobis distances, 21 pure chemicals commonly encountered in the pharmaceutical industry are identified using the Mahalanobis metric. When a single spectrum is examined, the Mahalanobis distance used is the one corresponding to the group to which the single spectral distance is being measured. When the distance between groups of spectra are measured, distances are also normalized using the RMS group sizes corresponding to the group to which the distance is being measured.

Normalized Mahalanobis distances correspond better to the statistical criterion that says a Mahalanobis distance of 1 corresponds to 1 standard deviation of the data. When normalized Mahalanobis distances are used, a distance greater than 3 really means that a spectrum is not a member of the training group, unlike the ordinary Mahalanobis distance calculation. Many criteria must still be met to use normalized Mahalanobis distances:

1. All wavelengths in a spectrum cannot be used in the calculation because the covariance matrix would not be invertible.
2. There must be more samples than wavelengths in the qualitative calibration. (As is the case with other metrics, inclusion of more wavelengths in a qualitative discriminant calibration not only does a better job of separating the different spectral clusters in hyperspace from one another, but it also tends to increase the distances of readings from a group mean inside a particular cluster, increasing the possibility of failure to classify a sample. The RMS group size increases with the number of wavelengths.)
3. When using Mahalanobis distances, setting spectral cluster boundaries at 3 standard deviations is correct only when the number of groups to be distinguished and the number of wavelengths used to distinguish them are approximately the same.

Using the normalized Mahalanobis distances enables one to set the cluster boundary at 3 standard deviations without concern for the number of spectral clusters. The number of wavelengths must still be less than the number of training samples; however, the sampling distribution of spectral-cluster shapes and orientations changes as the number of wavelengths increases, and the probability of the groups matching in shape and orientation decreases as the number of wavelengths increases. Mark states that the sampling distribution of shapes and orientations for multidimensional spectral data is unknown at present. The use of more than 25 samples in a training set cluster to determine a reliable, normalized Mahalanobis distance is recommended. Using several different samples of each material in the training set, including different lots and different suppliers of the material, allows the data to "spread out" in multidimensional space. It is important to note that the larger the particle size, the greater the variability of the absorbance readings. While constructing a training set library for qualitative spectral analysis, normalized Mahalanobis distances are indicated when some spectral clusters show more spatial dispersion than others.

Mark uses Mahalanobis distances to evaluate sample preparation methods for near-IR spectrophotometry in a 1987 report.[30] In this paper, root-mean-square group sizes computed from the Mahalanobis distances are used as a means of determining (before a calibration) the sample preparation technique to use for near-IR analysis. The principal advantage of this procedure is that it can be applied to optical data alone without the need for a reference laboratory analysis. Sample-

grinding represents the greatest problem in the near-IR analysis of solid materials. When the reflectance of a solid is measured, the measurements vary depending upon the presentation of the surface grains of the specimen. Coarse sample grinds give greater variation, appearing in the spectra with greater variability in the baseline and peak heights. Mark demonstrated this by grinding solid samples in a coffee grinder and a cyclotec grinder. The cyclotec grinder is known to produce a much more uniform grind than a coffee grinder. When these spectra are represented as groups of points in a multidimensional space, the groups representing the more reproducible grinds appear as clusters of points smaller in size and closer to the origin than the less reproducible spectral group. The normalization of Mahalanobis distances through RMS group size provides a means of determining the reproducibility of sample grinds.

Mark evaluates different methods of preparing beef samples for near-IR spectrophotometry using the RMS group size of near-IR spectral data clusters. Eight different methods of preparing the beef samples are analyzed, ranging from grinding the meat in a meat grinder at room temperature to running it in a blender with dry ice or grinding it in a food processor. Ten to 15 repacks of each sample preparation method are used to prepare spectral data clusters for analysis. Choice of analytical wavelengths is not as critical when only the RMS group size is desired for a sample preparation. In practical cases, one would select the sample preparation that provides the smallest RMS group size, and therefore the greatest reproducibility in presentation to the spectrometer. Mark notes that repack variation is different for different components of the beef samples, i.e., fat and protein.

The Upjohn Company has received U.S. Food and Drug Administration approval for a near-IR spectrophotometric method using Mahalanobis distances for qualitative analysis.[31] Upjohn uses near-IR methods for the analysis of an antibiotic (lincomycin) in an agricultural feed premix. This feed mixture contains 86% soybean meal, 9% water, 1% mineral oil, and 4% lincomycin. The near-IR method is based first on qualifying the spectrum for application of a secondary quantitative equation. A given unknown sample spectrum must be within a certain Mahalanobis distance of a training set in order to qualify for a prediction by the quantitative prediction equation belonging to that training set. A Mahalanobis distance criterion of three (for deciding whether an unknown belongs to a specified training set) may be too restrictive, depending on the sample size and number of wavelengths being used. As mentioned earlier, this problem arises because of the dependence of Mahalanobis distances on the number of training samples used and the number of wavelengths examined. Because Mahalanobis distances are not a full spectral technique, a certain subset of wavelengths must be chosen for examination, increasing the possibility of missing absorbance peaks of an unknown sample constituent.

In theory, a calibration set should be composed of samples that represent the range of concentrations of an analyte expected in unknown samples. The authors[31] report, however, that the range of samples available from the premix process is too narrow to permit development of an adequate prediction equation. Therefore, rather than basing the training set on random samples obtained from the process, the training set is composed of samples prepared in the laboratory as well as manufactured samples. The manufactured samples are selected using a calibration equation generated from laboratory standard samples. The combined manufactured sample and laboratory-prepared sample sets are then used to generate a second quantitative prediction equation. Lincomycin is then predicted by a two-wavelength equation obtained using a "best combination" algorithm. The standard error of estimate (SEE) for lincomycin is 1.65% and the multiple correlation coefficient is 0.9424. (This is the prediction equation developed only on the laboratory-prepared samples.) An optimized four-wavelength prediction equation for lincomycin gives a standard error of estimate of 1.73% and a standard error of performance (SEP) of 1.45%. The multiple correlation coefficient is 0.9530. This calibration is developed from a combined set of laboratory-prepared samples and manufactured samples. The fact that the SEP is less than the SEE is attributed to the difference in concentration range of the calibration set and validation set. The four wavelengths selected for the

optimized calibration equation are 1422, 1478, 1504, and 1684 nm. Lincomycin has peaks at 1422, 1478, and 1684 nm as well as a valley at 1504 nm. The valley is used for baseline determination.

The four-wavelength optimized prediction equation does respond to lincomycin content, but the precision is inadequate to determine lincomycin with the required confidence. Discriminant analysis using Mahalanobis distances is used to differentiate among those spectra that produce large residuals in the quantitative prediction equation and those that produce small residuals. Those spectra producing residuals that exceeded approximately 5% of the nominal concentration determined by the reference method are classified as either bad-low or bad-high, depending on the direction of the residual. Unfortunately, using two- and three-wavelength Mahalanobis models, consistent classification of good and bad spectra is not achieved. Solid samples are often rotated a number of times, with spectra obtained after each rotation, in order to average the effects of sample surface inhomogeneity. The Mahalanobis distance discriminant procedure can also be used to select sample spectra for calibration after rotation of the sample holder. A specification can be set in software to allow a certain number of spectra to be collected in the rotating/averaging process. The spectrometer can then be forced to continue collecting spectra, rotating the sample, and testing the Mahalanobis distance after each rotation until the required number of spectra within a certain Mahalanobis distance of the center of the calibration set is obtained. Samples excluded from the prediction model by the Mahalanobis test must be analyzed by the reference procedure until a population can be found that describes these samples. This new population can be used to develop another near-IR prediction model.

Ciurczak and Maldacker use spectral subtraction, spectral reconstruction, and discriminant analysis to identify active ingredients in multicomponent pharmaceutical dosage forms using near-IR spectrophotometry.[32] Among these methods, the most common for identifying active ingredients in pharmaceuticals is spectral subtraction. In this experiment, a tablet composed of aspirin, caffeine, and butalbital is prepared for scanning by grinding. The spectrum of inactive ingredients is subtracted from the spectrum of a total ground tablet to produce a spectrum of an active ingredient. This subtraction method can introduce artifacts into the spectra. Spectral reconstruction by cross-correlation is also used to identify active ingredients in the same ground tablet. Artifacts are fewer in cross-correlation than in simple spectral subtraction to recover constituent spectra. The last method employed to identify the presence of active ingredients in tablet powder is discriminant analysis using Mahalanobis distances. Three wavelengths are chosen (following 7 h of computer time) for the initial data analysis: 2258, 2314, and 2290 nm. Thirty-four samples are scanned at 701 wavelengths and the computer program is allowed to determine the best three wavelengths for analysis from this spectral set. Using discriminant analysis, tablet mixes containing all active ingredients and containing all ingredients except the aspirin, caffeine, and butalbital are identified. Asymmetry is noted in spectral clusters, especially in the clusters containing all ingredients and those containing no caffeine. More advanced discriminant analysis techniques are necessary to deal properly with this asymmetry.

A discriminant analysis method that is capable of dealing with spectral cluster asymmetry is described by Lodder and Hieftje.[33] The quantile BEAST (Bootstrap Error-Adjusted Single-sample Technique) is proposed as a method of detecting samples that contain a component not present in the training set. The BEAST produces a distance in multidimensional standard deviations that is equivalent to the Mahalanobis distance when the training set is normally distributed; however, the BEAST metric is able to adjust for skew in the training set and makes no assumptions about the shape, size, skew, or orientation of spectral clusters in hyperspace. Hence, the BEAST is ideally suited for answering the question, "Does the prediction equation apply to the current sample?" Furthermore, the BEAST metric can be used directly in quantitative analysis. The distance between a new sample spectrum and the center of the training set in BEAST standard deviations is proportional to the concentration of the analyte responsible for the displacement.

Many circumstances can lead to asymmetry

in spectral clusters in space. In general, pure components project as symmetric elliptical clusters in near-IR spectral data hyperspace. Complex mixtures, however, can appear as clusters with different shapes, particularly when components vary over a wide range. Biological samples are especially prone to producing asymmetric spectral data clusters. For example, the distribution of blood serum spectra in hyperspace is skewed in part because the triglyceride distribution in blood serum is skewed. Because Beer's Law is a monotone transformation of the underlying component concentration distribution, the spectral data cluster is skewed in the direction of the triglyceride spectrum.

Discriminant analysis involves several assumptions, including (1) that no discriminating variable is a linear combination of other discriminating variables; (2) that the covariance matrices for all spectral groups are approximately equal (unless special formulas are used); and (3) that each group has been drawn from a population that is normally distributed on the discriminating variables. Violations of these assumptions reduce the efficiency of discriminant analysis and increase the probability of misclassification of samples. For these reasons, the quantile BEAST often offers superior performance to discriminant analysis using Mahalanobis distances. The BEAST constructs a multidimensional spectral cluster in hyperspace using the reflectance of each training set sample at a number of wavelengths. New samples are projected into the same multidimensional space and a nonparametric confidence test is performed to determine whether the new sample is part of the training-set spectral cluster. The BEAST method makes qualitative identification of samples possible, permits quantitative analysis, and detects samples that are outside the domain of the training set.

The BEAST method is used to differentiate among four similar benzoic acid derivatives: benzoic acid, isophthalic acid, salicylic acid, and *p*-aminobenzoic acid. The method is calibrated with a set of 40 mixtures of the benzoic acid derivatives. The samples are examined at three selected wavelengths, and the BEAST is presented with the spectra of pure benzoic acid derivatives and other compounds to determine whether it detects these samples as outside the domain of the training set. The training mixtures range in concentration from 0 to 25% of each of the four benzoic acid derivatives (by weight). The remainder of the samples are made up of an aluminum oxide diluent. Component concentrations for each of the benzoic acid derivatives are selected by a random mixing algorithm. The BEAST distances for some compounds measured in standard deviations from the training set of benzoic acid derivatives include: (1) acetylsalicylic acid, which appears at 8.20 standard deviations from the mixture training set; (2) dextrose, which appears at 46.33 standard deviations from the mixture training set; (3) whole wheat flour, which appears at 12.71 standard deviations from the center of the mixture training set; and (4) sucrose, which appears at 3.88 standard deviations from the center of the mixture training set. In the final test of the BEAST algorithm, 20 mixture samples inside the domain of the training set are contaminated with a false-sample compound to determine whether they could be detected properly as being beyond the domain of the original training set. The contaminating compound is acetylsalicylic acid, ranging in concentration from 1 to 20%. The results of the studies performed using pure benzoic acid derivatives demonstrate perfect separation of the derivatives in an 18-dimensional space. Benzoic and salicylic acids are the closest spectral clusters, differing by only a single oxygen atom.

A large number of sample components, varying over a large concentration range, can have the effect of filling the analytical spectral hyperspace created by only a few wavelengths. Three wavelengths are used to separate acetylsalicylic acid, dextrose, whole wheat flour, and sucrose. Even though the Euclidean distances between these compounds (distances calculated from a pure reagent training set and a training set composed of mixtures of all the benzoic acids) are about the same, the distances in standard deviations between these four compounds and the two different types of training sets are quite different. This change occurs because the volume of a training set composed of mixtures in three-dimensional space is about 100 times the volume of a pure reagent training set in three-wavelength space. The increase in cluster size for training sets composed of mixtures of compounds high-

lights the need to be aware of the exact shape of the training-set envelope when small distances are to be classified as representing valid or misclassified samples. When mixtures that differ by only a few components are used to construct training sets, those training sets are likely to overlap as clusters in hyperspace. Overlap of clusters (see Figure 1) complicates the problem of assigning a new sample spectrum to a single training set.

A number of theoretical studies have been undertaken using the BEAST to examine synthetic samples generated by a computer. The parameters that affect the performance of the algorithm in experimental situations include: (1) the number of wavelengths used in the analysis; (2) the number of samples in the training set; (3) the selected radius of the hypercylinder (see Figure 2); and (4) the number of bootstrap replications in the training set employed. Using synthetic data randomly drawn from a multivariate normal population with a known group mean (center) and a known variance in all directions, a selected synthetic spectral point can be assigned both a true distance from the training set (given by the Euclidean distance divided by the variance of the training set in the direction of the synthetic sample spectral point) and an experimental distance (determined by the BEAST algorithm). The training set is formed by Monte Carlo integration of multivariate normal populations. The bias and mean square error of the BEAST as a point estimator of the directional variance of the training set is determined for particular combinations of the experimental parameters (number of wavelengths, training set size, hypercylinder radius, and number of bootstrap replications). In these theoretical studies, ten runs are made with each combination of parameters (1, 2, 3, and 5 wavelengths; training-set sizes from 10 to 200 samples; hypercylinder radii from 0.001 to 0.1 absorbance unit; and 50, 200, 1,000, and 10,000 bootstrap replications).

The bias (accuracy) of the BEAST estimator appears to be most strongly affected by the number of training set samples employed to establish the location and size of the training set in spectral hyperspace. The bias tends to be large (about 28%) when only 10 training samples are used; however, the bias drops steadily to under 1%

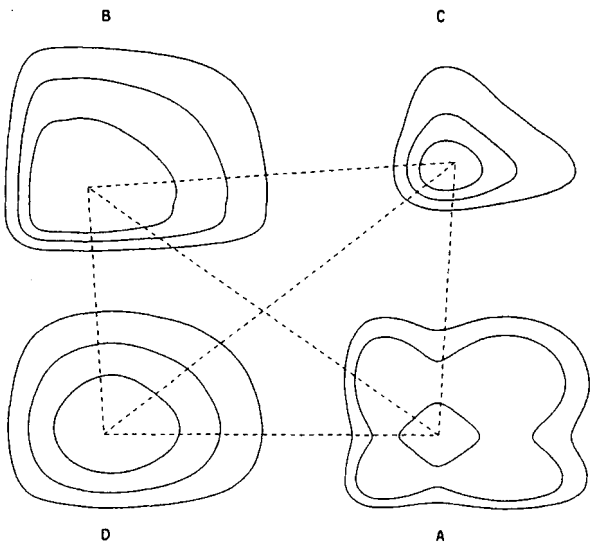

FIGURE 1. A projection of four hypothetical spectral clusters in hyperspace onto a plane.

when the training-set size reaches 200 samples. (These results are obtained using 1000 bootstrap replications, 2 wavelengths, and a hypercylinder radius of 0.001). The results of the tests for training-set size suggest that the prudent analyst will use as many training samples as possible, a result that is not startling in light of the requirements of other near-IR pattern-recognition techniques. The precision of the BEAST estimator also appears to be related to the size of the training set, although not as strongly. The relative standard deviation (RSD) of the method seems to be about 5% for training-set sizes between 10 and 200 samples when 1000 bootstrap replications, 2 wavelengths, and a hypercylinder radius of 0.001 are employed.

The number of bootstrap replications of the training set appears to be the most influential factor governing the RSD (precision) of the BEAST estimator of directional variance. The RSD drops rapidly as more bootstrap replications are performed. Using only 50 bootstrap replications of the training set can cause RSDs as high as 80% and biases as high as 25% in the BEAST estimator of variance. The RSD also tends to be a function of the number of dimensions in the analytical hyperspace. The use of more wavelengths in the analysis tends to demand the use of more replications of the training set to achieve a given RSD. The BEAST uses the bootstrap

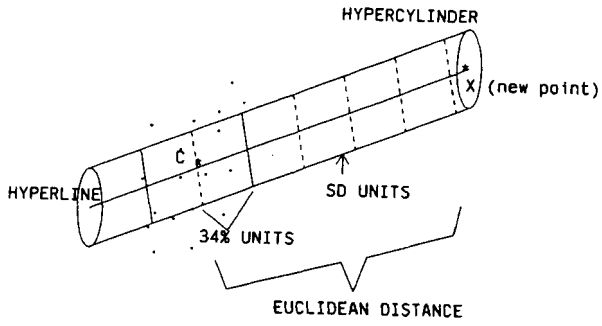

FIGURE 2. Constructing an analytical hypercylinder through the center of bootstrap replicates and a test spectral point.

replicate distribution to approximate a real sample distribution, and because the volume of the analytical space that must be described by these replicates increases as the number of wavelengths increases, more replicate points are required to describe a training set with more dimensions. The number of replications of the training set must always be large enough to produce an adequate number of spectral points in the analytical hypercylinder. Generally, at least 50 spectral data points in the hypercylinder are required to establish a directional variance. To a certain extent, the number of points in the hypercylinder can be increased by increasing the radius of the hypercylinder. However, this approach eventually results in a loss of directional selectivity that begins to bias the quantiles of the data in the hypercylinder, leading to a corresponding bias in the BEAST estimator of directional variance. The bias of the BEAST is not affected very strongly by the number of bootstrap replications of the training set. In the tests conducted in the paper,[33] increasing the radius of the hypercylinder decreases both the bias and RSD of the BEAST estimator of directional variance. However, the radius is never increased to the point that biasing of quantiles became a problem. (In fact, when the covariance between wavelengths is nearly zero, biasing of quantiles by increasing the radius of the hypercylinder is almost never a problem.)

Like many nonparametric methods, the BEAST method is based on a very simple procedure iterated a large number of times. The algorithm is easily vectorized and parallelized. The BEAST is useful in the analysis of pharmaceuticals, biological mixtures, and other complex samples.

Another bootstrap-based method, an extension of the BEAST technique, can be used to determine when a population of test samples is not the same as a calibration set of samples (as opposed to the ordinary BEAST procedure which tests a single sample to determine whether it comes from a particular calibration-set population). In typical near-IR statistical analyses, samples with similar spectra project as points in spectral hyperspace that cluster in certain regions. These clusters can vary in shape and size due to variations in particle size distributions, component concentrations, and drift factors. In such circumstances, discriminant analysis using simple distance metrics can produce a situation in which a discriminant result that places a particular point inside a particular cluster does not necessarily indicate that spectral point as an actual member of the cluster. Instead, the spectral point may be a member of a new, slightly different spectral cluster that overlaps the first. A new cluster can be created by factors such as low-level contamination. A bootstrap procedure (a variant of the BEAST algorithm) can be used to set nonparametric probability density contours inside as well as outside spectral clusters. When multiple spectral points formed by new samples begin to appear in a certain region of hyperspace, the perturbation of the training-set density contours can be detected and assigned a significance level. The detection of misclassified samples using this method is possible both within and beyond the three standard deviation limit commonly used to denote the surface of a training-set cluster. This density-contour procedure is able to detect contaminant levels of a few hundred parts per million in an intact over-the-counter drug capsule (see Figure 3). Furthermore, the method functions with as few as one or two wavelengths in the near-IR region, suggesting its possible application to very simple process sensors.

A major difference between the BEAST discriminant analysis and the Mahalanobis metrics is that a BEAST standard deviation can be symmetric or asymmetric. The Mahalanobis metric has often been referred to as a "rubber yardstick" whose length in hyperspace depends upon the orientation of the yardstick. The stretch of the

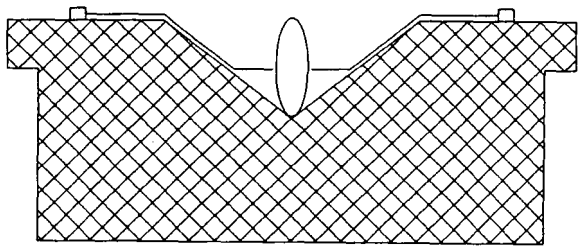

FIGURE 3. A conical reflector used for trace contaminant determinations in intact pharmaceutical capsules.

Mahalanobis distance is symmetric (i.e., grasping the yardstick at both ends and pulling produces identical increases in the length of the upper and lower halves of the stick). On the other hand, the stretches of the upper and lower halves of the BEAST yardstick are not necessarily equal and depend, in fact, upon the skew of the training set in the direction in which the yardstick is oriented. The BEAST metric can be thought of as a rubber yardstick held down by a nail located at the center of the calibration set. The presence of the nail enables the ends of the yardstick to stretch independently.

The ability to detect subpopulations of spectra inside the three standard deviation limit of a training set is what makes trace near-IR analyses possible. For this paper the authors tested the trace analysis capability of near-IR spectrometry and the subpopulation detection algorithm in pharmaceutical capsules containing various trace contaminants (such as aluminum dust and floor sweepings). Validation sets are prepared along with calibration sets of capsules by emptying all of the powder from over-the-counter capsules and repacking both the calibration-set and the validation-set capsules. In this way the effects of capsule repack on the final determination of contamination are minimized. Test-set capsules are packed with a small amount of contaminant mixed into the pool of capsule contents (powder). Three types of relationships between the training set and test set spectral clusters are explored: (1) the effects of pure location differences between the training set and test set; (2) the effects of pure scale or size differences between the training and test sets; and (3) the effects of simultaneous differences in location and scale between the training and test sets in hyperspace. The data analysis procedure includes the comparison of two integrals. The first integral is calculated from the training set by beginning the integration at the center of the training set and integrating outward in all directions at a constant rate until all of the training-set samples are included in the volume of the training-set integral. The second integral is formed by combining the training-set and test-set spectra. This second integral also begins at the center of the training set, encompassing both training- and test-set spectra as the integration is carried out at a constant rate, until all of the training-set and test-set spectra have been included. The correlation between these two integrals is used to determine whether the training-set and test-set spectral clusters are actually the same. High correlations between the two integrals indicate that the training and test sets are likely to be the same.

In addition to experiments conducted on contaminated capsules, experiments are conducted to determine the theoretical bias and RSD for the subpopulation detection procedure as a function of the number of training samples used, the number of bootstrap replications of the training set employed, and the number of wavelengths monitored in the spectra of the samples (the dimension of the hyperspace). Synthetic computer-generated data are used for both the training and test sets in these theoretical tests. When the test-set and training-set spectral clusters share the same size and shape, the location difference between the two spectral clusters can be determined by the subpopulation detection method when the centers of the two clusters were only one standard deviation apart. When the test set and training set share the same center but the test set is smaller than the training set, the difference between the two sets can be detected only when the test set is smaller than the training set by a factor of about two. When the test set is larger than the training set, the difference between the two can be detected when the test set is larger than the training set by slightly more than a factor of two. When simultaneous differences in the location and scale of the test set and training set are observed, the difference between the two is significant when the difference between centers of the two sets is as little as 0.2 standard deviations. Bias and RSD

values of <3% are attainable with fewer than 20 training-set and test-set samples. The bias and RSD of the subpopulation detection method are found to be largely independent of the number of bootstrap replications, at least when this number is >50. Bias and RSD are found to increase slowly as the number of spatial dimensions (e.g., wavelengths) is increased. In a five-dimensional space, however, the bias and RSD of the correlation coefficient between the two integrals is still <1%.

Near-IR spectrophotometry and the subpopulation detection algorithm are able to detect an average concentration of 296 ppm aluminum in intact capsules. Furthermore, the presence of a small amount (221 ppm) of floor sweepings in a set of intact pharmaceutical capsules is also detected by this method. It is clear that the method is able to detect misclassified samples inside the three standard deviation limit normally put on a training-set spectral cluster when multiple test samples are available. Thus, near-IR spectrophotometric analyses can be successfully carried out for constituents present in samples at low concentrations using only a small number of wavelengths (perhaps even one or two).

V. QUANTITATIVE THEORY AND CHEMOMETRICS

The effects of an absorbing sample matrix on near-IR diffuse reflectance spectra have been discussed in a quantitative context.[34] Many workers have reported that log 1/R values provide a more linear relationship to concentration in the near-IR region than Kubelka-Munk transformation of spectra. Olinger and Griffiths investigate this relationship in more depth. Quantitative spectrometric analysis relies upon a linear relationship between absorption or reflectance intensity and concentration. Many transformations have been investigated as a means of linearizing the relationship between spectra and analyte concentration. Kubelka-Munk and log 1/R transforms, however, represent the most commonly employed transformations. The Kubelka-Munk equation is used commonly in the UV, visible, near-IR, and far-IR regions for diffuse reflectance measurements.

The Kubelka-Munk equation includes a scattering coefficient that corrects for scattering in the absence of absorption. Nevertheless, since the early days of near-IR spectrophotometric analysis, almost all results have been reported as functions of log 1/R values rather than as functions of the Kubelka-Munk relationship. The Kubelka-Munk theory is meant to provide a quantitative description of absorption, reflection, and scattering of radiation within a sample. Beer's Law assumes that reflection and scattering of light within the sample are insignificant compared to absorption. Olinger and Griffiths pointed out that under such assumptions it is surprising that the Kubelka-Munk relationship is not more linear with concentration than the log 1/R relationship.

The Kubelka-Munk equation makes a number of assumptions, including:

1. The sample is illuminated with monochromatic radiation (this assumption is also made with Beer's Law).
2. The distribution of scattered light is uniform in all directions and specular reflection is ignored.
3. The particles in the sample layer are randomly distributed.
4. The particles are much smaller than the thickness of the sample layer.
5. The sample layer is subjected to only diffuse illumination.
6. The particles are much larger than the wavelength of the illuminating light (this makes the scattering coefficient independent of wavelength).
7. Edge effects are minimal because the sample is larger than the illuminating beam.

The Kubelka-Munk equation is a limiting equation, like Beer's Law, and only really applies to weak absorbance bands. Because molecules typically exhibit low molar absorptivity in the near-IR region, the Kubelka-Munk theory might be expected to apply even more strongly in the near-IR region.

In the near-IR region, however, the analyte is usually not separated from the sample matrix for analysis, and thus the sample matrix is generally absorbing about as much light as the analyte itself. Olinger and Griffiths' study was per-

formed on solid analytes in the near-IR region in both an absorbing sample matrix and a nonabsorbing matrix. Sodium chloride is used as the nonabsorbing sample matrix, while graphite is used as the absorbing matrix. Carbazole is selected as a typical organic analyte because it contains both carbon-hydrogen and nitrogen-hydrogen bonds. Quantitative training sets of binary mixtures of carbazole and sodium chloride are used to represent analysis in a nonabsorbing matrix, while analyses in the absorbing matrix are performed on mixtures of carbazole in sodium chloride with either a 1 or 5% graphite (by weight) added to the matrix to increase matrix absorption. The particle size of the samples ranged from 18 to 53 μm.

Absorption of light by the added graphite increases as particle size increases. Graphite particles have very high absorptivities in the near-IR region. Quantification of carbazole is performed using the carbazole band at 1672 nm. This band is very intense and is surrounded by a fairly flat (by near-IR standards) baseline. The Kubelka-Munk relationship is found to produce more linear plots with concentration for the pure sodium chloride (nonabsorbing) matrix. Log 1/R values produced more linear plots with concentration for the absorbing matrices with both 1 and 5% graphite. However, the 1% graphite in sodium chloride matrix is more linear than the 5% graphite in the same matrix.

The effects of referencing the recorded spectra to both a ceramic disk and a reference composed of the sample matrix without the carbazole analyte present are examined. The blanks are composed of the same chloride and sodium chloride-graphite mixtures used in the transformation studies. When the sample is 8% (by weight) carbazole in a 5% (by weight) graphite-NaCl matrix, the best baseline is obtained by referencing the spectra to a 5% by weight graphite-NaCl matrix. When referencing to a graphite-NaCl mixture of different weight-percent graphite, a graphite/NaCl mix provides a better baseline correction than referencing to NaCl alone. In the long wavelength region where baseline curvature occurs, one must be careful to compensate properly for the curvature when referencing against the sample matrix.

The effect of particle size on spectra is also discussed. Noise level increases with increased particle size, and the linear dynamic range is observed to decrease with increasing particle size. The increase in noise level is described as being due to larger particles giving a greater Beer's Law absorption per particle. The S/N ratio of the spectrum decreases as particle size increases. Adherence to Beer's Law occurs as a result of the fact that the ratio between the surface area illuminated by the beam in near-IR instruments and the depth of penetration is very large. Consequently, the composition of the surface layer (the ratio of analyte to matrix in the surface) determines the response at each wavelength.

When the analyte is present at low concentrations in a nonabsorbing matrix, the penetration depth of light into the sample can be as great as 1 mm. In most powders, however, the true penetration depth is about 300 μm.

In most applications of near-IR reflectance spectrometry, the sample matrix absorbs at all wavelengths in an amount that is similar to the absorption of the analyte. Thus, the requirements of the Kubelka-Munk theory are not met, and the log 1/R transformation of spectra is a more effective means of extracting quantitative information.

The meaning of the standard deviation in the reporting of the performance of a near-IR calibration has been described in the context of protein analysis in flour samples.[35] The typical repeatability (RSD) of Kjeldahl protein measurements on flour is about 0.15%. The near-IR method, however, is even more reproducible than the Kjeldahl result, with RSD values being typically about 0.05%. The mean measurements of repeated Kjeldahl and near-IR tests will converge to true values. These values, however, will not necessarily be identical. The difference between these two values often arises from the fitting process. The proper selection of calibration samples is critical to developing a calibration equation that will fit real samples routinely with an acceptable error. For example, if all the flour samples in a calibration set have 13% moisture, then no information exists in the calibration set about the dependence of near-IR spectra of flour on moisture. The standard error of estimate developed on a calibration set generally underestimates the error that will actually occur when

the calibration is used. Three reasons exist for this underestimate: (1) the standard deviation is almost always an underestimate of the real errors involved; (2) uncertainty exists concerning the actual values of the calibration constants and this uncertainty is larger when the number of samples in the training set is small; and (3) the entire calibration process depends entirely on the selected calibration set. If the chosen calibration set contains a chance correlation between two constituents that are not present in the population as a whole, the calibration equation may use these features to obtain an anomalous low standard deviation of estimation. Cross-validation is usually employed to detect these chance correlations. Cross-validation by data splitting or by leave-one-out methods are employed. In many instances cross-validation by data splitting provides a more robust estimate of the standard deviation of prediction that can be expected in real situations.

General principles for the selection of samples for use in the calibration sets in near-IR spectrophotometry have been discussed.[36] The work of Naes and Isaksson concerns the selection of samples for calibration using the general multivariate linear model. Regression techniques are employed to calculate the coefficients of the linear model.

Two basic principles should be followed: (1) all types of possible combinations of the independent variables must be present in the training set and (2) the independent variables must span the whole variation space. Satisfying these two principles enables one to obtain regression coefficients that provide adequate predictions over the entire range of interest. In real samples, the linear model often does not give an equally good fit over the entire range of analytical interest. The authors propose that the independent variables should be evenly spread over the range of their variation to allow fit to higher order (nonlinear) terms.

Many chemometric procedures are problematic when investigators are dealing with the situation in which there are more wavelengths available to describe samples than there are samples available for modeling. In these cases, principal axis transformation (PAT) techniques can be used to compress the data. PAT techniques eliminate the collinearity problem that arises in multiple linear regression and in discriminant analysis using Mahalanobis distances. The training-set requirements given above also apply to PAT space.

Factorial designs are proposed as a method of making certain that each independent variable is properly spanned. The use of multiplicative scatter correction to reduce differences between samples due to light scattering effects eases the spanning requirement. The authors[36] developed five different calibration sets from experimental data. In set 1, the samples are evenly spread over the whole experimental region using a Simplex-lattice structure. In set 2, the samples are selected from the corners of a lattice structure with a large empty region in the center. In set 3, samples are randomly selected from the whole experimental region in a way meant to simulate the natural population. In set 4, samples are selected from a limited region in the center of the experimental space. In set 5, the samples are selected to span only the variation of a particular analyte. The worst predictive results are observed with the fifth experimental design. The fourth design performs adequately as long as prediction samples are also in the middle of the region to be spanned; however, when the prediction samples move outside this region the fourth experimental design becomes inadequate. The authors conclude that designs 4 and 5 are useless when the prediction samples can be expected to cover the whole experimental region. The design with the even spread of points throughout the experimental space generally produces the best results. Both the evenly spaced design and the simulated natural population produce overall acceptable results. The principal differences between the two methods are found in their performance in the middle of the experimental region vs. their performance at the edges. Training sets developed with samples evenly spread over the entire experimental region offer similar performance in terms of RSD over the entire region; however, training sets developed with the randomly selected simulated natural population tend to perform better than the evenly spread samples in the middle of the prediction region.

Another sample selection method that is based on near-IR spectral subtraction techniques is de-

scribed by Honigs et al.[37] The spectral subtraction methods uses linear algebraic techniques to select spectrally unique samples from a large pool of potential candidates for inclusion in a near-IR training set. The method improves calibrations by developing a training set and a calibration that are more resistant to unexpected variations in the sample matrix, and by reducing the size of the training set, which reduces the number of reference analyses that must be performed.

The most important contribution of the linear algebraic sample-selection technique is that it makes near-IR spectroscopy the sample selection procedure rather than some complex laboratory reference procedure that would ordinarily be used to determine whether the samples evenly spanned the experimental region of interest. The speed of the near-IR technique makes it possible to rapidly screen a large number of samples and select a small portion for reference analysis, reducing the problem of developing an appropriate calibration set.

In the early days of near-IR reflectance analysis, a shortage of computing power made multiple linear regression a difficult process to employ in calibration.[38] Honigs et al. proposed solving the system of linear equations formed by the log 1/R values collected at a number of wavelengths by Gauss-Jordan linear algebra. The resulting algorithm was termed the row-reduction method. At the time, wavelength selection was commonly employed in calibration, using such schemes as all possible combinations of regressions, step-up regression, and step-down regression. These methods required a large number of iterations to select analytical wavelengths, and each iteration incorporated a multiple linear regression. The iteration step that produced the largest multiple correlation coefficient was designated the best, and the wavelengths used in that iteration were then used in the resulting calibration equation. The relatively small microcomputers available during that period were hard-pressed to calculate the number of multiple linear regressions required, leading to the development of the row-reduction algorithm as a means of rapidly approximating the results of multiple linear regression. Improvements in computation time ranging from a few percent up to a factor of 200 were reported using the row-reduction algorithm.

The row-reduction method is demonstrated with simulated spectra, spectra of solutions of methyl red and methyl orange, and the determination of protein in wheat. It is also used to analyze benzene and hydrocarbon mixtures.

Increased immunity to baseline drift and to overfitting (overfitting occurs when too many terms are used to develop a sample spectral correlation to concentration, leading to anomalous high correlations) are reported with the row-reduction algorithm.

The use of Fourier analysis to enhance near-IR diffuse reflectance spectrometry and quantitative analysis is reported by McClure et al.[39] Fourier coefficients for near-IR spectra of powdered agricultural products are computed and these coefficients are used to recalculate spectra. As few as 11 coefficients from the Fourier domain produce usable calibrations with stepwise multiple linear regression. The Fourier model greatly reduces the number of wavelengths that must be included in the stepwise process, thereby reducing the computation time for calibration by 96% and the storage space for spectral data by 98%. Removing the mean term from the Fourier model enables partial correction for particle-size effects encountered in solid samples. Most of the spectral information related to the chemical constituents of samples resides in the low-frequency components of the spectra, perhaps explaining why filter instruments containing 10 to 20 interference filters have found such broad applicability in near-IR reflectance analysis. FT-near-IR instruments make it possible to obtain Fourier coefficients faster and more directly, speeding the near-IR analysis process.

An assessment of the number of samples and wavelengths required to build a usable calibration set for near-IR spectroscopic analysis was published in 1984.[40] In the near-IR calibration process, a balance must be struck between having enough samples and wavelengths in the training set and having too many. The standard error of estimate and the standard error of performance calculated during cross-validation procedures provide a means for determining the optimum number of training samples and analytical wavelengths. Using too few wavelengths produces large prediction errors. On the other hand, the use of too many training samples increases the

number of reference analyses that must be performed to provide an information vector for the calibration process. Furthermore, the use of an insufficient number of wavelengths leads to analyses that are highly susceptible to changes in the sample matrix, and the use of too many wavelengths leads to calibrations that are overfitted and susceptible to noise. The calibration function is ordinarily calculated as the solution of a system of linear equations, and the system becomes indeterminate when the number of wavelengths used is greater than the number of apparent spectral constituents. In a practical sense, it is often best to use the smallest possible number of wavelengths; however, it is necessary to have at least as many wavelengths in the calibrations as independently varying contributions to the spectra. As the number of wavelengths is increased, the probability of encountering additional spectral interferences also increases. In addition, each successive wavelength provides a progressively smaller correction to the overall result, while adding a nearly constant amount of measurement noise. It is this measurement noise that causes the problem with overfitting (which appears as a small standard error of estimate and a large standard error of performance). As a general rule, a calibration works best when the number of wavelengths employed is approximately equal to the number of varying sample constituents.

When the noise level approaches the level of spectral contribution of the smallest spectral component, the noise level acts like an infinite number of additional independently varying spectral constituents. Once this noise level has been reached, the number of wavelengths used can never be matched by the number of spectral constituents. The authors[40] provide a graph that depicts the ratio of the standard error of estimate divided by the correlation coefficient vs. the number of samples. This graph can be used to determine the significance of a particular correlation.

The repacking of solid samples in closed cups for near-IR analysis introduces its own errors into near-IR calibrations.[41] The total variance (or error) generated by the prediction equation when applied to near-IR spectra is equal to the sum of the errors introduced by the reference laboratory method, the sampling error, the instrumental error (noise), and the repack error. The effects of repack error can be assessed by repeatedly reloading the same sample into the same sample cup and measuring the absorbance readings across the spectrum for each repack. The average of each set of repeated readings is certainly more stable than the individual readings. For this reason many near-IR calibrations are constructed using several repacks of the same sample. Calculating the amount of error due to repack enables one to estimate the improvement of results attainable by reducing the amount of repack error. In theoretical calculations it is shown that if the repack error measured by standard deviation is one half of the total error, then the potential reduction in total error by elimination of repack error is limited to a factor of 0.86. The repack standard deviation must be at least 70% of the total standard deviation in order to be able to reduce the total error to one half of its original value by elimination of the repack error component. In real situations, however, the improvement is less than theoretical calculations indicate. In the actual data, the spectra of breakfast cereal show the largest repack error. In this case, the fraction of variance due to repack was 78% of the total variance. The repack variance was reduced to 36% of the total by averaging successive repacks. If the calibration data are averaged, then the validation data and subsequent analyses must also be similarly averaged or the prediction results will be worse than if no averaging were performed at all.

In order to produce a calibration that is insensitive to a particular physical phenomenon, the phenomenon should be introduced into the calibration set over its broadest possible range of experimental values. This procedure will create a regression equation with biased coefficients that are insensitive to that phenomenon. In these cases, the results using this calibration to predict sample concentrations will be a little worse than the standard error of estimate obtained with a training set in which the physical phenomenon was held constant.

The effect of multiplicative scatter correction on the linearity of near-IR calibrations has been assessed on five different food products.[42] The multiplicative scatter correction was first pro-

posed by Martens and is based on linear regression of the variables in each spectrum against the average of the set of spectra. This procedure is used to correct simultaneously for multiplicative and additive scattering effects. In effect, this method takes the scatter level in each sample spectrum and adjusts it to the average scatter level in all of the samples. Thus, subsequent calibration methods, such as principal component regression or multiple linear regression, do not have to describe scattering effects as well as changes in sample constituents. This may not be an advantage if the scattering effects are correlated to a sample constituent or property of interest. In reported applications, the use of the multiplicative scatter correction produces improvements in correlations between spectra and analytes. Moreover, these improvements are greater in cases where the variation in the constituents of the samples is largest. The use of multiplicative scatter correction produces spectral data that conform more closely to Beer's Law.

Calibrating near-IR instruments when only a limited number of wavelengths are available (as is often the case with filter-based instruments) can be done with either local or global algorithms.[43] Osborne suggests 10 calibration samples for each term in the prediction equation (e.g., a linear equation with a single wavelength requires a minimum of 20 samples). A two-wavelength calibration with an intercept would require a minimum of 30 samples using this rule. A differentiation is made between "closed" and "open" populations. A closed population is one in which all the samples to be analyzed are available at the time of calibration. In an open population all samples are not available for examination. In describing open populations, a number of samples that encompass the range of expected variations in each sample constituent and sample property must be selected from known sources of error. For example, repack of solid samples must be included in the calibration samples to make the calibration resistant to error from the repack error source. Samples should be selected to include the known sources of error, but in a random fashion in order to prevent the accidental correlation of actual sample constituents to known error sources.

It is important to have good reference values (a good set of external standards) to calibrate the near-IR method. (This requirement is not peculiar to near-IR spectroscopy.) In general, samples should be presented in random order to both the reference and near-IR methods. This procedure reduces the possibility of drift being incorporated into the prediction model.

Particle size is often a problem with near-IR diffuse reflectance spectroscopy of solid samples. The multiplicative scatter correction, the use of first and second derivatives, and regression processes can all be used to reduce the effects of particle size variations on a given calibration.

Wavelength selection for instruments that record data at a small number of wavelengths is an important problem. In general, it is not a good idea to use multiple linear calibration equations that contain terms for each wavelength examined by the spectrometer. Forward stepwise regression and all-combinations searches are commonly used in near-IR spectroscopy as a means of reducing the number of wavelengths that are incorporated into the final model. In addition to stepwise regression and searching schemes, certain global calibration procedures such as principal component regression can also be applied to the limited number of wavelength vectors available on filter instruments.

The transfer of calibrations between filter instruments, using a limited number of wavelengths, is relatively simple and generally involves only the adjustment of a bias value that changes the intercept of the calibration equation. Manufacturers usually provide a simple procedure for transferring calibrations among their instruments. The adjustment of calibration equations between scanning instruments is more difficult and must take into account differences in wavelength accuracy and resolution of different instruments. The difference between detectors is also a significant factor in errors in the final calibration after transfer.

A tutorial on multiple linear regression (MLR), principal component (PCR), and partial least-squares (PLS) regressions was published by Geladi and Kowalski.[44] In the tutorial, PLS is shown to be a robust alternative to MLR and PCR. The theoretical foundation for PLS in terms of the singular value decomposition was derived

for the PLS algorithm by Lorber et al.[45] PLS is shown to be one of a continuum of variations of PCR in this article.[45] The similarity between the NIPALS algorithm and the power method for eigenvalues and eigenvectors is also described.

VI. SUMMARY

Near-IR spectroscopy is a rapid, nondestructive method of analysis that has proven to be relatively inexpensive when employed in large-scale analyses. Near-IR spectrometry offers greater safety than many alternative analyses because it eliminates the need for dangerous (and costly) solvents. Near-IR methods work with most solids or liquids and require little or no sample preparation in many cases. They are gaining broad acceptance in on-line or at-line applications as both quantitative and qualitative tests. New, non-invasive biological and clinical tests are always being performed by near-IR spectrometry, and additional discoveries are sure to follow.

REFERENCES

1. **Lodder, R. A. and Hieftje, G. M.,** *Appl. Spectrosc.*, 42, 1500, 1988.
2. **Buchanan, B. R. and Honigs, D. E.,** *Spectroscopy*, 1(7), 40, 1986.
3. **Watson, C. O.,** *Anal. Chem.*, 49(9), 835A, 1977.
4. **McDonald, R. F.,** *Anal. Chem.*, 58, 1906, 1986.
5. **Wetzel, D. L.,** *Anal. Chem.*, 55(12), 1165A, 1983.
6. **Wheeler, O. H.,** *J. Chem. Ed.*, 37, 234, 1960.
7. **McClure, W. F.,** *Anal. Proc.*, 21, 485, 1984.
8. **Tunnell, D. A.,** *Anal. Proc.*, 23, 299, 1986.
9. **Montalvo, J. G., Faughts, S. E., and Buco, S. M.,** *Am. Lab.*, 18(11), 37, 1986.
10. **Beebe, K. R. and Kowalski, B. R.,** *Anal. Chem.*, 59, 1007A, 1987.
11. **Callis, J. B., Illman, D. L., and Kowalski, B. R.,** *Anal. Chem.*, 59, 624A, 1987.
12. **Rotolo, T. K.,** *Cereal Foods World*, 24(3), 94, 1979.
13. **Shank, J. S., Landa, I., Hoover, M. R., and Westerhaus, M. O.,** *Crop Sci.*, 21, 355, 1981.
14. **Cooper, P. J.,** *Cereal Foods World*, 28, 241, 1983.
15. **Davies, A. M. C. and Grant, A.,** *Appl. Spectrosc.*, 41, 1248, 1987.
16. **Hirschfeld, T.,** *Appl. Spectrosc.*, 40, 1082, 1986.
17. **Lodder, R. A. and Hieftje, G. M.,** *Appl. Spectrosc.*, 42, 518, 1988.
18. **Yeo, P. L. and Honigs, D. E.,** *Appl. Spectrosc.*, 42, 1128, 1988.
19. **Ciurczak, E. W. and Weis, F. M. B.,** *Spectroscopy*, 2(10), 33, 1986.
20. **Linthicum, K. J., Bailey, C. L., Daves, F. G., and Tucker, C. J.,** *Science*, 235, 1656, 1987.
21. **Hallmark, V. M., Zimba, C. G., Swalen, J. D., and Rabolt, J. F.,** *Spectroscopy*, 2(6), 40, 1987.
22. **Waters, D. N.,** *Appl. Spectrosc.*, 41, 708, 1987.
23. **Ciurczak, E. W.,** *Spectroscopy*, 5(1), 10, 1990.
24. **Honigs, D. E., Hieftje, G. M., and Hirschfeld, T.,** *Appl. Spectrosc.*, 38, 317, 1984.
25. **Cowe, I. A. and McNicol, J. W.,** *Appl. Spectrosc.*, 39, 257, 1985.
26. **Puchwein, G.,** *Anal. Chem.*, 60, 569, 1988.
27. **Devaux, M. F., Bertrand, D., Robert, P., and Qannari, M.,** *Appl. Spectrosc.*, 42, 1015, 1988.
28. **Mark, H. L. and Tunnell, D.,** *Anal. Chem.*, 57, 1449, 1985.
29. **Mark, H.,** *Anal. Chem.*, 58, 379, 1986.
30. **Mark, H.,** *Anal. Chem.*, 59, 790, 1987.
31. **Whitfield, R. G., Gerger, M. E., and Sharp, R. L.,** *Appl. Spectrosc.*, 41, 1204, 1987.
32. **Ciurczak, E. W. and Maldacker, T. A.,** *Spectroscopy*, 1, 36, 1986.
33. **Lodder, R. A. and Hieftje, G. M.,** *Appl. Spectrosc.*, 42, 1351, 1988.
34. **Olinger, J. M. and Griffiths, P. R.,** *Anal. Chem.*, 60, 2427, 1988.
35. **Fearn, T.,** *Anal. Proc.*, 23, 123, 1986.
36. **Naes, T. and Isaksson, T.,** *Appl. Spectrosc.*, 43, 328, 1989.
37. **Honigs, D. E., Hieftje, G. M., Mark, H. L., and Hirschfeld, T. B.,** *Anal. Chem.*, 57, 2299, 1985.
38. **Honigs, D. E., Freelin, J. M., Hieftje, G. M., and Hirschfeld, T. B.,** *Appl. Spectrosc.*, 37, 491, 1983.
39. **McClure, W. F., Hamid, A., Giesbrecht, F. G., and Weeks, W. W.,** *Appl. Spectrosc.*, 38, 322, 1984.
40. **Honigs, D. E., Hieftje, G. M., and Hirschfeld, T.,** *Appl. Spectrosc.*, 38, 844, 1984.
41. **Mark, H. and Workman, J.,** *Anal. Chem.*, 58, 1454, 1986.
42. **Isaksson, T. and Naes, T.,** *Appl. Spectrosc.*, 42, 1273, 1988.
43. **Osborne, B. G.,** *Spectroscopy*, 4(4), 48, 1989.
44. **Geladi, P. and Kowalski, B. R.,** *Anal. Chim. Acta*, 185, 1, 1986.
45. **Lorber, A., Wangen, L. E., and Kowalski, B. R.,** *J. Chemometr.*, 1, 19, 1987.