Volume Fraction vs. Weight Fraction

Howard Mark*
Mark Electronics

CONTENTS
20.1 Historical Background ................................................................. 463
20.2 Problem Areas ........................................................................ 464
20.3 Proposed Root Causes .............................................................. 465
20.4 Previous Work ......................................................................... 468
20.5 New Developments ................................................................ 470
20.6 Experimental ......................................................................... 470
20.7 Results .................................................................................. 471
20.8 Summary of CLS Results .......................................................... 474
20.9 Principal Component Analysis of the Spectral Data—Using Reference Values ...... 475
20.10 Effects of Density on the Experimental Design ....................... 479
20.11 Summary ............................................................................. 481
20.12 Summary of CLS Results ........................................................ 482
20.13 Summary of PCA Results ........................................................ 482
20.14 Conclusions ........................................................................ 483
References .................................................................................. 483

20.1 HISTORICAL BACKGROUND

Ever since Karl Norris assembled the three key components (use of the NIR spectral region, diffuse reflection measurements of powdered samples, and use of what we now call multiple linear regression—MLR, the prototype algorithm for chemometrics) that ultimately turned into modern NIRS analysis, NIRS has existed as a self-contained, albeit internally consistent, analytical method. There was little or no connection to other existing chemical or spectroscopic science. Since Karl was affiliated with the US Department of Agriculture (USDA), his concerns were with the measurement of samples of interest to the department: largely foodstuffs and nonfood agricultural products such as cotton and tobacco. These types of samples differed greatly from samples of concern to academicians in some key respects: they were largely nonhomogeneous powdered solids as opposed to the clear liquids and gases that the theory of classical spectroscopy was based upon (e.g., Beer’s law, etc.).

It was generally not possible to separate the components of a sample (a kernel of wheat or a soybean, for example) so that the spectra of the “pure” components could be measured. It was at least as difficult to separate the components of something like a wheat kernel without changing them, as it would be to try to synthesize a wheat kernel from a collection of proteins, starches, vitamins, minerals, oils, and all the micro-ingredients that a natural wheat kernel contains. This was the task that Karl set out to address.

* hlmark@nearinfrared.com
I was working for Technicon Instruments Corp., one of the first three companies to commercialize the technology that Karl Norris had pioneered. When I joined the development group, the scientists already there were confused about how the technology was working. Compared with spectra in other spectral regions, the NIR spectra were anomalous. Baselines were nonexistent, absorbance peaks were at best broad and ill defined, standards were nonexistent, spectral interpretation was an unknown quantity, and what theory existed was known to be incomplete and likely somewhat incorrect. Since most of the samples of interest at that time were natural products (e.g., agricultural products such as wheat), spectra of pure constituents were unavailable; wheat “protein” could not be separated from a wheat kernel for example (at least, not and still remain recognizable as “wheat protein” in any way that could allow a meaningful measurement of a “wheat protein spectrum”). When we met scientists from the other NIR companies at conferences, or at the small “NIR meetings” that Karl would organize periodically, we found they shared our confusion and lack of understanding, as did Karl himself. Yet we all of us were well aware that from a practical, empirical, point of view, the technology “worked.” By the application of multivariate mathematics, it could be used to reliably analyze various products for their compositions.

Newcomers to the field, having had a course or two in classical spectroscopy, invariably had the same reaction: “I learned about spectroscopy in school. But here there’s no theory, no baseline, no reference beam, no standards, and no spectral peaks. Yet it works!! It must be magic!” All those missing pieces flew in the face of standard spectroscopic thought and practice. This distinguished and separated NIR analysis from the rest of science, making it appear to exist in a universe of its own. Yet, that “magic” made NIR an enormously successful commercial analytical method, being applied to many important industrial applications.

20.2 PROBLEM AREAS

The condition of existing in a universe of its own has persisted to this day. Very little progress (if any) has been made in connecting NIRS to the rest of the scientific universe. Despite its commercial success, many mysteries that were known back then still exist and still have no cogent (proven) explanations:

1. The apparent need for, and use of, more variables in calibrations than can reasonably be justified.
2. Algebra dictates that no more wavelengths or factors should be needed than variables present.
3. Difficulty in reproducing calibrations for the same constituents in the same type of samples.
4. Inability to reproduce wavelength selections (MLR models) in successive attempts.
5. Difficulty and/or inability in relating wavelengths chosen (for MLR) or prominent wavelength bands (for PCR/PLS) to spectral features.
6. Unexpected and/or unexplained “outliers.”
7. The standard error of calibration (SEC) and standard error of prediction (SEP) should drop precipitously when the number of wavelengths/factors equals the number of variations in the samples.
8. Spectroscopic measurements should be accurate over the entire range of concentrations, not only for dilute solutions or a small range of values.
9. Calibrations should be extrapolatable, with a calculated reduced accuracy at the extremes.
10. Calibration transfer should be as easily and readily performed as comparing two mid-IR spectra.

In practice, all the above difficulties were observed, but none of the proposed or expected explanations were. And, quite frankly, nobody seemed to care. Through the magic of Chemometrics,
“IT WORKED!” so why question success? And yet those unexplained “mysteries” were signs of trouble looming. Many potential root causes of the problems have been proposed.

### 20.3 PROPOSED ROOT CAUSES

1. Sample effects
   a. Inhomogeneity
   b. Polymorphism
   c. Interactions
   d. Optical scatter
2. Reference lab error
3. Instrument noise and/or variability
4. Instrumental nonlinearities:
   i. Stray light
   ii. Detector saturation
      a. Electronic nonlinearities
5. Calibration issues
   i. “Wrong” transforms of spectral data
   ii. “Wrong” calibration algorithm
   iii. Incorrect calibration parameters (number of wavelengths/factors)
6. Unknown relation between spectrum and composition (data transforms)

*Ad hoc* empirical solutions abound, often following this template:

1. One or more of the suspected above-listed root causes is targeted.
2. A (spectral) data transform is proposed to address the suspected root cause.
3. A calibration transfer experiment is designed to verify the correctness of the solution.
4. The experiment is conducted, and the data analyzed. “Outliers” are deleted.
5. The SEP of the transferred calibration is compared to the SEC of the calibration data.
6. Success is declared!

Over the years, many scientists have tried to address one or more of the suspected problems. Except for these studies, the attack usually consisted of hypothesizing a cause of the problem and taking an empirical approach by applying an *ad hoc* solution, usually as part of the routine chemometric algorithm being used for the application. Surprisingly, this approach often “worked,” insofar as the scientist was able to obtain satisfactory calibration and prediction performance for the analyte of interest in his samples and extrapolate that finding to the conclusion that his “fix” was actually correcting the situation. In contrast, even when that was the case, this approach usually gave little insight into the underlying (root) causes of the difficulties encountered. Even worse, the results obtained could rarely be generalized or extended to other applications of NIR or for other types of analysis. Every new application required a new research project, with the only common procedure being the application of chemometrics to the analytical situation.

Despite the practical usefulness of these *ad hoc* analytical methods, these exercises can hardly be called progress in the development of the technology. A given solution to an immediate analytical situation is often not applicable to another similar type of chemical analysis, even if the analyte and the sample types are the same, nor can they be applied to different types of analysis. No fundamental knowledge was developed through this work. The separation of an isolated method or calibration model from the rest of the scientific universe also isolates the method from further fundamental knowledge, so that there are no clues left as to which direction to make further explorations.
Often, the analytical “recipe” does not shed light on the fundamental science or processes underlying the particular analytical method used, nor does the analytical method used lend any insight into other potential applications of the analytical method, neither is the actual analytical method that is developed usable by other scientists. Indeed, not having explanations at that time for the effects seen, it made NIR analysis appear to be divorced from the mainstream of spectroscopy and existing in a universe of its own, separated from the universe of the rest of science.

The general NIR community, which at that time was quite small, shared the confusion over the behavior of NIR data. Attempts to explain this anomalous behavior went on for many years. Several of the common explanations that were proposed are listed above. Some of those common explanations have subcategories. For example, the use of the “wrong” calibration algorithm has, along with it, the use of an incorrect number of factors (for the principal components regression—PCR or partial least squares—PLS algorithm) or the wrong number of wavelengths (for the MLR algorithm). The use of MLR also carries with it the possibility of using the “wrong” wavelengths (even if the number of wavelengths is correct). Note: the equivalent is also true, for example, for PCR or PLS. That possibility usually goes unrecognized and unnoted, since most of the software packages available do not allow the user to select arbitrary subsets of the factors computed during the execution of the PCR or PLS algorithm.

For example, a user who decides to create a model that uses four principal components (PCs) is generally not free to select which four components to use, with all the common software packages for spectroscopic analysis, only the first four principal components may be used. This limitation is both a curse and a blessing. The curse is that the limitation prevents users from selecting just the minimum number of factors that are needed to optimally analyze the samples of interest. Instead, if a factor, say, the sixth, is important, the user cannot select four factors that include the sixth but are forced to include extra, unnecessary, factors such as the third and fifth, perhaps, that contribute to overfitting. The blessing is that a majority of users are already confused by needing to use complicated algorithms they are not comfortable dealing with and are happy and even relieved to not have another whole set of choices to make.

Matters remained thus for many years. Many successful applications of NIR and chemometrics were developed and published in the open literature, yet the “mysteries” remained. No general physical principles for determining a method to develop and optimize a calibration model, which could be applied by anyone, anywhere, were forthcoming. Only the general empirical methods seemed to “work” to develop a good calibration model.

Around 2010, a serendipitous discovery changed this picture [1]. It was found that the units in which the reference values for concentrations of an analyte are expressed actually matter. Use of the “wrong” units to express the analyte leads to large errors due to nonlinearity. Potential nonlinearity had previously been investigated [2], but the tools available at the time were inadequate, nor was attention directed to the right places. Nonlinearity can be covered up by the chemometrics, but it doesn’t go away. It lays hidden until something changes (e.g., a different set of samples) and then pops up, often in the form of one or more “outliers”. All the previous works for solving the problems of NIR analysis were directed toward investigations of problems with the spectral readings. Nobody had looked at the reference laboratory values except to disparage “bad reference laboratory results.” Nobody had concerned themselves with the question of whether they were measuring the right quantity in the first place.

The fundamental problem is illustrated in Figure 20.1, which shows the relationship between weight fraction and volume fractions of two different mixtures. Figure 20.1a shows the volume fractions of each component of a binary mixture, acetone and dichloropropane, as the weight fraction of the dichloropropane varies from zero to one (0%–100%). Even in the absence of chemical interactions between the two mixture ingredients, we can see that there is nonlinearity in the relationship between the two quantities: weight fraction and volume fraction. This nonlinearity is caused by the difference in density of the two mixture ingredients, causing the density of the mixture to vary as more of either ingredient is added.
Another way to approach the problem is to note that, since acetone is less dense than dichloropropane, adding a given weight of acetone corresponds to a greater volume of acetone and therefore, a greater total volume, than adding the same weight of dichloropropane would cause.

For a two-component mixture, Figure 20.1a clearly demonstrates the problem. For mixtures with more ingredients, the situation is even more complicated. The complication is due to the fact that, plotting the weight percent vs. volume percent for a mixture containing more than two ingredients, the details of the nature of the curvature seen depends on the densities of all the ingredients of the mixture. Figure 20.1b shows the relationship between the weight fraction and the volume fraction of dichloropropane in a ternary mixture as the relative amounts of the other two ingredients (acetone and dichloromethane) vary.

As the mixture changes, so does the curvature of the lines. In this case, a single line in Figure 20.1a becomes a family of lines, as seen in Figure 20.1b. Another property of mixtures with more than two ingredients is that there is no longer a one-to-one relationship between the weight fraction and the volume fraction; the volume fraction corresponding to a given weight fraction depends on the amounts of the other ingredients in the mixture and vice versa, as can also be seen in Figure 20.1b.
Note that the relationships in Figure 20.1 have nothing to do with spectroscopy, it is purely a matter of the physical chemistry of mixtures [1]. What DOES have to do with the spectroscopy is a fact that physicists have known about at least since James Clerk Maxwell formulated his eponymous equations: matter absorbs light in proportion to the volume of the absorbing material. Physicists knew this, but the knowledge seems not to have made it into the chemical and spectroscopic arenas. To be sure, sporadic papers do appear in the literature where the authors specify measuring concentrations of the analyte in their products as volume fraction, but these are rare exceptions. The vast majority of publications dealing with quantitative analysis specify weight fractions (or percentages) as the units of concentration. This is due to their historical backgrounds: the development of analytical chemistry has led chemists (except in medical applications, where the weight of analyte per deciliter is common) to consider weight fractions the “standard” way to perform measurements.

In the past, when the standard of measurement was based on gravimetric and volumetric wet chemistry, this was eminently satisfactory. To specify steel, for example, by the weight percent of chromium (Cr), Tungsten (W), and other metals it contains was a satisfactory way to set specifications for the steel. Then similar volumetric and/or gravimetric methods were used to determine whether a steel sample at hand met those specifications. Even then, when optical spectroscopy was applied to the measurement, spectroscopists knew that UV measurements could only be applied to samples with sufficiently low concentrations of the analyte or the instrument response would become nonlinear with respect to the concentration. Modern multivariate mathematical methods of computation were unknown then, but a univariate method could still be used, even in the face of the nonlinearity. All that was needed was to plot the response of a set of standard samples of known concentration. The response of the ultraviolet (UV) measurement could be compared to that graph to find the analyte concentration, whether it was linearly related to the UV response or not.

Similar procedures were used for mid-range infrared (MIR) analysis when they were applied to measuring organic materials. In both cases, it was important that the sample preparation be identical when measuring the unknown samples as when measuring the samples used to create the calibration curve.

Those methods were not applied to NIR analysis. As described above, “standard” samples could not be created, neither could special “calibration” samples be prepared for a calibration curve, nor could spectra be reduced to a single wavelength for measurement. Furthermore, the best efforts of the physics community have failed, for over 100 years, to describe a theory for computing the reflectance of an inhomogeneous solid sample from other properties of the sample.

In applying the multivariate mathematics to NIR spectra, the linearity of the responses of the interferences as well as of the analyte was simply assumed. No thought was given to the relationships between the spectroscopic readings (which can only see one thing) and the various ways that “concentration” can be specified. Hence the calibration algorithms have to work very hard to fit the nonlinear data into the structure of what it believes is a linear relationship [3].

20.4 PREVIOUS WORK

The work described in the aforementioned paper [1] and, in considerably more detail, in a series of columns in the magazine spectroscopy [4–17] explain how an experiment intended for a completely different purpose shed light on the questions raised above. For the sake of completeness, we summarize that work here.

The intended purpose of the experiment was to describe and explain the operation of the classical least squares (CLS) algorithm. Toward this end, the experiment was set up to eliminate or minimize all the ROOT CAUSES described above. The sample effects (inhomogeneity, polymorphism, interactions, and optical scatter) were addressed by using completely miscible mixtures of clear organic liquids with minimum polarity as the samples.

Reference laboratory errors were reduced (nearly eliminated) by weighing the liquids; weighing is one of the most precise and accurate measurement methods available, with precisions of as much
as $10^{-6}$ potentially available. No separate actual laboratory analyses were performed. Due care was exercised to prevent changes in the samples due to evaporation or other extraneous effects. Thus, we expect any errors from this source to be eliminated.

Instrumental effects were minimized by routinely checking the instrument characteristics and operating in a regime where the detector signal was neither too high (so as to saturate the detector) nor too low (where stray light or other extraneous effects might cause problems).

Calibration issues (incorrect transforms or number of PCR/PLS factors) were eliminated by using the CLS algorithm without any data transformations. CLS does not utilize any computed factors for performing the calibration; only the spectra of the pure materials comprising the samples are needed. Adherence to Beer’s law was verified by ascertaining that mixture spectra reconstructed from the CLS coefficients and pure material spectra agreed with the actual spectra of the corresponding mixtures.

A set of 15 mixtures of toluene, $n$-heptane, and dichloromethane, specified by a statistical mixture design, was made using the design specifications to determine the weights of each ingredient to be added to each mixture. Their transmission spectra were measured. No transformations were applied to the data. No samples stood out as potential outliers and none were eliminated.

The results were surprising and disappointing. The computed concentrations of the ingredients in most of the mixtures showed large discrepancies (as much as 10%–15%, in some cases) from the specified values of the carefully made-up samples. A review of the experimental work revealed no blunders or other experimental discrepancies.

Since the reconstruction of the mixture spectra demonstrated that Beer’s Law was operating as theory indicated, the only possible cause of the discrepancies was the use of an incorrect value for the actual mixture compositions. The care used in executing the experiment eliminated the possibility that the samples were made up incorrectly. The only other possibility that remained was that the optical results were being compared to the wrong property of the samples. As stated in that paper [1], “We expect the spectral values to correspond to some physical characteristic but there is no reason to expect that characteristic to be the weight of a sample component.” Weight, after all, depends on the nuclei of the atoms, while the spectroscopic interactions depend on the behavior of the electrons in the atomic and molecular orbitals. It remained, then, to determine what characteristic(s) of the mixtures would provide values that agreed with the spectroscopically measured values.

Several possibilities were proposed and tested. The results from one particular sample with nominal weight percent of the various ingredients are shown in Table 20.1, showing the results from five different ways of expressing the concentrations of the various mixture ingredients.

Interestingly, the value for toluene comes pretty close to the same value as the spectroscopic measurement no matter how the “concentration” is expressed. Only the volume percent gives a value for all three ingredients that is close to the corresponding spectrally measured value.

<p>| TABLE 20.1 |
|-----------------|-----------------|-----------------|
| <strong>Results from a Sample with Nominal Composition: Toluene = 25%, Dichloromethane = 25%, n-Heptane = 50%. All Values are wt %</strong> |</p>
<table>
<thead>
<tr>
<th>Spectral values</th>
<th>Toluene</th>
<th>Dichloromethane</th>
<th>n-Heptane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight percent</td>
<td>23.01</td>
<td>17.48</td>
<td>56.92</td>
</tr>
<tr>
<td>Mole percent</td>
<td>23.86</td>
<td>26.45</td>
<td>49.68</td>
</tr>
<tr>
<td>Mole % of hydrogen</td>
<td>24.29</td>
<td>29.21</td>
<td>46.50</td>
</tr>
<tr>
<td>Mole % H corrected&lt;sup&gt;a&lt;/sup&gt;</td>
<td>19.5</td>
<td>5.86</td>
<td>74.64</td>
</tr>
<tr>
<td>Volume percent</td>
<td>22.93</td>
<td>16.50</td>
<td>60.56</td>
</tr>
</tbody>
</table>

<sup>a</sup> Corrected for the density of the ingredient.
Table 20.1 presents the comparative information for one sample from the set of 15 used. In the original publication [1], the results from all 15 samples were examined and presented. The full set of results makes a compelling case that, just as with this single sample, volume percent is the only measurement unit that gives quantitative results for the concentrations of the various ingredients that consistently agree with the spectroscopic values.

What could not be done in the course of that work was to actually apply any of the modern calibration algorithms to the data used for that study, because the number of samples ($n = 15$) in the set available was simply too small.

20.5 NEW DEVELOPMENTS

Perusing the literature, a pair of papers about using spectroscopic analysis to examine clear liquid samples by Willem Windig attracted attention [18,19]. While Willem’s interest in the dataset was for testing the Simplisma algorithm, the data itself appeared to have properties that would make it useful for the purposes of further examining the effects of different units on calibration, i.e., enabling tests of conventional calibration algorithms along with CLS analysis similar to the previous experiment [20].

20.6 EXPERIMENTAL

Details of the sample preparation and measurement methods are described in [18]. With the assistance of Tormod Naes, Prof. Windig was contacted, and he graciously agreed to supply the data he used for the Simplisma tests described above. Table 20.2 describes the properties of the sample set.

Spectra were supplied as absorbance values from 1100 to 2500 nm. The data was of very high quality, so no samples were deleted for any reason. Also, no data transformations were applied to the spectra.

The experimental design had an interesting characteristic. With five different ingredients making up the mixtures, there are ten possible ways they can be plotted pair-wise, as shown in Figure 20.2.

With the experimental design used, all ten pairs of ingredients produce essentially the same plot except for minor variations in the dispersion of the points around each concentration level. I have been unable to find a description of the general type of statistical experimental design that was used to create this pattern. (Prof. Windig, himself, had himself lost track of the source of the design.)

CLS calculations and plots were performed using custom software written in MATLAB (Mathworks, Natick, MA), in order to verify that the data followed Beer’s Law and that results corresponding to those in the previous experiment [20] were achievable. CLS calculations were compared with the weight fraction values and with the (computed) volume fraction values as reference values; these results were plotted.

---

**TABLE 20.2**

**Characteristics of the Dataset Described in [18,19] and Used for this Work**

The sample set contained 70 mixtures, each containing five ingredients:

- Acetone
- 2-Butanol
- 1,2-Dichloropropane
- Methylene dichloride
- Methanol

Concentrations were defined gravimetrically as weight fractions, and a statistical design was used to specify the concentration of each ingredient in each of the 70 mixtures. The concentration of each ingredient in each sample was set to one of the following concentrations: 10, 22.5, 35, 47.5, and 60 wt %.

The spectrum of each mixture was measured twice.

Key characteristics of the mixture design are described in Figure 20.2.
Volume Fraction vs. Weight Fraction

PCA and PCR calculations and plots were performed using The Unscrambler (CAMO, Oslo, Norway) X 10.1. While as we will see, six factors were always enough to describe the mixture spectra, the PCA and PCR programs were allowed to compute up to as many as ten factors, to guard against potentially missing any interesting phenomena that might occur at higher numbers of factors.

20.7 RESULTS

CLS: spectral analysis without reference laboratory data.

As described in the Introduction, our first goal was to verify that these spectra perform the same way that our own previous work [20] on a smaller sample set indicated. To this end, our initial investigations were to apply CLS analysis to the data to verify that they behave the same way as the previous, simpler data set behaved, before applying the more sophisticated calibration algorithms. This provides reassurance that the properties and behavior of the two data sets are similar, and that the conclusions based on the results of the first data set were general and not idiosyncratic to that data set, and therefore could be expected to apply to other data sets. The volume fractions of the five chemical components of the mixture were calculated using the known weights of the ingredients added to each mixture and the densities of the pure materials shown in Table 20.3 [21].

The CLS results were evaluated in two ways: First was the ability to reproduce the spectra of several of the mixtures in the dataset from their pure-component spectra and the CLS coefficients computed for that mixture, compared to the actual spectrum of the mixture. This is also a test of whether Beer’s Law holds for these mixtures. Since the experimental design did not include the pure component spectra (although those were measured separately, and fortunately so, since they were needed for use with the CLS algorithm. The samples used as surrogates of the pure materials for this comparison were the ones containing each of the components at its maximum value (i.e., 60%) in the set of spectra, as surrogates for the pure component spectra.

These comparisons are shown in Figure 20.3, and we can see excellent agreement between the actual spectra of the various mixtures and the spectra computed from the pure components, which

![FIGURE 20.2 Overall characteristics of the orthogonal, balanced, symmetric statistical experimental mixture design used to specify the samples. The five different ingredients allow for 10 pairwise comparisons; each of the 10 possibilities provides a plot that looks like the one below.](image-url)
TABLE 20.3
Densities of the Components Used for the Five-Component Mixtures

<table>
<thead>
<tr>
<th>Component</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetone</td>
<td>0.792</td>
</tr>
<tr>
<td>2-Butanol</td>
<td>0.808</td>
</tr>
<tr>
<td>1,2-Dichloropropane</td>
<td>1.1593</td>
</tr>
<tr>
<td>Methanol</td>
<td>0.7928</td>
</tr>
<tr>
<td>Methylene dichloride</td>
<td>1.336</td>
</tr>
</tbody>
</table>

FIGURE 20.3  Comparison between actual spectra of mixtures (blue) and the corresponding spectra reconstructed from the pure components (green). The mixtures shown are the ones that have the maximum concentration (60 wt %) of the specified ingredient, with all other ingredients being at their minimum values (10 wt %).
verifies the applicability of Beer’s law to the mixtures, and also the lack of any more than minor interactions between the any of the mixture components [22].

The second way of evaluating the results was to see whether they conform to the theory previously developed to explain the relationships between the spectroscopic analysis, as reconstructed from the CLS algorithm, and the different units (weight fraction and volume fraction) used to determine the compositions of the mixtures, as developed during the study of the smaller experiment. These comparisons are shown in Figure 20.4.

The left-hand column of plots in Figure 20.4 shows that the data points in each cluster, representing those samples with the same nominal weight fraction specified by the experimental design, lie on a vertical or near-vertical, line. However, this characteristic creates large differences between the nominal reference laboratory value and the spectrally determined value. When the reference values are converted to volume fractions, as in the right-hand column of plots in Figure 20.4, the clusters of data points no longer lie on a vertical line. The predictability of the volume fractions used as surrogates for the concentration causes individual clusters to spread out parallel to the horizontal axis as well, in a manner that makes them agree with the spectroscopic values and conform to a hypothetical line describing the overall behavior of the data, albeit some with increased random scatter. This behavior, which all five components of the mixtures exhibit, agrees with the corresponding results found in the previous experiment that used mixtures with only three components [20].

FIGURE 20.4 Plots of the spectroscopic measure of concentration from the CLS algorithm versus the weight fraction or volume fraction representing the known concentrations of the components of the mixtures.

(Continued)
20.8 SUMMARY OF CLS RESULTS

Preparatory to applying other calibration algorithms to the data, the CLS algorithm demonstrated that each of the five components of the mixtures behaved similarly to the way the data in the previous experiment performed. The ability to reproduce the spectra of mixtures from the spectra of the mixture components and the improved agreement between spectrally calculated concentration values and the volume fractions of the mixture components compared to the weight fractions constitute confirmation of the two major characteristics for the use of the CLS algorithm, which were

FIGURE 20.4 (CONTINUED) Plots of the spectroscopic measure of concentration from the CLS algorithm versus the weight fraction or volume fraction representing the known concentrations of the components of the mixtures.
previously found. These findings provide assurance that the theoretical underpinnings explaining the behavior of the spectral data are correct and apply to both data sets.

20.9 PRINCIPAL COMPONENT ANALYSIS OF THE SPECTRAL DATA—USING REFERENCE VALUES

The mixtures each contain five chemical components, which therefore bring with them five degrees of freedom. In a mixture, however, the concentration of each component is dependent on the other four since the sum of all five must equal unity (or 100%); this relationship reduces the total number of degrees of freedom in the data set to four. In theory, therefore, four principal components (PCs) should suffice to accommodate the variance introduced by the several chemical components of the mixtures.

Figure 20.5 shows the residual variance of the spectra as a function of the number of PCs computed and applied to the data. Contrary to what theory tells us, Figure 20.5 indicates that not all the variance of the spectral data is accounted for until seven principal components have been calculated and applied to the spectral data.

Figure 20.6 shows a set of plots, one for each of the chemical components of the mixtures, which shows the amount of residual variance from fitting the principal components of the spectra to the compositions of the mixtures for each ingredient. These graphs depend slightly on the reference values of the mixtures, but plots made using either weight fractions and volume fractions for the reference values are almost indistinguishable to the eye. Several points are notable:

For all ingredients, the total variance (e.g., variance for zero PCs applied to the data) is greater when concentration is expressed as volume fractions than for weight fractions. All ingredients enter the experimental design with the same set of values, so we expect that the behavior of one ingredient should reflect the behavior of all. We note, however, that three ingredients have densities less than unity, while only two have densities greater than unity. Since conversion from weight basis to volume basis is nonlinear, this implies that over the entire set of samples, there is a larger total numeric value for volumes than there were for weight and the variance scales correspondingly.

FIGURE 20.5 PCA residual variance of the calibration spectra as a function of the number of principal components used.
FIGURE 20.6  PCA residual variances of ingredient concentrations as a function of the number of principal components used. Each part of this Figure 20.6 shows the residual variance plot when the ingredient is expressed as weight fraction (Column 1) and volume fraction (Column 2). The several mixture ingredients are plotted in order of increasing density of the ingredient.

(Continued)
Volume Fraction vs. Weight Fraction

Since variances are not scaled by the computations, this trivial numerical relationship can be the cause of the larger total variance. The computation of principal components scales the results to the total variance of the corresponding dataset; this would then scale all the results linearly and make the plots appear to be essentially the same although they are in fact (inversely) density dependent. We will examine this behavior further, later in this paper.

Second, for each ingredient, the graphs of the behavior of the principal components to explain the ingredient data per Figure 20.6 appear to the eye to be the same whether the ingredient concentration is expressed as weight fraction or volume fraction. This is not very surprising, since the principal component loadings are calculated entirely from the spectral data and therefore are themselves the same regardless of the chemical component that is being fitted.

Third, several of the graphs (notably the ones for acetone, 2-butanol, dichloropropane, and, arguably, methanol) contain sections (between points representing successive principal components used in the calibration) where the graph is horizontal or nearly horizontal. This indicates that for the two points marking the horizontal sections, the second point of each pair, corresponding to the larger number of principal components, is not removing any variance due to that analyte, from the fit of that loading to the spectral data.

**FIGURE 20.6 (CONTINUED)** PCA residual variances of ingredient concentrations as a function of the number of principal components used. Each part of this Figure 20.6 shows the residual variance plot when the ingredient is expressed as weight fraction (Column 1) and volume fraction (Column 2). The several mixture ingredients are plotted in order of increasing density of the ingredient.
Finally, unexpected and highly interesting results were noted when the graphs in Figure 20.6 were compared between the different ingredients. Here we point out that the graphs for the various mixture ingredients in Figure 20.6 are plotted in order of increasing density of the ingredient. The following characteristics of the plots were noted: the least dense ingredients (acetone, methanol, methylene chloride, and 2-butanol) had their error variance reduced to zero with four or fewer principal components, the remaining one (dichloropropane) required more than four principal components. Furthermore, the mixture ingredients with the lowest densities had a tendency to utilize the fewest and lowest numbered principal components. Again, this appears to fly in the face of the theoretical prediction that no more than four PCs should be needed to fit the spectrum for any ingredient.

Another relationship between the PCs was then noted: those PCs where the plot contained horizontal sections are the ones that required excess (i.e., >4) PCs. This tendency is especially pronounced in the case of dichloropropane, where the horizontal section is the one between the first and second PCs. This indicates that the second PC is not contributing to the model for dichloropropane, and we see that for this ingredient, only PCs one through five are reducing the error variance. In other words, only four PCs are needed to explain the variance due to the dichloropropane, but those are not the FIRST four PCs.

The other ingredients behave similarly. Acetone and butanol also include PCs that are not reducing the variance of the respective ingredients, and thus, a calibration for those ingredients would be as good without including the corresponding PC (actually better, arguably, since such a calibration would avoid the loss of degrees of freedom associated with including the unnecessary PC).

In the light of the plots of methanol and acetone, where acetone uses PCs 3 and 4 while methanol uses PCs 1 and 2, the statement above: “the mixture ingredients with the lowest densities had a tendency to utilize the fewest and lowest numbers of Principal Components” seems incongruous. We should expect methanol to be plotted before acetone, in that sequence. We note another property of these data: methanol and acetone have almost identical densities. Then, we should expect them to have the same eigenvalues in the principal component decomposition. This property would cause these two ingredients to be degenerate in the experimental design used.

A known property of degenerate eigenvalues (a property used in, e.g., quantum mechanics) is that indistinguishable degenerate states mix to form distinguishable states. Thus, acetone and methanol, being the two materials that have the lowest, and almost identical, densities in the set of ingredients, are both degenerate with respect to each other and contribute the largest amount of variance to the total; they therefore have virtually all their variance accounted for by the first four principal components, and each of these ingredients requires two PCs to account for their variance. We can see this in the plots.

Degenerate eigenvectors are also orthogonal; of necessity, they account for different portions of the variance, albeit the same amount. Recall that two key properties of principal components are that they are all orthogonal and that the first one accounts for the source of maximum variance in the data. Thus, the analysis uses the first two PCs to explain all the methanol variance while acetone, being orthogonal, includes none of the variance of the first two PCs and is completely explained by the subsequent two PCs. All this can be seen in Figure 20.6.

Note a careful distinction here: while all the variance for acetone and methanol are explained by those four PCs, that does not mean that those PCs are in any way “used up” in the sense of not including any more variance that could be available to help explain the effects of other ingredients in the mixture, although not each other. Indeed, each of the other three ingredients have some of their variance accounted for by one or more of the first four PCs.

Another noteworthy point is that dichloropropane, which has the second highest density in the set of ingredients, has two sections of the residual variance plot that show no decrease in variance. This is also the ingredient that most clearly shows a requirement for more than four PCs to fully accommodate its contribution to the total variance of the data. Yet, if we ignore the two sections of
the plot that show no decrease from the inclusion of the corresponding PCs into the model, we find that the dichloropropane also requires four PCs to accommodate its variance although those four are also not the first four that are computed from this data.

This explains two apparently contradictory observations: first, that mathematical theory tells us that no component should need no more than four factors (PCs) to create a calibration for any ingredient yet we find that we need seven for this set of data. The reason is that, as careful study of Figure 20.6 reveals, while no individual ingredient needs more than four PCs, those are not necessarily the first four, or even the same four, for all ingredients.

Not all the PCs contribute to the analysis of any particular ingredient. Horizontal or nearly horizontal sections of the residual variance plots of Figure 20.6 indicate that the higher-numbered PC from that section does not contribute to the analysis of the corresponding ingredient. Thus the first and second PCs do not contribute to analysis of acetone, the third PC to the analysis of methanol, the second and fifth PC to the analysis of dichloropropane, or the fourth PC to the analysis of methylene chloride. Because of that, sometimes it requires more than the first four PCs to be able to analyze all the ingredients in a given mixture, as it is here most clearly seen in the case of dichloropropane, which requires a total of five PCs in order to include the four actually needed to perform the analysis for that ingredient. Thus, an ingredient requiring N PCs for its calibration may still require more than N PCs to be computed, since some of those PCs may not be among the first “N” PCs.

20.10 EFFECTS OF DENSITY ON THE EXPERIMENTAL DESIGN

Figure 20.2 showed the diagrammatic experimental design that specified the weights, and therefore the weight fractions, of each of the ingredients. We noted above that the plot of the experimental design was essentially identical for any pair of ingredients that were plotted, regardless of the amount of the other materials in the mixture, and therefore all the data points for a given concentration pair all overlapped. This is not the case when the actual spectroscopically active units (volume fractions) are used to specify the design.

Figure 20.7a illustrates how the design is affected when volume fractions are plotted rather than weight fractions. In Figure 20.7a, volume fractions of butanol and methylene chloride are plotted against each other. It is clear that the volume fractions for mixtures with the same amount (by weight) are no longer the same, despite the same weight of each ingredient being plotted, because the volume fractions change depending on the other ingredients in the mixture. The groups of data points with the same amounts of butanol and methylene chloride have the same relationship to each other as they did when the weight fractions were being plotted, but within each group of data points, dispersion of the data has been introduced. The plot of each of the 10 combinations of ingredients has these same properties although the different common ingredient changes the variances of the data points in the plot. This is an important point.

An even more important effect is that the plots are no longer identical for the different ingredient pairs, as they were in Figure 20.2. While all the plots of the different pairs of ingredients are spread out as shown in Figure 20.7a, they are spread out differently and by different amounts due to the different variances induced in the data. This is another illustration of the same property presented earlier in the discussion of Figure 20.6, that less dense ingredients occupy larger volumes and therefore bring larger amounts of variance into the dataset; this is demonstrated in Figure 20.7b, where butanol volume fractions are plotted against volume fractions of the other four materials. Since the plot of butanol versus each of the other four ingredients would look very similar, it would be very difficult to distinguish them for comparison purposes. In order to be able to compare the different ingredient pairs, therefore, different colors are used to distinguish the four other ingredients from each other. We can now easily see that the plots from the different ingredients are of different sizes due to the differences in the variances of the different ingredients.

We also see now that the data points cover most of the area of the plot. More subtly, but more importantly, is that the extreme Y point for each of the other ingredients is at a different value with
the intermediate mixtures changing proportionately. This indicates that each ingredient introduces a different amount of variance into the design and, from the arguments above, we can now see that the low-density ingredients occupy more space on the plot (as they do in the samples). Since variance increases as the square of the amount of dispersion of the data, this effect magnifies the effect.
of the varying densities. While the weight variances of all the ingredients are the same, the different ingredients contribute a different amount of variance to the volume and to the spectroscopic response.

We also now see that, despite the use of a mixture statistical design that appears to be completely balanced for the five mixture ingredients, those ingredients with higher density add less variance to the dataset than ingredients with low density. Note that in Figure 20.7b, for the samples where the value of butanol = 0.1, the highest $Y$ values for acetone and methanol (the two ingredients with the lowest density—red and green in Figure 20.7b), the highest spectral ($Y$) reading is roughly 0.65. The highest $Y$ value for methylene chloride (the densest ingredient, black in Figure 20.7b) is roughly 0.5.

Thus, despite all good intentions, the design is NOT balanced among the five ingredients as intended. This effect is distorting the nature of the actual experimental design affecting the data being measured: what you see is NOT what you get when using weight fractions in the design for spectroscopic analysis. Compare Figure 20.2 with either part of Figure 20.7. When implementing the experimental design by weight, Figure 20.2 implies that all ingredients included with the same weight have the same influence on the results, but that is not the case when performing spectroscopic analysis. The actual design that is effectively controlling the experiment is the one presented in Figure 20.7. Interestingly, since the densities of the methanol and acetone are the same, their effects on the spectroscopic behavior are still subject to the degeneracy described above; note that in Figure 20.7b, the points representing acetone (green) and methanol (red) overlap each other. Also, if the design were plotted against an ingredient other than butanol as being common to the other four, then the entire design would appear to be different from Figure 20.7b.

We can extrapolate this finding to any experimental design used to specify design points for any experiment where spectroscopic readings are used for measurement. This discrepancy between different units to specify an experimental design needs to be taken into account when an experimental design is developed. The amount of the various ingredients in a mixture needs to be adjusted to ensure that specified volumes, rather than specified weights, of the different ingredients are used in creating the samples in order for the design to be properly balanced. Otherwise, the actual design implemented in the specification of the design points is not what it appears, or is intended, to be. Experimenters in the future, designing experiments in which spectroscopy plays a key role, need to take this effect of density of the ingredients into account in their experimental designs. One way this can be accomplished is by specifying the ingredient composition so as to use equal volumes of ingredients; the ingredients could even be measured and dispensed using classical volumetric equipment. Alternatively, if ingredients are specified by weight, this can be accomplished by specifying the ingredient weights that correspond to the needed volumes. Alternatively it can be accomplished post facto by using the various ingredient densities to correct the experiment to the actual ingredient volumes used.

### 20.11 SUMMARY

We summarize this as follows:

1. The experimental design is balanced (symmetric) in weight.
2. Densities differ, and therefore it is not balanced for volumes.
3. For the same weight, less dense liquids have greater volumes.
4. The design would therefore be distorted and differ for each ingredient pair.
5. Spectroscopy responds in proportion to volume and therefore is more sensitive to lower-density ingredients. This breaks the symmetry.
6. In the design, less dense ingredients have greater volume.
7. Acetone and methanol have the same density.
8. Acetone and methanol principal components are degenerate.
9. Less dense ingredients have greater optical variability (variance).
10. Principal components are sensitive to variance: first component represents the greatest variance and therefore the least dense ingredient.
11. Therefore, principal components are otherwise calculated in order of increasing density of the ingredient.

The various observations discussed above all lead to the following explanations: the theoretical mixture design was balanced, but the design and its properties were based on the use of the weight fractions to specify the various ingredients. Since the various materials have different densities, the specified design is NOT balanced with regard to the volume fractions of the various ingredients in the mixtures. Ingredients with the lowest densities (acetone and methanol) are present with higher volumes in the actual mixtures and therefore make a larger contribution to the total variance in the mixtures (and consequently the spectra) than the ingredients with higher densities. This can be deduced from the behavior of the principal components and their relations and can actually be seen in Figure 20.6.

This allows us to recognize that there is a direct connection between a physical property of the ingredients, in this case the density, and mathematical properties of the data set, in this case the order of calculation of the principal components. The PCs corresponding to the low-density ingredients are the ones that contribute the most to the variance to the data set and are consequently the ones computed first.

20.12 SUMMARY OF CLS RESULTS

Preparatory to applying other calibration algorithms to the data, the CLS algorithm demonstrated that each of the five components of the mixtures behaved similarly to the way the data in the previous experiment performed. The ability to reproduce the spectra of mixtures from the spectra of the mixture ingredients verifies the applicability of Beer’s law, and the improved agreement between spectrally calculated concentration values and the volume fractions of the mixture components compared to the weight fractions constitutes the two major characteristics of the use of the CLS algorithm that were previously found. These findings provide assurance that the theoretical underpinnings explaining the behavior of the spectral data are correct.

20.13 SUMMARY OF PCA RESULTS

The intended purpose of the experimental design used was to ensure balance between the five ingredients’ contribution to the variance of the data. This intent was subverted by the differences in the densities of the various components, the property affecting what the volume (the property the measurement is actually sensitive to [20]), which is not the same for the various ingredients in the mixture. Since the mixtures were made according to the weights of the various ingredients included, the corresponding volumes differed, and therefore the low-density ingredients affected the total variance more than the high-density ingredients did, effectively destroying the expected balance of the experiment. The residual variance plots of Figure 20.6 are in order of increasing ingredient density, so it becomes clear by scanning those plots that there is a distinct tendency for the PCs contributing to the spectrum of a given ingredient to move to the higher PCs as the density of the corresponding ingredient increases. This is in accord with theoretical expectations.

Not all the PCs contribute to the analysis of any particular ingredient. Horizontal, or nearly horizontal, sections of the residual variance plots of Figure 20.6 indicate that the higher-numbered PCs from that section do not contribute appreciably to the analysis of the corresponding ingredient. The first and second PCs do not contribute to the analysis of acetone, the third PC does not contribute to the analysis of methanol, nor do the second and fifth PCs to the analysis of dichloropropane, nor the fourth PC to the analysis of methylene chloride. Because of this effect and, despite the fact that
each ingredient in the mixture requires only four PCs, it sometimes requires more than the first four PCs to be able to analyze all the ingredients in a given mixture. In the case of dichloropropane, we require a total of five PCs in order to include the four active ones actually needed to perform the analysis for that ingredient.

This explains the common observation that more total PCs often need to be computed for a calibration than theory would seem to indicate because a calibration using N PCs does not always find them among the first N PCs that are computed. The excess PCs that are often computed, which are not required for individual analyses, may be profitably deleted from the analysis for the corresponding ingredient. The previous conclusion [20] about the nonlinear relationship is now additionally seen to be a consequence of the direct relationship between the physical property (density) and mathematical properties (order of calculating PCs) of the spectroscopic data measured on samples affected by that physical property, and thus the nonlinearity found previously [20]) is only part of the explanation for the need for “extra” factors.

20.14 CONCLUSIONS

The principal component analysis of the data conforms to the theory, albeit in an unexpected way. Each of the five ingredients does in fact have their variance accounted for by four principal components although not necessarily the first four. Since PCA is performed on the spectra only, this is true whether correct (volume fractions) units or incorrect units are used to express analyte concentrations.

Due to the inverse relationship between ingredient density and the corresponding volume for equal weights of an ingredient, with everything else being equal, principal components were calculated in order of decreasing density, for the contribution of an ingredient to the variance of the spectral data.

When a statistical experimental design is implemented for spectroscopic measurement, the design parameters need to take into account the differences between specifications of the design in different units. Otherwise, analysis of the data will produce incorrect (or at least, unexpected) results for ingredients that are expected to have similar effects (i.e., due to equal weights) but in fact are affected by variables that have different spectroscopic activity. Corrections to the implementation of the design should be made to ensure that the design specifications represent the volume fractions of corresponding ingredients, so that the statistical analysis to mean what it is expected to mean.

The sensitivity of the spectroscopic measurement to the volume fraction of the ingredients of a mixture distorts the nature of an experimental design-based mixture made according to weight. Ingredients that might be expected to have effects of similar magnitudes because equal weights were added to the mixture will actually have different magnitudes of their effects. This is above and beyond any effects that could be attributed to the actual spectral differences, a question we have not considered.

REFERENCES