33

MULTIVARIATE ANALYSIS FOR PHARMACEUTICAL DEVELOPMENT

Frederick H. Long
Spectroscopic Solutions, LLC, Randolph, NJ, USA

Multivariate analysis (MVA) is the statistical analysis of many variables at once. Many problems in the pharmaceutical industry are multivariate in nature. The importance of MVA has been recognized by the U.S. FDA in the recent guidance on process analytical technology [1]. MVA has been made much easier with the development of inexpensive, fast computers, and powerful analytical software. Chemometrics is the statistical analysis of chemical data, which is an important area of MVA. Spectral data from modern instruments is fundamentally multivariate in character. Typically pharmaceutical process monitoring requires more than one variable. Furthermore, the powerful statistical methods of chemometrics are essential for the analysis and application of spectral data including NIR and Raman. In this chapter, we will review the subject of chemometrics and MVA and its application in the pharmaceutical industry.

With spectral data, it is not uncommon to measure several thousand variables at one time. However, it is often hard to conceptualize so many variables; therefore, we will begin our discussion of MVA with a few simple examples that illustrate important statistical concepts which are essential in chemometrics. The first problem is a set of pharmaceutical quality data. Measurements of density and assay have been measured for 43 lots of material. The data is shown in Table 33.1. Inspection of the data reveals that the density values are near 1.0, while the assay values are closer to 100. A goal of the data analysis is to understand the variation within the data set. It will be advantageous to have the two variables in the data set with similar magnitudes; therefore, we will scale each of the two variables by its own standard derivation. The standard deviation, s, of a set of measurements ($x_1, \ldots, x_n$) is given by

$$\sigma = \left(\frac{\sum (x_i - \overline{x})^2}{n-1} \right)^{1/2} \tag{33.1}$$

where $\overline{x}$ is the average value of the n measurements. The denominator in equation 33.1 is $n-1$, because once the average is calculated, there are $n-1$ degrees of freedom. We note that the standard deviation has the same units as the variable of interest.

A plot of the scaled data is shown in Figure 33.1. The x-axis is the scaled density and the y-axis is the scaled assay values. Each point represents 1 of the 43 lots of material. From the plot in Figure 33.1, one data point is far away from all of the others. Statisticians call data points that do not belong to the data set outliers. Outliers are important to identify and remove from the analysis of the data set, because a single outlier can greatly influence the statistical analysis and obscure underlying trends in the data. We note that while outliers are often removed in a research and development environment during method development, great caution must be used in removing outliers during validation or use in actual production.

The scaled data are replotted in Figure 33.2, with the outlier point removed. The reader will also note that the origin of the graph has been moved to the center point of the data set. This operation is called *mean centering*, when the average of the overall data set is subtracted from the data. As mentioned earlier, in MVA we are concerned with

Chemical Engineering in the Pharmaceutical Industry: R&D to Manufacturing Edited by David J. am Ende

TABLE 33.1 Pharmaceutical Quality Data Example

Density (g/cm^3)	Assay (mg)
0.801	121.410
0.824	127.700
0.841	129.200
0.816	131.800
0.840	135.100
0.842	131.500
0.820	126.700
0.802	115.100
0.828	130.800
0.819	124.600
0.826	118.310
0.802	114.200
0.810	120.300
0.802	115.700
0.832	117.510
0.796	109.810
0.759	109.100
0.770	115.100
0.759	118.310
0.772	112.600
0.806	116.200
0.803	118.000
0.845	131.000
0.822	125.700
0.971	126.100
0.816	125.800
0.836	125.500
0.815	127.800
0.822	130.500
0.822	127.900
0.843	123.900
0.824	124.100
0.788	120.800
0.782	107.400
0.795	120.700
0.805	121.910
0.836	122.310
0.788	110.600
0.772	103.510
0.776	110.710
0.758	113.800

investigation of the variation within the data set. The average values of the data set are not of primary importance. Two arrows in the figure illustrate the two directions of variation within the data set. P1 is the largest direction of variation and P2 is the second direction of variation. It is important to note that P1 and P2 are perpendicular to each other. In MVA, P1 and P2 are the first and second principal components of the data set, respectively.

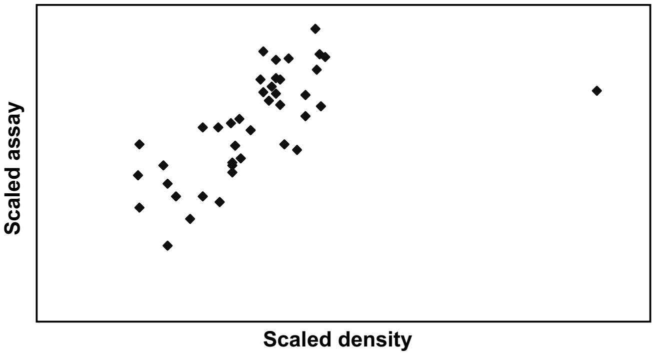

FIGURE 33.1 Scaled pharmaceutical quality data. Both the density and assay are scaled by the standard deviation of the data for each variable. Because the variables are scaled by the standard deviation, they are dimensionless.

For each one of the data points, the projection of the data point onto the P1 or P2 vector is called a score value. Plots of score values for different principal components, typically P1 versus P2 are called score plots. Score plots provide important information about how different samples are related to each other. Principal component plots, also called loading plots, provide information about how different variables are related to each other. Because we are working with scaled variables, the PCs and scores are dimensionless variables.

The mathematics of PCA can be clearly described using linear algebra [2]. An excellent discussion of linear algebra can be found in the references [3]. By convention, the data matrix, **X**, has p columns and n rows, and each column represents another variable and new rows for each observation or sample. The average data matrix, $\overline{\mathbf{X}}$, is the average of each individual column (i.e., variable) in the data set. Mean centering is written as

$$(\mathbf{X}-\overline{\mathbf{X}}) \qquad (33.2)$$

The covariance matrix is written as

$$\mathbf{C} = (\mathbf{X}-\overline{\mathbf{X}})^{\mathbf{T}}(\mathbf{X}-\overline{\mathbf{X}}) \qquad (33.3)$$

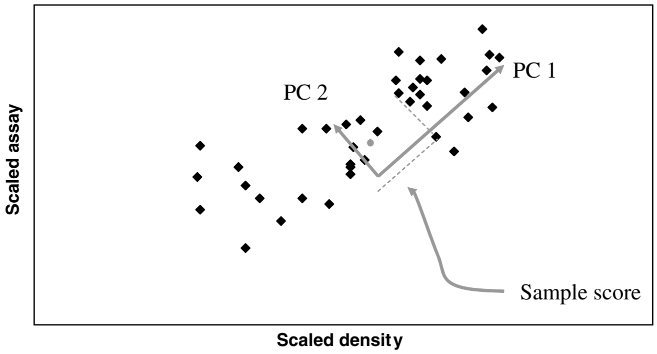

FIGURE 33.2 Scaled pharmaceutical quality data showing both the first and the second principal components for the data set. The first principal component is the direction of the maximum variation within the data set. The second principal component is perpendicular to the first PC. The scores for each sample point are given by the projection of the data point onto the principal component vector.

where an upper script **T** represents a matrix transpose. The covariance matrix is a square, symmetric, $p \times p$ matrix. The covariance matrix provides information about the relationship between different variables. For example, the *i,j* element of the covariance matrix quantifies the relative change between the *i,j* variables. If an element of the covariance matrix is zero, there is no relationship (correlation) between the two variables.

Related to the covariance matrix is the correlation matrix where all the variables have been scaled for their standard deviations. The correlation matrix is useful when one or more of the variables have much higher numerical values than the other variables. The scaling of the variables means that all variables will contribute to the analysis in roughly the same way. Mathematically the correlation matrix, **R**, is written as

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} = \frac{\sum_{i=1}^{n}(x_{ji}-\overline{x}_i)(x_{jk}-\overline{x}_k)}{\sqrt{(x_{ji}-\overline{x}_i)^2}\sqrt{(x_{jk}-\overline{x}_k)^2}} \qquad (33.4)$$

$$\mathbf{R} = \begin{pmatrix} 1 & & r_{1p} \\ & 1 & \\ r_{p1} & & 1 \end{pmatrix}$$

where the elements of the correlation matrix are given by r_{ij}. R is a square $p \times p$ matrix, where p is the number of variables. The diagonal elements of R are equal to one.

PCA is the systematic analysis of the covariance or correlation matrix. It can be shown that the eigenvalues are positive and the eigenvectors are orthogonal for both matrices [4]. The eigenvector equation for **C** is

$$\mathbf{C}u_i = \lambda_i u_i \qquad (33.5)$$

where u_i is the *i*th eigenvector and λ_i is the corresponding eigenvalue. By convention, the eigenvalues are placed in descending order, where λ_1 is the largest eigenvalue. In PCA, the eigenvectors are also called principal components. It can be shown that the first PC represents the largest source of variance in the data set. The percentage variation explained by the *i*th PC is given by

$$100 \times \frac{\lambda_1}{\sum_i \lambda_i} \qquad (33.6)$$

It is common with spectral data that the data set can be well approximated by a few principal components. As explained earlier, score values provide information about the relationship between different observations. The PCs form a basic set which can be used to approximate the original data set. For a single mean-centered observation, x_j,

$$x_j = \sum_{i=1}^{p} t_{ji}\mathrm{PC}_i = \sum_{i=1}^{A} t_{ji}\mathrm{PC}_i + E \qquad (33.7)$$

where t_{ji} are the score values, A is the number of principal components, E is the error when the number of principal components is less than the number of variables. Because the PCs are orthogonal, a direct expression for the score values can be given by the following equation.

$$t_{ji} = (\mathbf{x}_j - \overline{\mathbf{X}}) \bullet \mathrm{PC}_i \qquad (33.8)$$

Equation 33.7 is derivable from equation 33.6 by taking a dot product of both sides and exploiting the orthogonality of the principal components. The previous example is somewhat trivial because only two variables were involved.

Let us now consider another example with more variables. In Table 33.2, a set of data describing the properties of 43 raw materials is shown. The variables that describe the raw materials are labeled QV1–QV8. The variables QV1–QV8 describe different properties of the raw material such as moisture, assay, and particle size. Using commercial software, we can do a PCA analysis of the data set using the same approach that was used for the first data set, that is, scaling by standard deviation and mean centering. A few of the critical results are shown in Figures 33.3 and 33.4. The loading (principle component) plot shows some results that are clearly interpretable, Figure 33.3. The principle component plot shows how different variables relate to each other. In the plot the reader can observe that QV5 and QV8 are close to each other and therefore are well correlated to each other. QV1 and QV7 are also correlated. A plot of the score values for each 1 of the 43 raw materials is shown in Figure 33.4. The origin of the score plot corresponds to the average of the entire data set. The samples that are farther away from the origin are more likely to be possible outliers. The ellipse in Figure 33.4 is called the Hotelling T^2 ellipse and is showing the 95% probability level for outliers. The Hotelling T^2 ellipse is based on scaled, squared score values [2]. The T^2 value for observation i given below.

$$T_i^2 = \sum_{a=1}^{A} \frac{t_{ia}^2}{S_{ta}^2} \qquad S_{ta}^2 = \frac{\sum_{i=1}^{N} t_{ia}^2}{N} \qquad (33.9)$$

where A is the number of principal components and t_{ia} is the *a*th principal component score value for the *i*th sample. S_{ta}^2 is the variance of t_a, because the average of the score values is zero [2]. T^2 is closely related to the often-used parameter Mahalanobis distance. An important property of the T^2 statistic is that it is directly proportional to an F value, which is a statistical parameter that is rigorously related to a probability value.[1] The numerical value of the F value is dependent on the number of samples, principal components,

[1] $T_i^2 \frac{(N-A)N}{A(N^2-1)}$ is approximately F-distributed, see Ref. 2.

TABLE 33.2 Multivariable Quality Data Set

Primary ID	QV1	QV2	QV3	QV4	QV5	QV6	QV7	QV8
1	110	2	2	180	1.5	10.5	10	70
2	110	6	2	290	2	17	1	105
3	110	1	1	180	0	12	13	55
4	110	1	1	180	0	12	13	65
5	110	1	1	280	0	15	9	45
6	110	3	1	250	1.5	11.5	10	90
7	110	2	1	260	0	21	3	40
8	110	2	1	180	0	12	12	55
9	100	2	1	220	2	15	6	90
10	130	3	2	170	1.5	13.5	10	120
11	100	3	2	140	2.5	8	140	m
12	110	2	1	200	0	21	3	35
13	140	3	1	190	4	15	14	230
14	100	3	1	200	3	16	3	110
15	110	1	1	140	0	13	12	25
16	100	3	1	200	3	17	3	110
17	110	2	1	200	1	16	8	60
18	70	4	1	260	9	7	5	320
19	110	2	0	125	1	11	14	30
20	100	2	0	290	1	21	2	35
21	110	1	0	90	1	13	12	20
22	110	3	3	140	4	10	7	160
23	110	2	0	220	1	21	3	30
24	110	2	1	125	1	11	13	30
25	110	1	0	200	1	14	11	25
26	100	3	0	0	3	14	7	100
27	120	3	0	240	5	14	12	190
28	110	2	1	170	1	17	6	60
29	160	3	2	150	3	17	13	160
30	120	2	1	190	0	15	9	40
31	140	3	2	220	3	21	7	130
32	90	3	0	170	3	18	2	90
33	100	3	0	320	1	20	3	45
34	120	3	1	210	5	14	12	240
35	110	2	0	290	0	22	3	35
36	110	2	1	70	1	9	15	40
37	110	6	0	230	1	16	3	55
38	120	1	2	220	0	12	12	35
39	120	1	2	220	1	12	11	45
40	100	4	2	150	2	12	6	95
41	50	1	0	0	0	13	0	15
42	50	2	0	0	1	10	0	50
43	100	5	2	0	2.7	1	1	110

and probability level desired, α. Examination of equation 33.9 for two PCs shows that

$$F(\alpha) = C\left(\frac{t_1^2}{S_1^2} + \frac{t_2^2}{S_2^2}\right) \qquad (33.10)$$

where C is a constant. Equation 33.9 is an equation for an ellipse in the t_1, t_2 space. By convention, the Hotelling T^2 ellipse is usually drawn at the 95% probability level.

PCA can be viewed as a method for approximating the original data set. The approximation is based on a linear combination of the principle components where the amplitude coefficients are the previously described scores. The approximation is exact when the number of principle components equals the number of variables in the data set. For most spectral data sets, a small number of principle components (also called factors) can be used to approximate the spectral data set very well. The determination of the correct

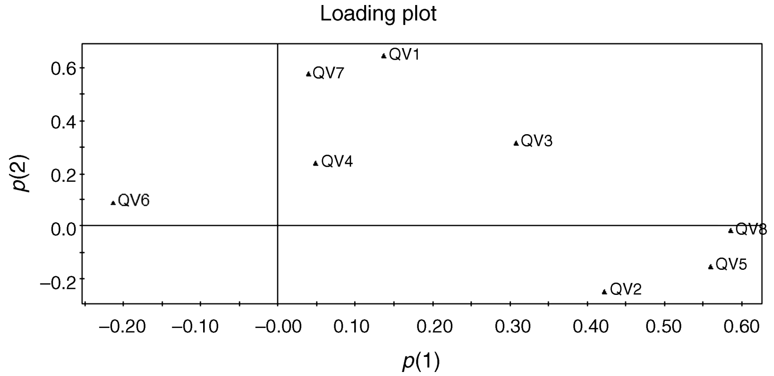

FIGURE 33.3 Loading plot for the data set in Table 33.2. The first principal component is plotted on the x-axis and the second principal component is on the y-axis. Variables that are close to each other are highly correlated.

number of factors can be done by a variety of numerical methods. Too many factors in the PCA model will over fit the data and the model will not predict reliably. Most multivariate analysis software packages will suggest a suitable number of principle components. The suggested number is usually a good starting point; however, it is best practice to verify the optimum number of principal components with additional independent test data.

Classification is an important application of chemometrics. Classification is the sorting of data into different groups. These groups can be quite diverse such as different sources or different quality grades of the same raw material. Chemometrics methods for raw material identification using NIR or Raman spectra as important but are relatively simple and are discussed elsewhere [5]. In this chapter, we will discuss soft independent modeling of class analogies (SIMCA) [6]. A method for classification of similar classes using multivariate analysis. PCA score plots sometimes show data sets to consist of several subgroups. For example, Figure 33.5 shows the score plot for the mid-IR spectra for a series of oils. Color coding the score plot clearly illustrates the differences between the four oils (olive, corn, safflower, and corn margarine).

SIMCA is designed to improve on this separation of classes by using the residuals from the PCA analysis. Residuals are the difference between the PCA model and the data. In the SIMCA analysis, a separate PCA model is built for each class in the training set. The average residual value for each class (S_0) is also calculated. Test or validation data are then fit to each PCA class model. The correct class is the class that has the best fit to the PCA model. The comparison is quantified by the use of the scaled residual S_0 (Dmod$X$) values. The equations are given below

$$
\begin{aligned}
S_i &= \sqrt{\frac{\sum_{k=1}^{K} e_{ik}^2}{(K-A)}} \\
S_0 &= \left(\frac{\sum_{i,j} e_{ij}^2}{(N-A-1)(K-A)} \right)^{1/2}
\end{aligned}
\qquad (33.11)
$$

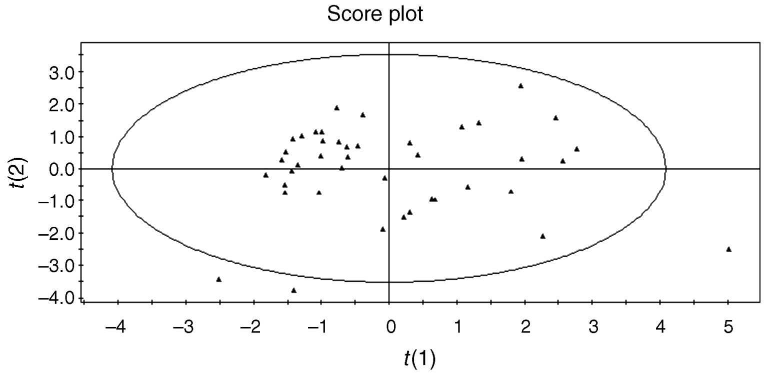

FIGURE 33.4 Score plot of the data set in Table 33.2. The ellipse is the Hotelling T^2 ellipse at 95% probability level. Samples outside the ellipse have a probability of greater than 95% of being statistical outliers.

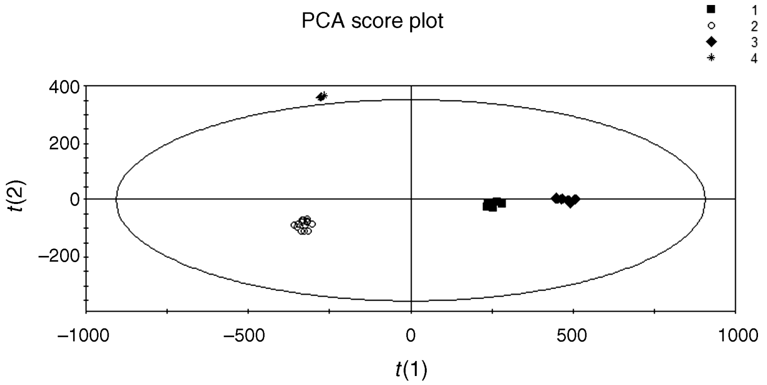

FIGURE 33.5 Score plot of mid-IR spectral data for a series of oils. Legend group 1: corn oil, group 2: olive, group 3: safflower, and group 4: corn margarine.

N is the number of samples, A is the number of principal components, K is the number of variables, s_i is the root mean square residual value for the ith sample, and e_{ij} is the spectral residual, that is, the difference between the spectra and the PCA model for observation i and variable j. If the test sample residual is close to the average residual for the entire class, then the sample has a high probability of belonging to the class. The relationship to actual probability values is possible because the scaled residual values $(S_i/S_0)^2$ in equation 33.11 are described by an F-distribution. The results of a SIMCA analysis are often displayed in a Cooman's plot. In a Cooman's plot, two classes are compared as shown in Figure 33.6.

A typical Cooman's plot is shown below. PCA models for corn oil and olive oil are used to predict the classification of a set of test samples. The test samples include olive, corn, corn margarine, safflower, and walnut oils. The different classes are color coded as shown in the legend. The x-axis on the Cooman's plot is the Dmod$X$ (distance to model) value for the corn oil; the y-axis is the same for olive oil. The red vertical line is the 5% probability level for the corn oil model, samples to the right of this line are probable outliers for the corn oil models. The red vertical line is the same for olive oil. Note most olive and corn oil test samples are correctly classified. Test samples form other classes are well separated from the oil and corn oil groups.

In many cases, spectral data requires mathematical transformations before multivariate analysis is performed [7]. The mathematical transformations are collectively referred to as spectral preprocessing. Derivative preprocessing is the most common form of spectra preprocessing with NIR spectra. Derivative preprocessing will eliminate or at least minimize the background variation associated with the NIR spectra of many pharmaceutical materials. The effects of a first derivative preprocessing on a typical NIR spectra are shown in Figure 33.7 (top and bottom). The first derivative removes the slowly varying baseline typical of NIR spectra of powders, positive and negative peaks correspond to regions where the slope of the raw spectrum has a positive or negative value. There are several methods for the calculation of spectral derivatives; however, they all start with the definition of first or second derivative from elementary calculus.

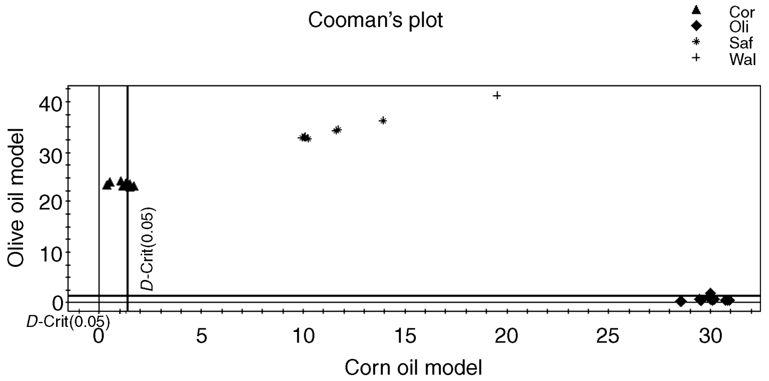

FIGURE 33.6 Cooman's plot comparing the olive and corn oil classes using a test set. Legend triangle: corn oil, diamond: olive oil, asterisk: safflower oil, and plus: walnut.

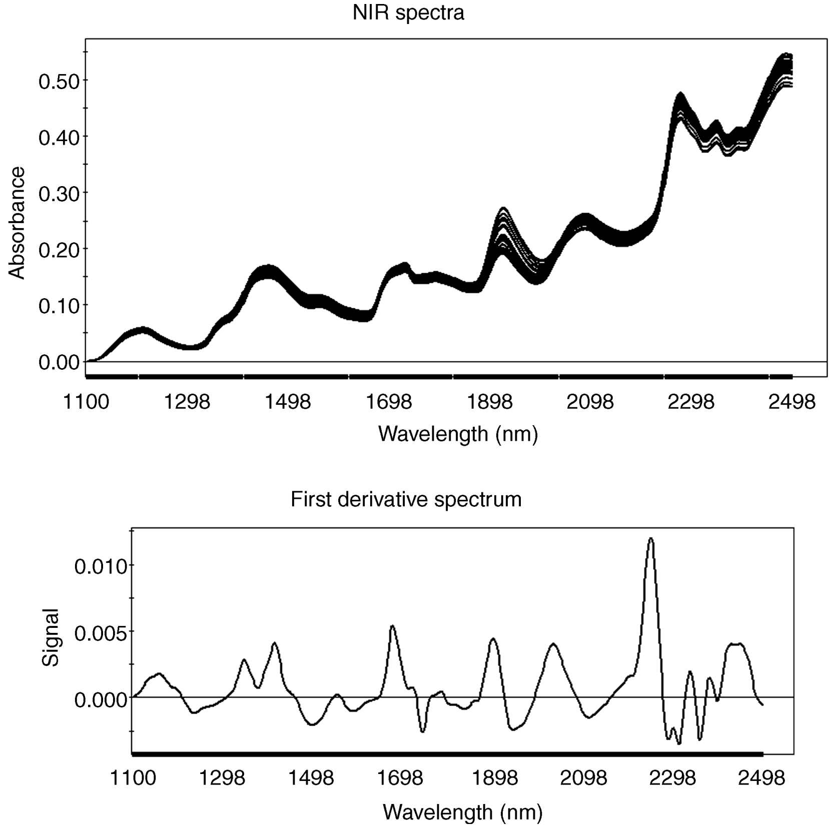

FIGURE 33.7 Effects of first derivative spectral preprocessing. *Top*: Several raw NIR spectra. *Bottom*: First derivative spectrum.

$$f'(x) = \frac{f(x+\Delta x)-f(x)}{\Delta x}$$

$$f''(x) = \frac{f(x+2\Delta x)-2f(x+\Delta x)+f(x)}{(\Delta x)^2} \qquad (33.12)$$

$$f''(x+\Delta x) \approx f''(x) + (\Delta x) f^{(3)}(x)$$

The two most common approaches for the calculation of derivatives are gap or Savitzky–Golay derivatives [8]. A gap derivative is based on the calculation of a running average for n points to the right and the left of a center point. The average values for the right- and left-hand sides are then used to calculate a finite difference derivative. There is an optional gap or separation between the right- and left-hand sides. Gap derivatives are described in Figure 33.8. A first-order gap derivative uses an n-point average to calculate a finite difference first derivative. Commonly the gap is set to zero in many applications.

Savitzky–Golay derivatives are based on fitting N points of the data to either a quadratic or cubic polynomial. The derivative is found by differentiation of the polynomial. For both methods of numerical differentiation, it is important to properly determine the number of points used in the averaging. Too few points can compromise the signal to noise; too many points will filter out important high frequency components of the data.

Standard normal variant (SNV) is a preprocessing method that is used to autoscale individual spectra [9]. The equation for SNV is given below.

$$\frac{x-\mu}{\sigma} \qquad (33.13)$$

where μ is the average value for the spectrum of interest and σ is the standard derivation of the numbers, which make up the spectrum. During SNV preprocessing, the average value

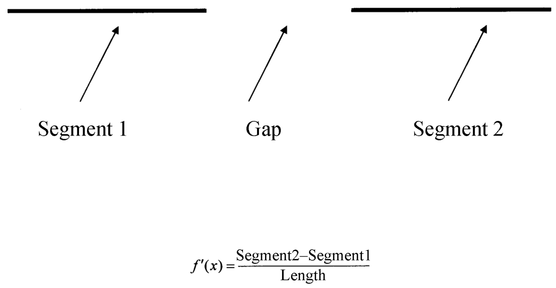

FIGURE 33.8 Illustration of gap derivative algorithm.

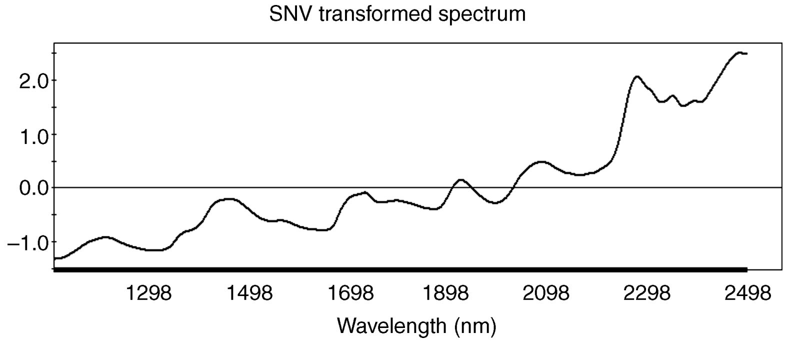

FIGURE 33.9 Effects of SNV preprocessing on spectrum from Figure 33.7.

for each spectra is subtracted, then the spectrum is divided by the standard derivation for the sample spectrum. After SNV preprocessing, the range of each spectra will be approximately −2 to 2. SNV processing can be used to correct for laser intensity variation in Raman spectra, and several kinds of path length variation in NIR spectra. The effects of SNV preprocessing are illustrated in Figure 33.9.

Multiplicative scatter correction (MSC) is a preprocessing method designed to eliminate background variation in NIR spectra due to scattering [10]. The effects of MSC are similar to SNV in many real-world applications; however, it is a distinct method. MSC uses the average spectrum of the entire data set and not individual spectra. The sample spectra are then regressed against the average spectrum producing slope and offset values at each wavelength for all samples in the data set. The slope and offset values are then used to correct the data set. Results of MSC preprocessing are illustrated in Figure 33.10. MSC preprocessing will remove the variation due to scattering in the data set but not change the average spectral value as SNV preprocessing does. The equations for MSC preprocessing are given below.

$$\begin{aligned} x_j &= a_j\mathbf{1} + b_j\boldsymbol{\mu} + \boldsymbol{\varepsilon}_j \\ x'_j &= (x_j - a_j)/b_j \end{aligned} \qquad (33.14)$$

$\boldsymbol{\mu}$ is the average spectrum, b_j is the slope, and a_j is the offset values for each wavelength. An important advantage of MSC preprocessing is that it can be used on filter wheel data, where the wavelength spacing is irregular and only a few wavelengths are typically measured.

There are many other preprocessing method used in chemometrics such as wavelets, orthogonal signal correction, and extended MSC (EMSC) [3, 11]. In practice, combinations of different preprocessing can also be used. However, the three methods discussed derivatives, SNV, and MSC are still the most commonly used preprocessed methods in the real-world applications.

Partial least squares (PLS) is an extension of PCA where both the *X* and *Y* data are considered [12, 13]. In PCA, only the *X* data is considered. The goal of the PLS analysis is to build an equation that predicts *Y* values (laboratory data) based on *X* (spectral) data. The PLS equation or calibration is based on decomposing both the *X* and *Y* data into a set of scores and loadings, similar to PCA. However, the scores for both the *X* and *Y* data are not selected based on the direction of maximum variation but are selected in order to maximize the correlation between the scores for both the *X* and *Y* variables. As with PCA, in the PLS regression development the number of components or factors is an important practical consideration. A short description of the PLS algorithm is given below, a more detailed discussion of the PLS algorithm can be found elsewhere [12, 13]. Commercial software can used to construct and optimize both PCA and PLS calibration models.

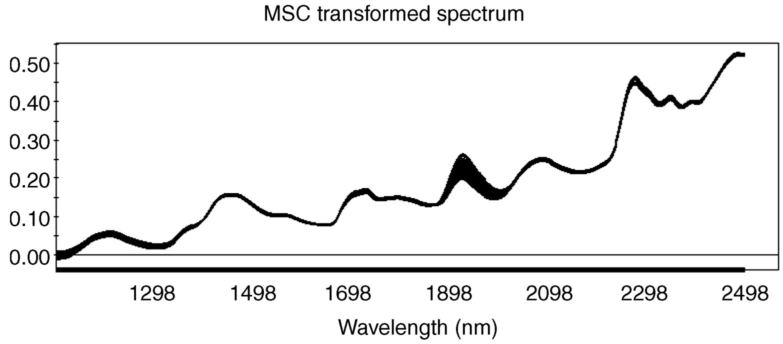

FIGURE 33.10 Effects of MSC preprocessing on spectrum from Figure 33.7. Note difference in *y*-axis from Figure 33.9.

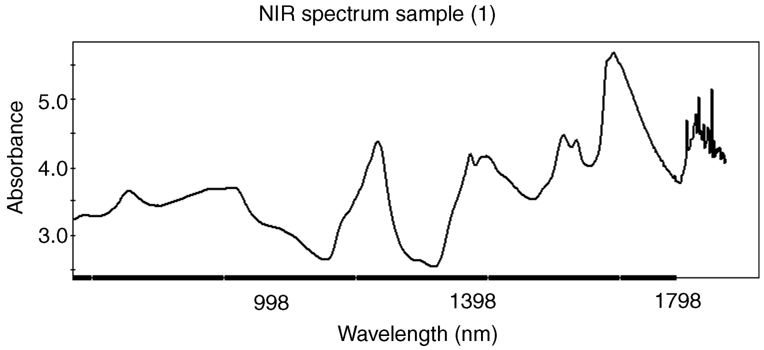

FIGURE 33.11 NIR transmission spectrum of a pharmaceutical tablet.

PLS decomposition of both *X* and *Y* data into scores and loadings is given in equation 33.15.

$$\begin{aligned}\mathbf{X} &= \mathbf{TP}^{\mathrm{T}} + \mathbf{E}\\ \mathbf{Y} &= \mathbf{UQ}^{\mathrm{T}} + \mathbf{f}\end{aligned} \tag{33.15}$$

The score matrices for **X** and **Y**, that is, **T** and **U**, are calculated together. This self-consistent approach allows for a set of scores and loadings that represent the variation in the *Y* data set. Therefore, the scores and loadings are much better than PCA scores and loadings for quantitative prediction. The algorithm proceeds by mean centering the data and then finding the first loading spectrum and first component scores. The prediction of a PLS method is summarized in the regression vector or coefficient, **B**. The predictions are related to the **x** sample data by

$$y = \mathbf{B} \cdot \mathbf{x} \tag{33.16}$$

We will now consider an example of a PLS calibration using NIR data. NIR transmission spectra from 155 tablets have been measured [14]. The tablet calibration set included samples with a range of assay values and lots of production samples in order to capture the typical variations seen in the tablets. After scanning with the NIR instrument, the amount of active ingredient in each tablet was measured by HPLC. The weight of the tablet was about 800 mg and the target value for the drug content was 200 mg. We will use chemometrics to develop a model for the amount of active. This model could be used to monitor the stability of tablets over time in a nondestructive manner. For brevity, we will only outline the analysis procedure. Typical NIR transmission spectra for the pharmaceutical tablet are shown in Figure 33.11. The broad, overlapping spectra with a considerable background is typical of NIR spectra. Derivative preprocessing can be used to remove the unnecessary background and elucidate the underlying peaks in the spectra. A first derivative spectrum is shown in Figure 33.12.

A calibration curve showing the predictions of the PLS model versus the laboratory data is shown in Figure 33.13. The clear quality of the calibration curve is evident. The calibration curve can be evaluated by several methods including outlier detection and removal and optimization of the spectral range used for PLS calibration. A detailed discussion of these issues can be found in the references [12, 13]. Common examples of quantitative methods done with NIR data and PLS regression are moisture, particle size, and assay [7].

Method validation for NIR or Raman spectroscopic methods using chemometrics is outlined in United States Pharmacopoeia (USP) Chapter ⟨1119⟩ [15]. The criteria for method validation are the same as other quantitative analytical methods, such as accuracy, precision, intermediate precision, linearity, specificity, and robustness. Since these methods are statistical in nature and are based on a previously

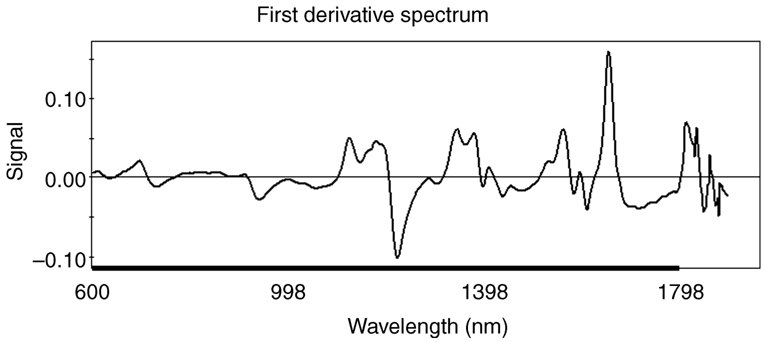

FIGURE 33.12 Spectrum from Figure 33.11 after first-derivative preprocessing.

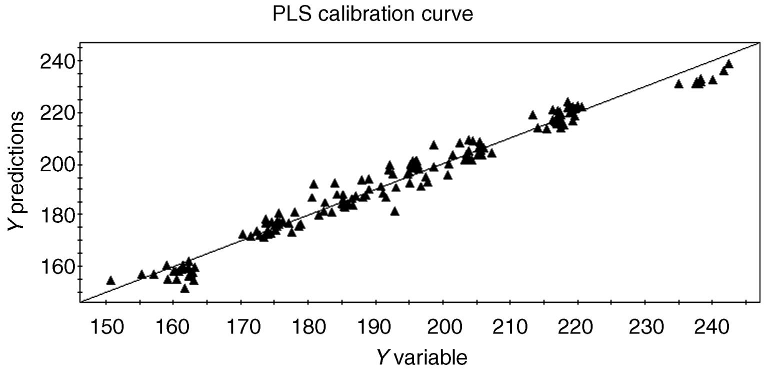

FIGURE 33.13 Calibration curve for PLS method for tablet assay value.

validated analytical method, the validation of MVA methods is somewhat different than traditional analytical methods. In this chapter, we will briefly discuss chemometric method validation, a more detailed discussion can be found elsewhere.

Accuracy of the MVA method refers to how closely the MVA method and the original laboratory method compare. The accuracy of a chemometric method is evaluated by comparing the predictions of the MVA model with the actual laboratory data for a set of validation samples. The validation samples should be from lots of material not used in the original calibration set. There are several mathematical ways to express the accuracy. The most commonly used approach is the standard error of prediction (SEP). The SEP is defined in equation 33.14.

$$\mathrm{SEP} = \sqrt{\sum \frac{(\mathrm{NIR}-\mathrm{LAB})^2}{n}} \qquad (33.17)$$

where n is the number of validation samples. The SEP value should be close to the actual error of the original laboratory method. The actual error of the laboratory method should include normal sources of variation such as different analysts, different instruments, different materials analyzed on different days.

The *linearity* of a multivariate method is an important topic. Typically the linearity of a chromatographic method is evaluated by the R^2 (coefficient of determination) value of a recovery measurement. R^2 is the fraction of variation in the y-variable explained by the linear fit; r is the correlation coefficient that quantifies the correlation between the x and y variables [16]. R^2 is often used in the analysis of chromatography recovery studies [16]. In contrast, R^2 is not a good statistical parameter for multivariate methods. The linearity of a multivariate method is evaluated by the inspection of the residual values, that is, the difference between the predictions of the multivariate model and the actual laboratory data. A linear model will have residuals that are random, that is, normally distributed. A nonlinear model will have residuals that are not normally distributed. The USP Chapter ⟨1119⟩ states that the linearity should be evaluated by examination of the residuals, but no specific threshold or criteria are given. In the opinion of this author, visual inspection of the residuals using a normality plot is recommended. In Figure 33.14, a normality plot of residuals is shown. The data points in Figure 33.14 do follow a straight line, indicating a normal distribution of residuals, consistent with a linear model or

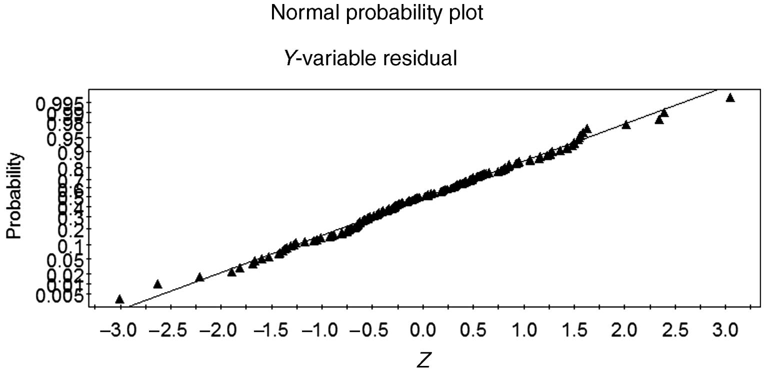

FIGURE 33.14 Normal probability plot for residuals. When the residuals fall on straight line, the calibration under consideration is linear.

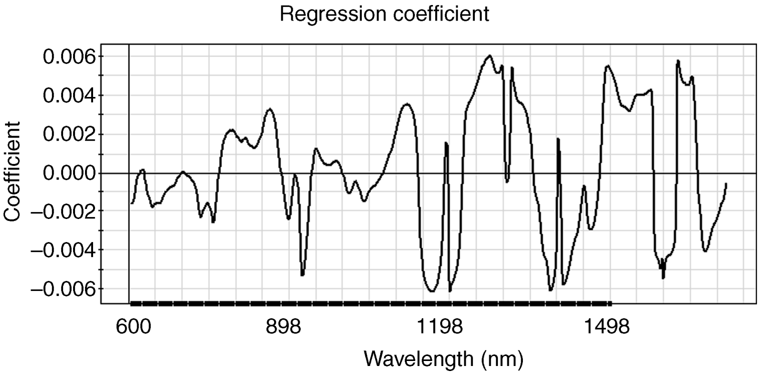

FIGURE 33.15 Regression coefficient from the PLS model for tablet assay described earlier. The regression coefficient is a method for documenting and examination of which wavelengths are most important for the PLS calibration.

calibration [13]. In some cases, the linearity of the model can be improved by removing some of the points in the normality graph, which are probable outliers.

Method *specificity* is the extent the multivariate calibration is specific to the analyte of interest. With a PLS calibration, the specificity is documented by the regression coefficient of the calibration. The regression coefficient shows which wavelengths are most important for the PLS calibration. Important wavelengths may have either positive or negative regression coefficient values. The most important wavelengths should correspond to the absorption peaks of the analyte of interest. For example, the regression coefficient for a moisture model will have peaks at the known water absorbance band locations. In practice, the regression coefficient is often documented in the method development report. A regression coefficient from the PLS calibration for tablet assay described earlier in this chapter is shown in Figure 33.15.

The *range* of a multivariate calibration method is determined by the range of laboratory values in the calibration and validation data sets. A method is validated over the range of laboratory values of the samples used in the independent validation set. The range of the validation samples can also depend upon the application of the method. For example, in-process testing or testing where a limited number of samples are available may require a fairly small range of values because samples outside of a small range are not available or do not exist.

This chapter has briefly summarized the essential principles of chemometrics and their application to spectral data. There are many applications of chemometrics that have not been discussed here due to space limitations. Two important examples of this are chemical imaging and batch monitoring. Raman and NIR chemical imaging have been applied to pharmaceutical products including tablets and drug-coated stents [17]. Both Raman and NIR chemical imaging methods typically require chemometrics for the creation of useful images. Batch monitoring involves the use of multivariate control charts based on score plots developed from a collection of good batches [18]. Batch monitoring can be used with spectral or process data. Batch monitoring has been used to monitor a variety of complex pharmaceutical products to improve yields and provide improved process understanding [2]. In summary, chemometrics is a vital part of process analytical technology, quality by design, and the overall future of both pharmaceutical development and manufacturing.

REFERENCES

1. US Food and Drug Administration. *Guidance for Industry Process Analytical Technology*, 2004.
2. Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikström C, Wold S. *Multi- and Megavariate Data Analysis*, Umetrics AB, Umeå, Sweden, 2006.
3. Strang, G. *Computational Science and Engineering*, Wellesley-Cambridge Press, Wellesley, MA, 2007.
4. Johnson RA, Wichern DW. *Applied Multivariate Statistics*, Prentice Hall, Upper Saddle River, NJ, 2002.
5. Long F. *Am. Pharm. Rev.* 2008; Sept/Oct.
6. Wold S. *Pattern Recogn.* 1976;8:127–139.
7. Siesler H, Ozaki Y, Kawata Y, Heise H. *Near-Infrared Spectroscopy: Principles, Instruments, Applications*, Wiley-VCH, Weinheim, 2002.
8. Savitzky A, Golay MJE. *Anal. Chem.* 1962;36:1627–1639.
9. Barnes RJ, Dhanoa MS, Lister SJ. *Appl. Spec.* 1989;43: 772–777.
10. Martens H, Jensen SA, Geladi P. *Proceedings from Nordic Symposium on Applied Statistics,* Stokkand Forlag Publishers, Norway, 1983, pp. 205–233.
11. Martens H, Stark EJ. *Pharmaceut. Biomed. Anal.* 1991;9: 625–635.
12. Wold S, Josefson M. Multivariate analysis of analytical data. In: Meyers R, editor, *Encyclopedia of Analytical Chemistry*, Wiley, New York, 2000, pp. 9710–9736.

13. Esbensen K. *Multivariate Data Analysis—In Practice*, CAMO Process AS, Oslo, Norway, 2002.
14. Ritchie G. *Software Shootout Data,* International Diffuse Reflection Conference, Chambersburg, PA, 2002.
15. US Pharmacopoeia. *Near IR Spectroscopy*, Chap. ⟨1119⟩.
16. Miller J, Miller J. *Statistics and Chemometrics for Analytical Chemistry*, Prentice Hall, Harrow, UK, 2000, p. 137.
17. Balss K, Long F, Veselov V, Akerman E, Papandreou G, Maryanoff C. *Anal. Chem.* 2008;80:4853–4859.
18. Wold S, Cheney J, Kettaneh N, McCready C. *Chemometr. Intell. Lab. Syst.* 2006;84:159–163.