

Processing Techniques for Handling Multiple Spectra

David L. Wooton



This article discusses several mathematical processing methods for handling, assessing, and studying groups containing multiple spectra. The processing techniques feature standard mathematical tools that are used to study a series of numerical results, namely averaging, mean centering, and standard deviation.

The role of infrared (IR) spectroscopy for the study and control of processes has expanded during recent years.

People are finding that this analytical tool is probably one of the most valuable for the control of plant processes and for the monitoring of product streams when compared to other analytical techniques (1, 2). Classical interpretation techniques are very effective when it comes to the study of a single IR spectrum (3). When it comes to having to deal with a number of spectra, the classical techniques need a little help. Handling data in the form of an IR spectrum is different than results obtained from elemental analyses or tests such as acid number. The spectrum is an array of numbers, and may be expressed as a vector instead of a single scalar value for each measurement. This numeric array must be treated as a vector quantity and requires different numerical approaches when one is studying a series of samples. This article is designed to provide analysts with a few more tools to handle the interpretation, and for the study of sets of IR spectra obtained from a series of related samples.

Discussion

When using an analytical technique such as elemental analysis or acid number, a single scalar value for each sample is obtained. Good analytical chemists will test the sample analytically multiple times to determine the precision

and repeatability for the method as used. This approach is also used for IR analyses if the method of determination involves the measurement of a single peak height or peak area.

As an example, the results from the series of analyses shown in Table I will be discussed.

Pair-wise Data Set Comparisons

One approach to study the data is to compare two results with a pair-wise approach. This is achieved by the comparison of neighboring pairs of results. Are the values similar or are they very different? As an example, we compare the first two results, 54.2 and 53.2. These results have an average value of 53.7 and a numerical difference of 1.0 (1.9% difference). The same approach can be used with IR spectra.

A simple visual comparison of the spectra is obtained from a spectral overlay, as shown in Figure 1. In the visual comparison, one looks to see if all the peaks line up with each other and that they are generally the same intensity, and if there is a presence (additional) or absence of peaks. In most cases one knows the chemistry of the samples that are being studied. The objective is to assess differences

Table I. Numerical results.

Sample number	Analyses results	Cell deviation
1	54.2	-0.3
2	53.2	0.7
3	55.2	-1.3
4	52.8	1.1
5	53.7	0.2
6	54.1	-0.2
7	54.3	-0.4
8	54.8	-0.9
9	53.9	0.0
10	54.1	-0.2
11	53.1	-0.8
12	54.9	-1.0
13	54.3	-0.4
14	52.7	1.2
15	53.4	0.5

David L. Wooton

is a consultant with Wooton-Consulting, 17145 Tulip Poplar Road, Beaverdam, VA 23015, (804) 227-3418, fax (804) 227-9426, e-mail dave_wooton@att.net.

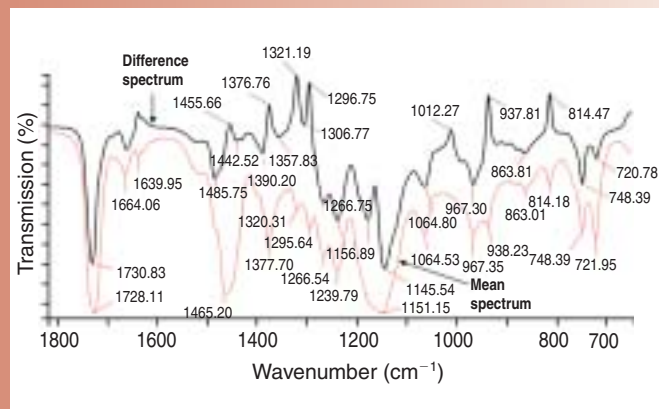
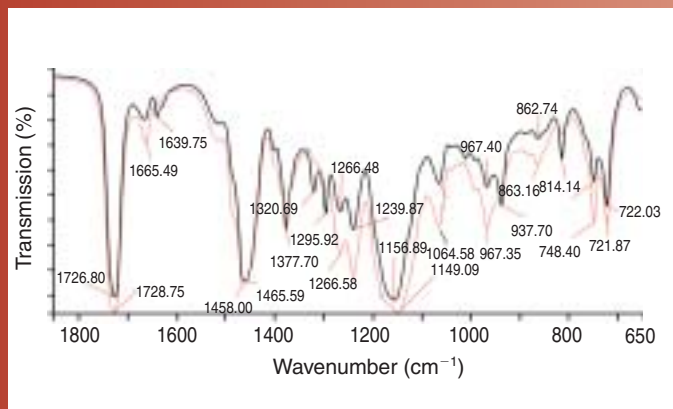
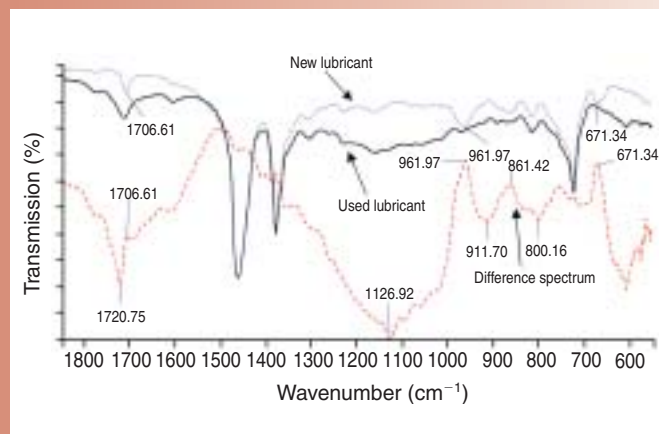


Figure 1 (above). Special overlay of polymethacrylates.

Figure 2 (upper right). Mean spectrum and difference of polymethacrylates spectrum.

Figure 3 (lower right). New lubricant spectrum, used lubricant spectrum, and difference.



that exist between production batches or source variability with raw materials. With the two examples shown in Figure 1, there are some spectral differences that are readily observed with spectral frequencies of 1639 cm⁻¹, 1320 cm⁻¹, 1295 cm⁻¹, 967 cm⁻¹, and 937 cm⁻¹. There are other, more subtle differences, and these become more difficult to observe as they get smaller.

Just as in the case of two numerical values discussed earlier, the two spectra can be averaged together to produce a mean spectrum or subtracted to obtain a difference spectrum (Figure 2). The analyses of used lubricants have long been an application for the difference spectrum technique (4, 5). The analyst studies the difference spectrum of the new oil from the used oil (Figure 3). This allows the removal of any unchanged features, thus expanding the changes that occurred during the lubricant's operation. The evaluation of the difference spectrum is not as intuitive as a normal spectrum because individual peaks can be both positive and negative. One has to question whether a peak is a real peak, or whether it is a distortion resulting from the subtraction. For example, in Figure 3, are the negative absorbance peaks at 961 cm⁻¹ and 861 cm⁻¹ real peaks caused by the absence (or a reduction in concentration) of a component in the sample, and likewise, is the positive absorbance peak at 911 cm⁻¹ the real peak caused by the addition or increase in concentration of a component?

These questions always permeate the entire difference spectrum, making them more difficult to interpret when compared to traditional transmittance or absorbance spectra. Sometimes, as with this example, the mean spectrum will assist in the peak assignments. The difference spectrum allows us the ability to observe the variations between the two spectra. If the peaks can be assigned, then one may assume that the sample is a mixture of two or more components.

The difference and mean spectra are obtained in the same manner as with single value results shown in Table I, except that with the IR spectrum, it was necessary to handle the two spectral arrays of numbers as vector entities. Before any calculations are performed on the spectra, it is important to ensure that the intensity scale for the spectra is in absorbance and not percent transmittance. Calculating the average (equation 1) or differences (equation 2) of the intensity at each wavelength (or digitization interval) for the two spectra as single values achieves this goal. A new array of these calculated wavelength-intensities is constructed that represents the new spectrum.

$$Mean_I_j = (I_{a,j} + I_{b,j}) / 2 \quad [1]$$

$$Diff_I_j = (I_{a,j} - I_{b,j}) \quad [2]$$

where j is the wavelength or index value in the array, $I_{a,j}$ is the intensity from spectrum a at wavelength j , $I_{b,j}$ is the intensity from spectrum b at wavelength j . $Mean_I_j$ is the mean intensity for the wavelength j and $Diff_I_j$ is the difference intensity for the wavelength j . The $Mean_I$ array becomes the mean spectrum and the $Diff_I$ array becomes the difference spectrum.

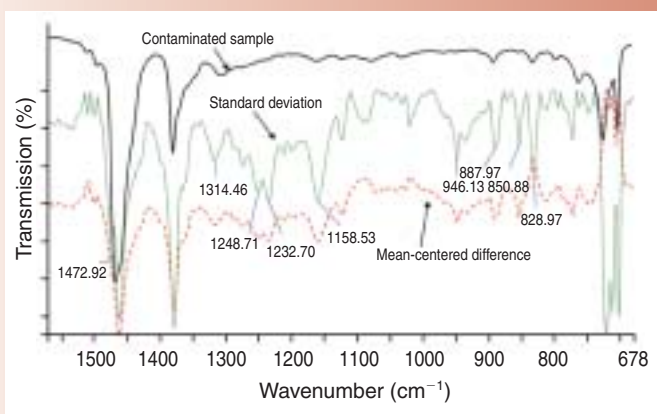
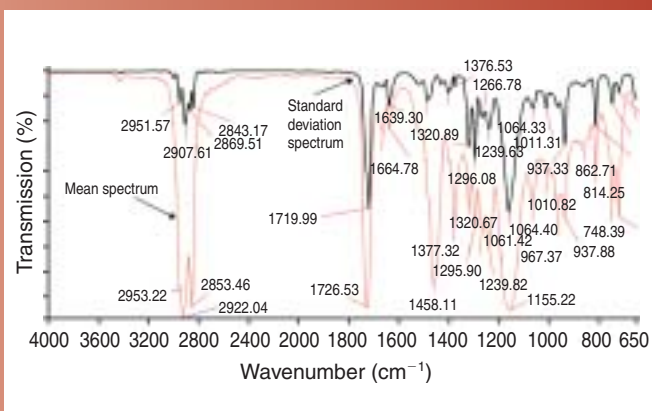
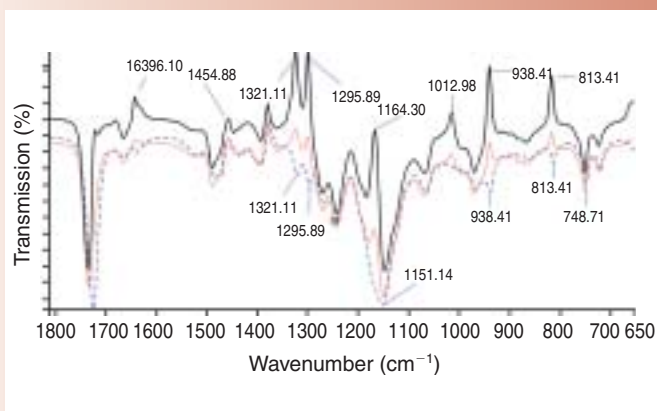


Figure 4 (upper left). Mean centering difference spectra.

Figure 5 (upper right). Mean spectrum and standard deviation.

Figure 6 (left). Spectra of contaminated sample, mean-centered difference, and standard deviation. The difference and standard deviation spectra have been auto-expanded for better viewing.

$$\text{Mean}_I_j = \sum_{k=1}^n (I_{k,j})/n \quad [3]$$

where k represents the individual spectra in the set of n spectra, j is the wavelength (or wavelength index), $I_{k,j}$ is the intensity from spectrum k at wavelength j , and n is the total number of spectra being studied. The Mean_I array becomes the mean spectrum.

When one is working with a series of spectra, one is typically not interested in baseline shifts between the spectra. A baseline shift can be looked at as a bias in the intensity data between the spectra, similar to lab-to-lab bias data. All the intensity data within the spectra is shifted by this bias value. Consequently, unless one is interested in the effects of this spectra bias, it can be removed from the spectral series by first applying a baseline correction before doing any calculations on the spectra.

This means spectra can be used and studied in a similar manner to the average of an analytical results series. As with Figure 2, the mean spectrum can be used to indicate spectral features that are real, when combined with the difference spectrum, providing a better understanding of the variations in chemistry.

In the comparison of a series of repetitive analyses, the American Society of Testing and Materials (ASTM) procedure E-691 addresses the determination of the average of the series of analysis results and the cell deviation (6). Like the difference between two numbers, the cell deviation is the difference of the results from each analysis when compared to the mean or average value. This cell deviation tells how much the individual value differs from the average of the set of analyses (Table I). When an individual sample has a large

Multiple-Analyses Data Set

As the number of analyses performed for a method increases, it is still important to know the average or mean of the data set. If the data were taken from multiple analyses of the same sample, the average provides a better value for the real number (assuming that the method is accurate). If the data are taken from the analysis of multiple different samples, the average of the sample set is a value equivalent to the mixture of all the samples blended together. This average value can be used to set the target value for a specification, as used for a certificate of analysis, or it can help to define control limits for statistical process control of the product quality.

With a series of IR spectra, it is possible to evaluate our analysis results by a visual overlay of the multiple spectra. However, as the number of spectra increases, observing the individual spectra and the changes within the individual spectra becomes more difficult. When the difference between samples is major, one can observe it. However, as the differences get smaller and the number of samples gets larger, they become even more difficult to observe and assess without some form of spectral processing or number crunching. When we treat the series of spectra as we do, the series of numbers provides the ability to obtain the average or mean spectrum, even with a large number of spectra. This averaging is designed to show an average spectral response for the sample data set (equation 3).

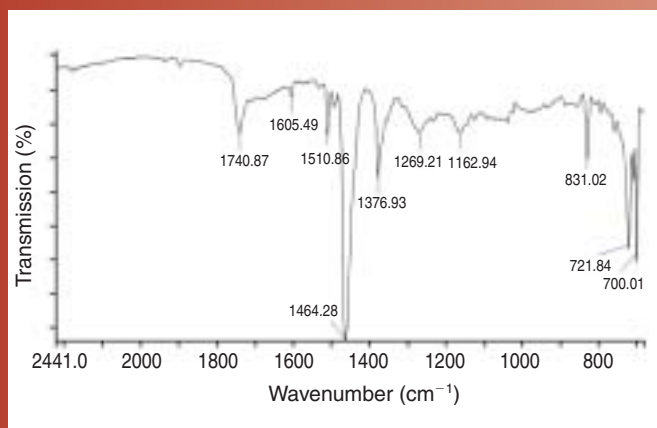


Figure 7. Standard deviation spectrum of a series of process-produced olefins showing chemistry of variation.

cell deviation, one might want to scrutinize this sample for analysis error or sample impurities. The same approach can be applied to a series of spectra (equation 4).

$$\text{Mean_Diff_}I_j = \text{Mean_}I - I_{k,j} \quad [4]$$

This approach is sometimes referred to as “mean centering” of the difference spectra (7). It can generate difference spectra for each analysis, showing the variation of the individual spectra from the set mean or average spectrum (Figure 4). With mean centering of the difference spectra, the differences between the individual spectra are magnified by the removal of the spectral similarities. This allows a rapid visual assessment of the possible differences. As an overall technique, mean centering is an accepted approach for handling spectral arrays with chemometrics. It can also be an excellent approach when looking for a contamination in a series of samples and the noncontaminated sample cannot be identified. The limited number of difference spectra from the array of spectra in Figure 4 shows both components that are in reduced or increased concentration from the data set mean. If analysts were attempting to do quality control on this series of samples, the reduced or increased concentration seen can be used for the process control criteria. As the number of spectra in the series grows, the series of difference spectra becomes very complex, and so there is a need for a method to control its growth while still looking for spectral differences. There is also need to have a method to determine whether a spectral difference is within the normal difference limits for a given set of samples.

Returning to ASTM E-691 for the determination of the variance of a series of results, one can determine the *standard deviation* (6), a measure of how widely the results are dispersed around the average value (the mean). It is also defined as the square root of the variance. For the series of results in Table I, the standard deviation is 0.8. This provides a measure of the variability of the data set. If the data from Table I are a normal population of data, then about 68% of the observations fall within 53.9 ± 0.8 and 95% fall within 53.9 ± 1.6 (8). When studying a series of spectra, the same

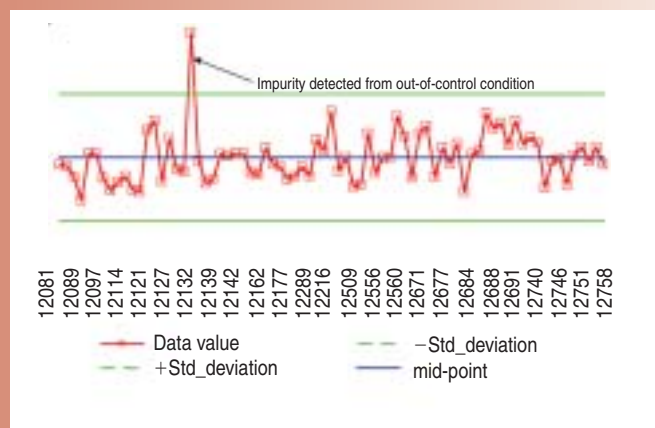


Figure 8. Moving average control chart showing out-of-control data.

determination can be applied. A standard deviation spectrum (Figure 5) can be generated using equation 5 (8):

$$\text{Std_dev}_j = \sqrt{\frac{\left(n \sum_{k=1}^n I_{k,j}^2 - \left(\sum_{k=1}^n I_{k,j} \right)^2 \right)}{n(n-1)}} \quad [5]$$

where k represents the individual spectra in the set of n spectra, j is the wavelength (or wavelength index), $I_{k,j}$ is the intensity from spectrum k at wavelength j , and n is the total number of spectra being studied. The Std_dev_j array becomes the standard deviation spectrum.

This approach to the study spectra provides the ability to develop quality control specifications around the use of spectral data.

This standard deviation spectrum has the same characteristics as the standard deviation generated from Table I. The spectral intensity variations at any wavelength are related to the variation seen within the set of spectra. The larger the intensity in the standard deviation spectrum, the greater that wavelength's intensity varies in the data set and consequently the larger the variation in the concentration of the component that is being measured. The data set of samples used to generate the spectra from Figure 5 had large differences in their chemistries. This is reflected in the fact that the standard deviation spectrum has a reasonably high intensity. If there were only very small chemistry differences between the samples in our data set, the standard deviation spectrum would be very small or near zero when the samples are close to being identical.

Examples

With the techniques discussed in mind, one can make comparisons of multiple spectra using the mean and standard deviation spectra, as well as study the individual samples with the aid of the mean-centered difference spectra. Looking at a series of spectra from multiple production lots or multiple sources (such as raw materials), one can do so with the knowledge of location of the center point of the series and by how much the series varies. If the variation is greater than the data set variation, one knows that the samples have differences (in composition) that may need to be addressed.

This approach to the study spectra provides the ability to develop quality control specifications around the use of spectral data. Analyzing the same sample multiple times, as one does for any new analytical technique, provides the means to determine the repeatability of the new method. The repeatability can be defined by the standard deviation of the series of repeated analyses. The larger the intensity of the standard deviation spectrum, the worse the repeatability of the method. If the repeatability provides deviations greater than required for the specification, then it is necessary to consider method improvement.

This type of numerical processing of spectral data can lead to a very effective method for quality control of a process.

Once a method is established with known variance, it is possible to study the plant production variance or sample-to-sample (batch-to-batch) variances. This can then be achieved by the analysis of multiple production lots or multiple samples. The intensity of the standard deviation spectrum is the guide to the variation in the sample differences. The advantage here is that the results help to provide an understanding of the chemistry that is causing this variance in the product stream or the set of samples.

Statistically speaking, it is not a normal practice to determine the standard deviation between two samples. However, if one determines the standard deviation between the mean and the sample in question (similar to the mean-centering difference spectrum) useful information can be abstracted; especially if one is attempting to provide quality control of the sample. As an example, a series of known good-quality samples from our plant process is evaluated. From this series of samples a mean spectrum and a standard deviation spectrum are generated. This defines the average sample spectrum for the process and how much this spectrum is expected to vary.

A sample of unknown quality is compared with these spectra. From the standard deviation obtained from this un-

known sample spectrum and the mean spectrum, it is possible to determine how much this sample varies from the set of good samples. If the spectral intensities of the standard deviation spectrum are smaller than or equal to the good samples' standard deviation spectrum, our unknown sample can be defined as good quality. On the other hand, if the spectral intensities are larger the sample can be defined as poor quality.

Figure 6 shows an example of an impurity in an olefin sample. The mean-centered difference spectrum shows some differences. The standard deviation spectrum as compared against the mean-centered difference spectrum showed peaks that were larger than the standard deviation spectrum for the set of olefins. If the spectral features of the mean-centered difference spectrum are larger than that of the standard deviation spectrum, that feature will be outside the normal distribution limits for the data set. One can therefore define this sample as bad or poor quality. For this bad quality sample, valuable process information is obtained from the standard deviation spectrum or mean-centered difference spectrum. This will lead to an assessment of the chemistry and a determination of what is causing it to be bad. The impurity in this example was identified as an alkyl phenol by the study of its spectrum in relation to known spectral libraries. In addition, the intensity of the peak can be used to determine the concentration of the component change, whether it is an addition or an omission (or reduction in level) to the sample. With this example, the use of known standards allowed the determination of the concentration of the alkyl phenol to be 0.1%.

There is always a strong need to study the variation of samples produced from plant processing operation. Every time a process is run in the plant, there are always product variations. There are many technical reasons for these variations and knowing the source of these can allow plant operations to improve the product quality. Figure 7 shows an example of a variation in a mixed olefin stream. In this example, one can observe a small variation in olefin content and type from the typical olefinic bond vibrations (1605, 1510, and 831 cm^{-1}). However, one can also observe the presence of an ester (1741, 1269, and 1163 cm^{-1}). This chemistry is not part of the process variables and thus comes from a contamination of the sample during processing or an oxidation reaction of the olefin. With this information, plant personnel know where to start looking for the process improvements, and continuous application of the standard deviation spectra will allow the ability to show the improvements in progress.

This type of numerical processing of spectral data can lead to a very effective method for quality control of a process. Using both the standard deviation spectrum and the mean-centered difference spectrum in combination works better than using the difference spectrum alone. One can produce a trend plot for a series of production samples with the maximum spectral intensities for the entire spectrum or from a defined spectral region within the spectrum. Using this method, upper and lower control limits based on production variations can be established. Figure 8 shows an example of a

run chart (8), which is a graphic display of the analysis results for a series of samples. Measuring the maximum peak intensity within a spectral region of the mean-centered difference spectrum developed the run chart in this figure. These maximum peak intensities were plotted for each sample in the series and became the run chart evaluation criteria. Along with knowledge of the potential contamination and the peak intensity for the standard deviation spectrum, one can determine good spectral regions to study for impurities. This figure also displayed the upper and lower control limits that were determined from the standard deviation of the sample series. One of the samples in the series has results that are above the control limits. This out-of-control result is an indication of a contamination in the sample. Figure 8 shows the mean-centered difference spectrum for this out-of-control sample along with the standard deviation spectrum for the series. One can observe from the difference spectrum that the contamination in this questionable sample is a carboxylic acid.

Conclusion

In conclusion, when studying a series of multiple samples by IR spectral analysis, several numerical processing techniques can help. These techniques are the same numerical or statistical methods used by analytical chemists for the evaluation of results from a series of singular value (scalar) results. Dif-

ference, mean centering, and standard deviation techniques are all valuable tools for spectral processing. They allow spectra to be used in the traditional statistical processing techniques that are in common use for many quality control regimes.

Although these techniques have been applied to IR spectra, they can also be applied equally to any x, y -data analyses. This includes ultraviolet-visible, near-IR, and nuclear magnetic resonance spectra, as well as high performance liquid chromatograms, gas chromatograms, and titration plots.

References

1. J.J. Workman, *Appl. Spectrosc. Rev.* 34(1), 1–89 (1999).
2. D.L. Wooton, *Appl. Spectrosc. Rev.* 36(4), 315–332 (2001).
3. N.B. Colthup, L.H. Daly, and S.E. Wiberley, *Introduction to Infrared and Raman Spectroscopy, 3rd Ed.* (Academic Press, New York, 1990).
4. D.L. Wooton, J.L. Milner, J.G. Damrath, and K. Yatsunami, *Japan Petrol. Inst.* 62, 19 (1987).
5. J.P. Coates and L.C. Setti, *ASLE Trans.* 29(3), 394 (1986).
6. *Annual Book of ASTM Standards, Vol. 14.02, "E 691-99"* (American Society for Testing and Materials, West Conshohocken, PA, 2000).
7. J. Duckworth, *Applied Spectroscopy, a Compact Reference for Practitioners* (J. Workman and A. Springsteen, eds.), Chapter 5, p. 165, (Academic Press, New York, 1998).
8. E.L. Bauer, *A Statistical Manual for Chemists, 2nd Ed.* (Academic Press, New York, 1971). ■