Application of New Least-squares Methods for the Quantitative Infrared Analysis of Multicomponent Samples

DAVID M. HAALAND and ROBERT G. EASTERLING
Sandia National Laboratories, Albuquerque, New Mexico 87185

Improvements have been made in previous least-squares regression analyses of infrared spectra for the quantitative estimation of concentrations of multicomponent mixtures. Spectral baselines are fitted by least-squares methods, and overlapping spectral features are accounted for in the fitting procedure. Selection of peaks above a threshold value reduces computation time and data storage requirements. Four weighted least-squares methods incorporating different baseline assumptions were investigated using FT-IR spectra of the three pure xylene isomers and their mixtures. By fitting only regions of the spectra that follow Beer's Law, accurate results can be obtained using three of the fitting methods even when baselines are not corrected to zero. Accurate results can also be obtained using one of the fits even in the presence of Beer's Law deviations. This is a consequence of pooling the weighted results for each spectral peak such that the greatest weighting is automatically given to those peaks that adhere to Beer's Law. It has been shown with the xylene spectra that semiquantitative results can be obtained even when all the major components are not known or when expected components are not present. This improvement over previous methods greatly expands the utility of quantitative least-squares analyses.

Index Headings: Fourier transform infrared spectroscopy; Quantitative infrared spectroscopy; Least-squares analysis; Xylene.

Received 1 February 1982.

INTRODUCTION

Quantitative infrared spectroscopy has gained in popularity with the capability of obtaining digitized infrared

spectra. Antoon *et al.*[1] have applied linear least-squares regression methods to the infrared spectra of multicomponent mixtures in order to analyze quantitatively the composition of the individual components. Assuming a linear relationship between concentration and absorbance (i.e., assuming Beer's Law is obeyed), these techniques have been successful in the quantitative analysis of multicomponent mixtures even in those cases where there is complete overlap of the infrared spectral features. The inclusion of all the data in the spectral region of interest also significantly improves the precision and accuracy of the results. The methods of Antoon *et al.*[1] have been applied with success in the quantitative analysis of polymer components and the mineral composition present in coal.[2]

Earlier, we presented least-squares methods for improving the sensitivity of the quantitative analyses of the infrared spectra in regions of the spectrum where only single components were present.[3] We were able to improve the sensitivity by simultaneous least-squares fits of the spectra and the baselines. Alternatively, a least-squares derivative fit of the spectra was used to correct for slow nonlinear variations in the baseline. These methods were shown to improve the detection of trace gases by factors of 5 to 7 for gas molecules of low molecular weight and allowed detection even in those cases where the signal was less than the noise. It is useful to combine our earlier methods with those of Antoon *et al.*[1] so that automatic spectral baseline corrections can be incorporated into the least-squares regression analysis of multicomponent mixtures with overlapping spectral peaks. This combined method of analysis would improve sensitivity and eliminate the more subjective baseline corrections required for the sample and reference spectra. Baseline corrections for quantitative analysis are especially difficult and subjective in those cases where there is scattering by the sample or where significant spectral overlap occurs. Incorrect baseline corrections are equivalent to a breakdown in Beer's Law. Therefore, accurate baseline corrections are of fundamental importance. This paper outlines the extension of our earlier least-squares methods to the case of multicomponent mixtures with overlapping spectral features and its application to artificial and real mixtures.

I. EXPERIMENTAL

A Nicolet 7199 Fourier transform infrared (FT-IR) spectrometer was used with a liquid nitrogen cooled Hg-Cd-Te detector with a range from 400 to 5000 cm^{-1}. Interferograms were collected to yield ~2 cm^{-1} resolution after transforming the data with triangular apodization. In order to compare results with those of Antoon *et al.*,[1] mixtures of the three xylene isomers (dimethylbenzenes) were used to evaluate the least-squares analyses. The xylenes used were chromatographic standards from Poly Science Corp. with isomeric purity of ≥99.5%. The reference spectra were obtained from the same 15-μm path length liquid cell to assure constant path length. Two hundred fifty-six interferograms were signal averaged in each case. The sample consisted of an accurately weighed (±0.0001 g) mixture of nearly equal weights (~0.15 g) of the three xylenes. Again the same liquid cell was used to eliminate path length variations. The spectrometer was well purged with dry N_2 gas to eliminate H_2O and CO_2 interferences.

The least-squares analysis was applied both to artificial and real xylene mixtures. The artificial mixtures were created by digitally adding known fractions of each xylene reference. Noise at various levels was then added to complete the artificial sample spectrum. The noise was generated by taking two separate single beam spectra of one scan with no sample in the IR beam. These were then ratioed to create a noise spectrum with the normal instrumental noise characteristics. Greater amounts of noise were created by multiplying this noise spectrum by constant factors. These artificial spectra assured that Beer's Law was followed over the entire spectral region since they were created from a linear combination of the reference spectra. The spectra of the real mixtures were used to determine the effects of possible non-Beer's Law behavior on the least-squares analysis. To determine the effectiveness of least-squares baseline corrections, the baselines of the samples were either left unchanged or corrected to zero with the interactive Nicolet software before applying the least-squares program to the data.

II. THEORY

Beer's Law is used as the basis for relating the concentration (c) of an absorbing species to its infrared absorbance (A) at each specified energy. That is,

$$A = abc \tag{1}$$

where a is the absorptivity and b is the path length. Beer's Law generally requires that the resolution of the spectrometer is sufficiently high to avoid significant instrumental broadening of the spectral bands being measured.[4] Anderson and Griffiths[5] also show that the instrumental line shape can often affect the measured peak absorption, and they recommend the application of Beer's Law only when $A \leq 0.7$. Eq. (1) also requires that each individual component in a mixture is not affected by the presence of other components (*i.e.*, noninteracting components). As has been shown previously[1] and confirmed in this study, the absorptions due to certain vibrations are often not influenced by the presence of component interactions that affect other vibrations of the same species. These unaffected vibrations will therefore follow Beer's Law.

Least-squares regression analyses are specifically developed here for samples of multicomponent mixtures with overlapping spectral features. Each analysis uses all the spectral data above a selected absorbance threshold in the spectral region of interest. The least-squares analysis includes a fit of the spectral baselines as well as a quantitative determination of each component of the sample. Since the sample spectrum is fitted to a least-squares linear combination of pure reference spectra, no assumptions about spectral line shapes are required. As before,[3] we have developed and tested four different least-squares fitting procedures. The differences between the four are in the assumptions made about the baselines of the spectra. The assumptions are: (I) the baseline is zero, i.e., the case developed previously by Antoon *et al.*[1] with no baseline fit; (II) the baseline is linear across the spectral region to be fit; (III) the baseline is linear over each spectral peak in the fit; and (IV) there is a negligible

baseline shift between successive data points in the fit. This last analysis is equivalent to a least-squares fit between first derivatives of the sample and reference spectra. In general, method I will require a correction of the baseline to zero absorbance before curve fitting since even reflectance losses will result in nonzero baselines. If the baseline assumptions are valid in methods II to IV for both the sample and reference spectra, then no preliminary baseline corrections are necessary for either the sample or reference spectra. However, with sloping baselines it may be desirable to baseline correct the reference spectra prior to curve fitting since this aids in the proper selection of peaks for the fitting program. In general, the baselines of pure reference spectra with high signal-to-noise ratios (S/N) are readily corrected to near zero by the software accompanying most commercial computerized infrared spectrometers.

In this study there are two points of departure in method I from that presented by Antoon *et al.*[1] The first is that only peaks in the reference spectra that are above a specified threshold value in the desired spectral region are selected for the least-squares analysis. Therefore, only those regions of the spectra where spectral information is present will be included in the fit. This eliminates data that add little or nothing to the analysis and thus reduces memory requirements and computation time. The selection of peaks is a formal requirement for method III since a linear baseline is fitted under each peak. The second deviation from the method of Antoon *et al.*[1] occurs in the weighting factors applied to each datum. The weighting factor should be related to the noise response expected for the particular type of detector used in the spectrometer. In particular, the weighting factor will be the reciprocal of the variance of the random noise. For most detectors used in infrared spectrometers (e.g., thermocouple, pyroelectric, or semiconductor), the noise is constant and independent of the signal level. Since the signal is converted to absorbance before being fitted, the variance of the noise in terms of absorbance must be determined. If the signal in absorbance is expanded as a Taylor series about the transmission and only the first two terms are kept, it is found that the variance of the noise is proportional to the inverse of the square of the transmittance (i.e., T^{-2}). Thus the weighting factor, z_i, used in these analyses, is calculated from the sample spectrum at each frequency i and is equal to T_i^2. Antoon *et al.*[1] used $(A_i)^{-1}$ as the weighting factor in their analyses. This weighting factor is more appropriate for absorption spectra using a photomultiplier detector where the noise is proportional to the square root of the absorbance. Its use is apparently a result of applying methods developed for gamma-ray spectra[6] where photomultipliers are used for detection.

The appropriate mathematical equations describing the four methods of least-squares analysis and the assumptions about the baselines are given in Table I. A_i^s represents the absorbance of the sample at frequency i while A_{ij}^r represents the absorbance of the jth reference at the same frequency. The parameters k_j are the ratios of the concentration of the jth component in the sample, c_j^s, to the concentration of the jth reference, c_j^r. (If path length variations are present, k_j represents the ratio of the path length concentration product (bc) for the sample and reference). The k_j terms are therefore the scaling parameters which can be used directly in spectral subtractions of the references from the samples. The equation for method I in Table I simply represents the linear combination of reference spectra which make up the composite sample spectra as expected from Beer's Law. The e_i term is the random error present at frequency i. The random error is assumed to be normally distributed with an expectation of zero and a variance proportional to T^{-2} as discussed earlier. A term that is linear in frequency (i.e., $a + b\nu_i$) has been added in method II in order to fit a linear baseline over the desired spectral region. Method III has the same linear baseline fit except that a separate linear baseline is fitted for each spectral peak. A different set of $\hat{k}_j$ is estimated from each peak and the final $\hat{k}_j$'s are determined by pooling the individual $\hat{k}_j$'s weighted inversely by the estimated variance calculated for each component in each peak. Method IV, which fits the difference between successive data points at constant frequency separation, is equivalent to a least-squares fit of the first derivatives of the reference and sample specta.

Table II presents the same equations in matrix form. The details of the computational methods for each fit are given in the Appendix. In particular, the least-squares

TABLE I. Least-squares methods for quantitative analysis of multicomponent mixtures by infrared spectroscopy.

Method	Fit[a]	Base line assumptions
I	$A_i^s = \sum_{j=1}^{m} k_j A_{ij}^r + e_i$	Baselines of sample and reference spectra are zero
II	$A_i^s = a + b\nu_i + \sum_{j=1}^{m} k_j A_{ij}^r + e_i$	Baselines of sample and reference spectra are linear over the spectral range fit
III	$A_{ip}^s = a_p + b_p\nu_{ip} + \sum_{j=1}^{m} k_j A_{ipj}^r + e_{ip}$	Baselines of sample and reference spectra are linear over each peak
IV	$\Delta A_i^s = \sum_{j=1}^{m} k_j \Delta A_{ij}^r + \Delta e_i$	Negligible baseline shift between successive data points for both sample and reference spectra

[a] Symbols defined as follows: A_i^s, sample absorbance at frequency i; A_{ij}^r, absorbance of jth reference at frequency i; k_j, ratio of concentrations of the jth component in the sample and the jth reference; ν_i, frequency; a, b, intercept and slope for linear baseline; e_i, noise at frequency i; p, subscript indicating values pertaining to a particular peak; ΔA_i^s, difference in sample absorbances at frequencies i and $i + 1$; ΔA_{ij}^r, difference in absorbances of the jth reference at frequencies i and $i + 1$.

TABLE II. Matrix representation of least-squares methods for quantitative analysis of multicomponent mixtures by infrared spectroscopy.

Least-squares fit	Matrix model[a]
I	$A_{(n\times 1)}^s = A_{(n\times m)}^r k_{(m\times 1)} + e_{(n\times 1)}$
II and III[b]	$A_{(n\times 1)}^s = U_{[n\times(m+2)]}\,\theta_{[(m+2)\times 1]} + e_{(n\times 1)}$ where $U = \begin{bmatrix} 1 & \nu_1 & \\ 1 & \nu_2 & \\ \vdots & \vdots & A^r \\ 1 & \nu_n & \end{bmatrix}$ and $\theta = \begin{bmatrix} a \\ b \\ k_{(m\times 1)} \end{bmatrix}$
IV	$\Delta A_{(n\times 1)}^s = \Delta A_{(n\times m)}^r k_{(m\times 1)} + \Delta e_{(n\times 1)}$

[a] Symbols defined as follows: $A_{(n\times 1)}^s$, vector of absorbance equally spaced frequencies; $A_{(n\times m)}^r$, matrix of absorbance values for the m references at n equally spaced frequencies; $k_{(m\times 1)}$, vector consisting of the ratio of concentrations of the m components in the sample and the concentration of the corresponding reference; ν, frequency; a, b, intercept and slope for linear baseline; $e_{(n\times 1)}$, vector of noise in spectrum at each frequency; Δ, indicates the matrix is composed of differences in absorbances between frequencies i and $i + 1$.

[b] Fit II uses one linear baseline over the entire spectral region, whereas Fit III is obtained by applying Fit II to each individual peak, then pooling the results over all peaks.

solution for method I in matrix form is

$$\hat{k} = (A_r'ZA_r)^{-1}(A_r'ZA_s) \quad (2)$$

where A_r and A_s are the matrices representing the reference and sample absorbances, the prime indicates the transpose of the matrix, and the Z matrix is the $n \times n$ diagonal matrix of weights (i.e., $z_{ii} = T_i^2 = 10^{-2A_i}$, where the T_i term is the transmittance of the sample at frequency i). If c_r is the vector matrix of known reference concentrations, then the concentration of each component in the sample is calculated by

$$\hat{c}_s = \hat{k}c_r \quad (3)$$

The details of calculation of the variance of $\hat{k}$, the error variance, σ^2, and the standard error of the estimated concentration, $SE(\hat{c}_j^s)$, are given in the Appendix. For large n, to a close approximation a 95% confidence interval on the true concentration in the sample is given by $\hat{c}_j^s \pm 1.96\ SE(\hat{c}_j^s)$.

III. PROGRAM

The methods presented above have been programmed in Basic language for the Nicolet 1180 computer, and the program is currently available with the standard Nicolet software package. The spectra in absorbance are stored on magnetic disk for the sample and each reference. The file name, concentration, spectral region of interest, and a threshold value for selecting peaks are input for each reference. After inputting the sample file name, any or all of the four fits may be selected. The starting and ending frequencies for all spectral bands above the absorbance threshold are determined for each reference. These are then compared between references, and new peaks are defined to include only the first and last frequencies of each set of overlapping peaks. All the corresponding absorbance data within peaks for each reference and sample spectrum are then stored in core memory and used in the fit. At each frequency i in these peaks, the absorbance data of each reference are included in the fit even if the absorbance of one or more references is below the selected threshold for that reference. This is necessary since large differences in component concentrations in the sample might result in an appreciable contribution from a low intensity peak in the component with a high relative concentration. Storage of the data completely in core memory reduces computation time by eliminating disk-to-memory transfers. However, it can also reduce the number of references or the spectral range that can be fitted. This limitation is not generally a problem since the proper selection of peaks greatly reduces the data that needs to be included in the fit.

The output of the program includes the number of peaks to be fitted and the number of data points involved. The least-squares estimates of the concentration of each component in the sample mixture are then listed for the selected fit. Ninety-five percent confidence intervals are listed along with the standard error of the estimated concentration. As written, the program can take up to three references, but straightforward changes in the dimension statements and one additional program statement are all that are required to increase the maximum number of references to be fitted. These changes are documented in the program listing.

IV. RESULTS AND DISCUSSION

The infrared spectra of the three pure xylene isomers and a mixture of nearly equal weights of the three xylenes are presented in Fig. 1. It is clear from these spectra that conventional methods of quantitative infrared spectroscopy which require isolated infrared bands for analysis must rely on the out-of-plane C—H bending vibrations (i.e., 600 to 850 cm^{-1}) since these are the only bands which have the requisite lack of spectral overlap in the mixture. As pointed out by Antoon *et al.*[1] and demonstrated later in this paper, these bands exhibit the greatest deviation from Beer's Law and, therefore, quantitative accuracy will suffer. Of course a large series of known mixtures could be run to determine calibration curves in this spectral region, but this is a time-consuming process which will only be applicable over narrow concentration ranges.

In order to evaluate the least-squares analysis in the absence of deviations in Beer's Law, the analyses were performed on the artificially generated spectra discussed earlier. Three artificial spectra with varying quantities of noise are presented in Fig. 2. The S/N ratios are defined as the peak signal to the average peak-to-peak noise. Spectrum A represents the spectrum that would be obtained with no signal averaging (i.e., with one scan each of the sample and background). This is the lowest S/N ratio to be expected for this type of sample and the S/N is degraded by a factor of 16 over that actually obtained with the 256 scan signal averaging used in this study. The S/N ratio has been artificially degraded by successive factors of 10 in Fig. 2, B and C.

The results of the least-squares analyses are presented

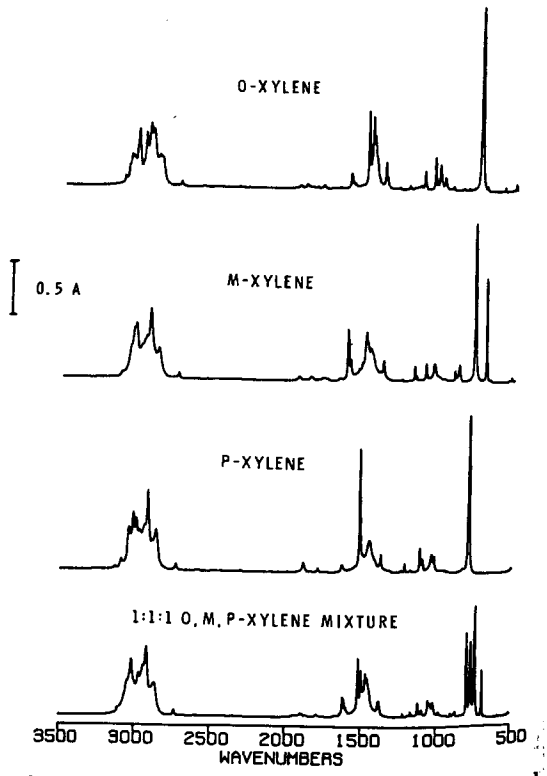

FIG. 1. Spectra of each of the three pure xylene isomers and accurately weighed 1:1:1 mixture of three xylenes.

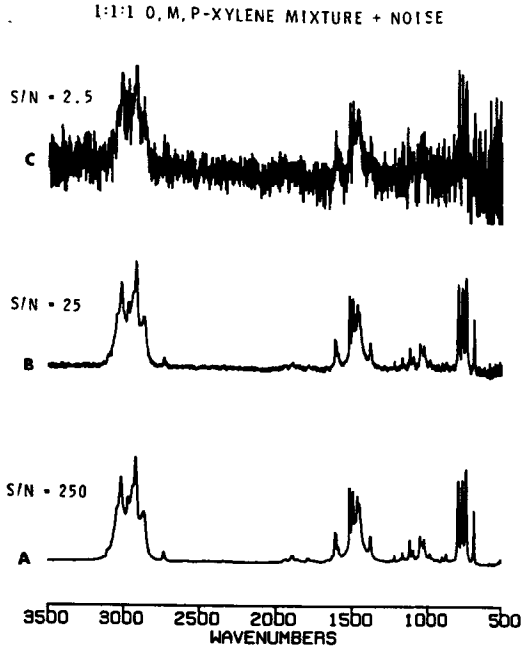

FIG. 2. Artificially synthesized 1:1:1 mixture of the three xylene isomers with noise added. A, S/N = 250; B, S/N = 25; C, S/N = 2.5.

in Table III. No corrections for nonzero baselines have been made for either the reference or sample spectra used in obtaining these data. Table III gives the percent relative error for each of the xylenes using least-squares methods II (linear baseline under the entire spectral range) and IV (the derivative fit). No attempt has been made to force the total percentage to 100%. The results for method I will be dependent on the position of the baseline, whereas the results for fit III show relative errors approximately 20% greater than those of fit II. This greater imprecision is due to the lower precision with which the individual peaks are fitted relative to the single fit of method II. The derivative fit generally yields greater average relative errors and yields lower precision than fit II since method IV fits differences between two successive data points which each contain random error. However, it should be remembered that the accuracy of fits III and IV may be greater in those cases where the more strict linear baseline assumption of fit II fails.

The results of Table III show that the average relative error is approximately 0.1% for a fit by method II when S/N = 250. Even when the S/N ratio is only 2.5, quantitative results with an average relative error of <15% are obtained with fit II. Thus, the least-squares analysis using all the spectral information above a threshold of 0.15 *A* (i.e., 15 infrared bands included in the fit) yields accurate quantitative results for the three xylenes when Beer's Law is known to be valid over the entire spectral region. An effort was also made to analyze concentrations where the signal was less than the noise. This analysis yielded estimated concentrations that were not significantly different from zero at the 5% statistical level; that is, the 95% confidence interval on the actual concentration included zero. Therefore, the least-squares analyses applied to this mixture of xylenes, which exhibits significant spectral overlap, does not allow concentrations to be reliably detected when the signal is below the noise level. Detection when S/N < 1 was possible when these least-squares methods were applied to gas phase spectra of low molecular weight molecules with no spectral interferences.[3] The lower number of peaks and broader nature of the IR bands for the xylene isomers also contributes to this reduced sensitivity from the high-resolution gas phase spectra. However, these methods are invariably superior to conventional techniques which do not use the least-squares regression analyses over wide spectral regions.

The least-squares analyses were subsequently applied to the real mixture of nearly equal molar volumes of the three xylenes. Antoon *et al.*[1] have pointed out that the fingerprint region (450 to 1400 cm^{-1}) is more likely to experience deviations in Beer's Law than the C—H stretching region. With the spectral addition capabilities of computerized infrared spectrometers, adherence to Beer's Law can be readily checked for real mixtures. If an accurately weighed mixture of the components is prepared and its infrared spectrum obtained, then this spectrum can be compared with one which is artificially generated by adding the single component reference spectra in the same molar concentration as prepared in the real mixture. This artificially generated spectrum will correspond to that of an ideal mixture for which Beer's Law is exactly followed. A spectral subtraction of the artificial spectrum from that of the real mixture will then indicate which regions of the spectrum do not follow Beer's Law. An example of this procedure is shown in Fig. 3 for the mixture of the three xylenes in nearly equal mole fractions. The residual spectrum in Fig. 3 obtained after subtraction of the appropriate amount of each pure xylene reference spectra has been scale expanded by a factor of 2 and clearly indicates those regions where the Beer's Law assumption breaks down. These regions are primarily the in-plane and out-of-plane C—H bending vibrations on the aromatic ring. This failure of Beer's Law is in part due to the strong and narrow nature of these bands as well as the greater molecular perturbation of the bending vibrations of the aryl hydrogens. Avoidance of these vibrations in the least-squares analysis should, therefore, yield more accurate results. This is confirmed by the results of the least-squares analysis applied to the real mixture of xylenes as demonstrated in Table IV for a variety of spectral ranges. These results were obtained from spectra that were not corrected for nonzero baselines.

TABLE III. Percent relative error for the least squares analysis applied to an artificially generated spectrum of 1:1:1 *o*, *m*, *p*-xylenes.[a]

S/N of sample spectrum	% Relative error[b]		
	o-Xylene II (IV)	*m*-Xylene II (IV)	*p*-Xylene II (IV)
250:1	−0.02 (0.20)	0.08 (0.28)	−0.21 (−0.02)
25:1	−0.21 (1.9)	0.75 (2.7)	−2.2 (−0.27)
2.5:1	−10.3 (13.1)	5.4 (17.1)	−27.5 (−4.8)

[a] Samples generated by adding 0.3333 of each reference, plus appropriate noise. Spectral region fit was 550 to 3100 cm^{-1} with an absorbance threshold of 0.15 A. Fit included 15 peaks with a total of 818 data points. No preliminary baseline correction of either sample or reference spectra were made.

[b] Values without parentheses are relative errors obtained by applying Fit II. Values in parentheses are relative errors obtained by applying Fit IV.

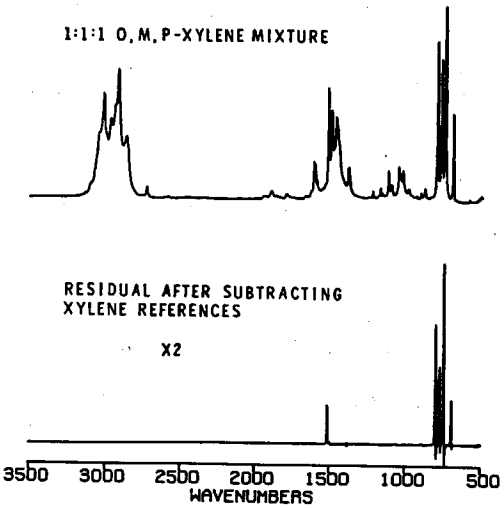

FIG. 3. Top, Accurately weighed 1:1:1 mixture of the three xylene isomers. Bottom, Residual spectrum obtained by subtracting the pure xylene reference spectra according to their mole fractions in the mixture from the mixture spectrum. Scale is expanded by a factor of 2 over that of the upper spectrum.

Table IV presents both the estimated concentrations, with their precision at the 95% level for fit III, and the percent relative error from the known concentrations. The data in this table illustrate that the fitting is less accurate in those spectral regions (600 to 900 cm^{-1} and 1250 to 1750 cm^{-1}) where deviations from Beer's Law are expected from the results of Fig. 3. The C—H stretching region (2800 to 3100 cm^{-1}) and the fingerprint region between 850 and 1250 cm^{-1} both give more accurate results as expected from Fig. 3. The combination band regions 1750 to 1950 cm^{-1} and 2000 to 2700 cm^{-1} also yield quite accurate results even though the spectral peaks in these regions are factors of ~10 and ~50 less intense, respectively, than found with the C—H stretching vibrations. This illustrates both the high S/N of FT-IR and the value of simultaneous least-squares fitting of the baselines. Examination of the results of Antoon *et al.*[1] shows that their analysis of the 1750 to 1950 cm^{-1} region with subjective baseline corrections exhibited the greatest absolute error of the spectral ranges they studied. The peak intensities in this overtone region are quite low, and therefore, the effects of poor baseline corrections are greatly magnified and the quality of the results severely affected. In addition, it is found that the average error obtained by Antoon *et al.*[1] using the C—H stretching region and the zero baseline assumption was a factor of 2 greater than found here when the baselines are fitted by least-squares methods. A subjective effort to baseline correct the spectra in this study also resulted in greater errors than achieved by fitting linear baselines. Therefore, combining a fit of the baseline with the least-squares regression analysis of the concentrations eliminates the subjective baseline corrections, is faster, and yields more accurate results.

It is interesting to note the high accuracy of fit III (see Table IV) when the entire spectral range is included in the fit. This is in spite of the fact that several intense peaks which do not adhere to Beer's Law are included. The high accuracy of this particular fit is a result of calculating a set of concentrations for each peak and pooling the results over all peaks according to the inverse of the variance determined for each component in each peak. Naturally, those regions which experience Beer's Law deviations have significantly greater variances and are given low weighting. The pooled result is therefore strongly biased to those peaks that follow Beer's Law. Thus, fit III allows the experimenter more leeway in selecting spectral regions for the fit. However, since greater computation time is required for the larger spectral regions, if many samples are to be analyzed it may be preferable to determine in advance those spectral regions which follow Beer's Law.

The errors in fit I are of course related to the extent of nonzero baselines. A shift in baseline values between the sample and reference spectra of 0.08 *A* (i.e., the shift encountered in reflection from the cell windows) results in relative errors of 50%. Even using the same liquid cell for each reference and sample with approximately equal baseline shifts in each case results in calculated errors 1.5 to 10 times those listed in Table IV (i.e., 2% to 8% average relative error). The error associated with fit II is comparable to that of fit III when narrow spectral ranges are examined. However, the weighted pooling of the results in fit III as discussed above and the more valid assumption of linear baselines across each peak contribute to the result that the error with fit III is one-tenth that of fit II when peaks from the entire spectral range (550 to 3100 cm^{-1}) are used in the fit. The average relative error for fit IV is ~1.5 times that of fit III except again when the entire spectral region is fit. In this latter case, the average relative error for the three xylenes is ~17% for fit IV.

In addition to the necessary validity of the baseline assumptions and the requirements that Beer's Law apply, it is implicitly assumed that all components of the sample mixture are known and that their pure reference spectra are used in the fit. This latter restriction is not required if those components that are not included in the fit contribute negligibly to the infrared bands used in the analysis. It has been shown above that as a consequence of the weighted pooling of results for each peak in fit III, the Beer's Law assumption does not have to be valid over the entire spectral region to yield accurate results. The requirement of including all components in the fit

TABLE IV. Results obtained from applying least-squares fit III to various spectral regions of the real 1:1:1 mixture of xylene isomers.[a]

Spectral region fit (cm^{-1})	Absorbance threshold (A)	No. of peaks	Calculated mole fractions (% relative error)		
			o-Xylene 0.3395[b]	*m*-Xylene 0.3313[b]	*p*-Xylene 0.3292[b]
2800–3100	0.2	1	0.3401 ± 0.0009 (0.18)	0.3259 ± 0.0013 (−1.6)	0.3294 ± 0.0010 (0.06)
2000–2700	0.083	5	0.3284 ± 0.0043 (−3.3)	0.3349 ± 0.0048 (1.1)	0.3370 ± 0.0038 (2.4)
1750–1950	0.1	5	0.3412 ± 0.0027 (0.50)	0.3439 ± 0.0044 (3.8)	0.3278 ± 0.0016 (−0.43)
1250–1750	0.2	3	0.3305 ± 0.0052 (−2.7)	0.3341 ± 0.0020 (0.85)	0.3704 ± 0.0052 (12.5)
850–1250	0.15	7	0.3368 ± 0.0029 (−0.80)	0.3410 ± 0.0036 (2.9)	0.3447 ± 0.0040 (4.7)
600–900	0.2	3	0.3810 ± 0.0350 (12.2)	0.3960 ± 0.0219 (19.5)	0.3786 ± 0.0379 (15.0)
550–3100	0.2	14	0.3398 ± 0.0009 (0.09)	0.3285 ± 0.0011 (−0.85)	0.3311 ± 0.0010 (0.58)

[a] No preliminary baseline corrections of sample or reference spectra were made.
[b] Actual mole fraction in the mixture of xylene isomers.

can also be relaxed for the same reason and by virtue of the fact that at least a portion of the spectral overlap from unknown components can be placed in the linear baseline for each peak.

To test the ability of obtaining accurate fits in the absence of complete knowledge of the components present in the mixture, fits of the three-component xylene mixture were attempted with only one or two of the components as references. The percent relative error is presented in Table V after applying fit III to the real 1:1:1 mixture of the three xylene isomers using only one or two of the three xylenes as references. The data in Table V demonstrate that the ortho and meta xylenes are quite accurately determined even though one or two of the major components in the spectra are missing. The relative errors for *p*-xylene range from 17% to 45%. At least a portion of this error is due to non-Beer's Law behavior since the fit of the artificially constructed spectrum (Fig. 2A) results in relative errors ranging from only 4% to 11% when components are eliminated from the fit. It is important to note that the errors presented in Table V are sensitively dependent on the threshold absorbance selected since this value determines the infrared bands to be included in the fit. Surprisingly, the derivative fit yields results with relative errors ranging from 11% to 26%. These results, which are only slightly dependent on the threshold absorbance selected, would suggest that the least-squares derivative fit (fit IV) is also not strongly dependent on the inclusion of all components in the fit. When one or more of the sample components are missing from the fit, both fits III and IV yield results which are reasonably accurate although they do not, in general, yield the high quantitative accuracy possible when all components are fit. This is in contrast to the large errors found with fit I applied in the same manner. With fit I the relative errors range from 40% to 160%. Thus the greater sophistication of fits III and IV greatly expands the utility of the least-squares routines by relaxing the requirement that all components of a sample be known. In fact, a least-squares fit of several components in a mixture will yield fairly accurate subtraction factors for use in spectral stripping. After subtracting the appropriate amounts of the references, the remaining major spectral components can be more easily identified.

Another severe test of these least-squares techniques involves the linear independence of the reference spectra. If reference spectra are used in the fit of a sample which does not contain the references as components, then their composition should be zero within the precision of the calculation if the reference and sample spectra are linearly independent. A simple test of this independence is obtained by selecting two of the xylene isomers as references in an attempt to fit a sample composed only of the third. Antoon *et al.*[1] performed this test and calculated mole fractions of the two xylenes which summed to 0.9 to 1.1 when only the third xylene was present in the sample. These large nonzero values clearly demonstrate the need to account for every component accurately in the sample when the baseline is assumed to be zero. We find similarly high errors with our fit I (i.e., total fractional composition of the two xylenes of 0.48 to 0.56 when the spectral region 550 to 3100 cm^{-1} was fit). However, fits III and IV yield quite accurate results as demonstrated in Table VI. Although the calculated concentrations are often nonzero by slightly more than the 95% precision limits, they are quite small, yielding a maximum mole fraction of 0.023. It must be remembered that the isomeric purity of the samples is ~99.5%, which would allow for xylene mole fractions of 0.005 to be present in each sample. The derivative fit IV yields slightly more accurate results suggesting a greater linear independence in the derivative spectra than in the direct infrared spectra. The quality of these results adds further confidence in the application of these least-squares fitting routines to unknown samples that may not contain all the expected components.

V. SUMMARY

The development of new least-squares procedures for fitting reference and sample infrared spectra has greatly improved the available methods of quantitative determination of the components in samples of multicomponent mixtures with overlapping spectral peaks. Four least-squares fitting procedures were described and tested in this report. The four differed in assumptions made about the spectral baselines. One method was based on assuming a zero baseline, whereas two of the methods provided for least-squares fitting of the spectral baselines. The fourth method fitted the first derivatives of the reference spectrum to that of the sample spectrum. The least-squares fitting of spectral baselines yields greater accuracy and eliminates subjective baseline corrections. Fitting of baselines is especially valuable when spectra of low absorbance are fit since in these cases baseline errors are magnified and can result in large errors in the estimated concentrations. Greater linear independence of the reference spectra is also achieved by fitting the baselines.

TABLE V. Percent relative error obtained by applying least-squares fit III to a real 1:1:1 mixture of xylene isomers if one or two of the components are not included in the fit.[a]

Reference spectra used in fit III	% relative error		
	o-Xylene	*m*-Xylene	*p*-Xylene
o, *m*-Xylene	0.9	0.7	...
o, *p*-Xylene	−6.9	...	17
m, *p*-Xylene	...	11	45
o-Xylene	3.3	...	...
m-Xylene	...	4.0	...
p-Xylene	...	...	45

[a] Spectral region fit was 550 to 3100 cm^{-1} with an absorbance threshold of 0.2 A. No preliminary baseline corrections of sample or reference spectra were made.

TABLE VI. Amount of xylenes calculated by applying least-squares fits III and IV to pure xylene samples using the other two xylene isomers as references.[a]

Reference spectra	Sample spectrum	Fit	Mole fraction of xylene calculated[b]		
m, *p*	*o*	III	...	−0.0062 ± 0.0050	0.0234 ± 0.0126
		IV	...	−0.0059 ± 0.0066	0.0022 ± 0.0052
o, *m*	*p*	III	−0.0082 ± 0.0035	−0.0077 ± 0.0006	...
		IV	0.0083 ± 0.0038	−0.0029 ± 0.0054	...
o, *p*	*m*	III	−0.0045 ± 0.0085	...	0.0182 ± 0.0060
		IV	−0.0033 ± 0.0033	...	0.0125 ± 0.0039

[a] Spectral region fit was 550 to 3100 cm^{-1} with an absorbance threshold of 0.2 A. No preliminary baseline corrections of sample or reference spectra were made.
[b] The limits given are for 95% confidence intervals on the mole fractions.

A simple method for determining which regions of the sample spectra follow Beer's Law has been demonstrated. Thus those regions which adhere to Beer's Law can be selected for inclusion in the fitting procedure. However, it was also shown that accurate quantitative results are possible by simultaneously fitting linear baselines under each peak (fit III) since the method heavily weights those regions where Beer's Law is followed. The power of these new methods is further illustrated and emphasized by the demonstration that semiquantitative results are possible even when all major components of the sample are not known or when expected components are not present. These latter findings greatly expand the applicability of the quantitative least-squares fitting methods. Finally, the methods developed here have been applied to infrared spectra. However, they are not limited to the infrared but can be applied to any quantitative spectral analysis where there is a known relationship between the sample concentration and spectral intensity.

ACKNOWLEDGMENTS

The authors wish to acknowledge G. E. Rivord for writing a portion of the least-squares program. P. J. Rodacy performed the GC purity analysis of the three xylene isomers.

1. M. K. Antoon, J. H. Koenig, and J. L. Koenig, Appl. Spectrosc. 31, 518 (1977).
2. P. C. Painter, S. M. Rimmer, R. W. Snyder, and A. Davis, Appl. Spectrosc. 35, 102 (1981).
3. D. M. Haaland and R. G. Easterling, Appl. Spectrosc. 34, 539 (1980).
4. D. A. Ramsey, J. Am. Chem. Soc. 74, 72 (1952).
5. R. J. Anderson and P. R. Griffiths, Anal. Chem. 47, 2339 (1975).
6. J. A. Blackburn, Anal. Chem. 37, 1000 (1965).
7. D. W. Marquardt and R. D. Snee, Am. Stat. 29, 3 (1975).

Appendix

Analysis Details

A. Weighted Least-Squares Regression. Estimating the concentrations of the individual components in a multicomponent mixture, by the methods described in Table I, was done by a weighted least-squares regression analysis. This appendix describes the basic methodology and the variations required for each of the four methods considered.

In general a linear regression model can be written as

$$Y = X\beta + e,$$

where Y is a $n \times 1$ vector of observations; X is a $n \times s$ matrix of known constants, sometimes called predictors or explanatory variables; β is a $s \times 1$ vector of unknown constants; and e is an $n \times 1$ vector of errors or noise. If the noise vector is modeled as a random observation from a probability distribution that has a mean of zero and a variance of $\sigma^2 V$, where V is a known $n \times n$ matrix then the weighted least-squares estimate of β is

$$\hat{\beta} = (X'V^{-1}X)^{-1}(X'V^{-1}Y).$$

The $s \times s$ covariance matrix of $\hat{\beta}$ is

$$C(\hat{\beta}) = \sigma^2(X'V^{-1}X)^{-1}$$

and the unknown σ^2 is estimated by

$$\hat{\sigma}^2 = \frac{Y'Y - \hat{\beta}'(X'V^{-1}Y)}{n - s}$$

Thus the diagonal elements of $C(\hat{\beta})$, with σ^2 replaced by $\hat{\sigma}^2$, yield the estimated variances of each estimated coefficient and the square root of these estimated variances is the standard error of the estimate.

Weighted regression analysis is simplest when the matrix V is diagonal. This situation corresponds to an assumption that the errors are independent but with possibly different variances. That assumption is appropriate here. The diagonal elements, however, are not known but can be estimated by T_i^{-2}, as described in section II of this paper. Using estimated weights means that the above standard errors are not correct because they are based on the assumption of fixed weights. The effect, however, should be small because of the wide range in the values of T_i^{-2} and because of the threshold used in specifying the data used in the least-squares analyses.

The analyses for methods I to IV all follow this general approach, but are modified to reflect the different baseline assumptions considered.

B. Fit I: Zero Baseline. In this case the vector Y of observations is the sample spectra A_s, and the columns of X are the reference spectra considered. The vector β is the vector of relative concentrations, k_j. In order to best achieve computational accuracy before inverting $X'V^{-1}X$ using our Nicolet 1180 computer program, this matrix is scaled so that the diagonal elements equal 1.0. It should be noted that $X'V^{-1}X$ and $X'V^{-1}Y$ can be constructed by reading one observation at a time, forming all the weighted squares and products (among the sample and reference spectra for that frequency), and accumulating these results. Thus, it is not necessary to store X and V^{-1} which are large matrices. In order to complete the fit, there must be at least $r + 1$ data points available where r is the number of references used in the program.

C. Fit II: Linear Baseline. This case differs from method I in that the X matrix of reference spectra is augmented by a column of ones, representing the intercept, α, of the linear baseline, and a column of frequencies for the spectral region being considered. For the sake of computational accuracy the matrix $X'V^{-1}X$ is reduced, in the program, to "correlation form." This means in essence that the columns of X have been centered and scaled so that the weighted sum of each column is zero and the weighted sum of squares of each is one (see Ref. 7). Fit II requires the presence of at least $r + 3$ data points.

D. Fit III: Linear Baseline across Each Peak. In this case method II is applied to the spectral data from each peak. This results in an estimated relative concentration, $\hat{k}_{pj}$, for reference j in peak p, a matrix $(X'V^{-1}X)_p^{-1}$ which is used in estimating the variance of $\hat{k}_{pj}$, and an estimate of σ^2, say $\hat{\sigma}_p^2$, based on peak p's data. Let s_p^{jj} be the diagonal element of $(X'V^{-1}X)_p^{-1}$ corresponding to $\hat{k}_{pj}$. The method used to estimate the relative concentration of component j in the mixture is to take a weighted average of the $\hat{k}_{pj}$'s, where the weight is the reciprocal of s_p^{jj}. To obtain a standard error of the resulting $\hat{k}_j$ requires an estimate of σ^2, which is assumed to be constant across peaks. This estimate is the following weighted average of

the $\hat{\sigma}_p^2$'s:

$$\hat{\sigma}^2 = \frac{\sum_p (n_p - r_p - 2)\hat{\sigma}_p^2}{\sum_p (n_p - r_p - 2)}$$

where n_p is the number of observations for peak p and r_p is the number of references included in that peak. The weights, $n_p - r_p - 2$, are the number of degrees of freedom associated with $\hat{\sigma}_p^2$. The standard error of $\hat{k}_j$ is then the square root of

$$\hat{var}(\hat{k}_j) = \frac{\hat{\sigma}^2}{\sum_p (1/s_p^{jj})}.$$

This fit requires at least $r_p + 3$ data points to be present in each peak. The program automatically discards any peak with fewer data points.

E. Fit IV: Negligible Baseline Shift. In this case the matrix Y is a vector of successive differences between the sample absorbances, X is a matrix of successive differences for the r references, and β is the vector of relative concentrations. Thus, initially, this model resembles that of Fit I, but there are some complicating factors that arise because the matrix V is not diagonal. Successive differences are correlated. The correlations are easily obtained but to apply the analysis described above would require inverting V which is, in general, a very large matrix. Thus, an alternative analysis was chosen.

The variance of a difference between successive independent observations is the sum of the variances. As described in section II of this paper, the variance of an absorbance measurement is approximately proportional to T_i^{-2}. Thus, the variance of a successive difference is $T_i^{-2} + T_{i+1}^{-2}$. Let D be a diagonal matrix with these variances as diagonal elements. To obtain estimates of the relative concentrations, D was substituted for V. That is, we take

$$\hat{\beta} = (X'D^{-1}X)^{-1}(X'D^{-1}Y),$$

where X and Y contain differences, as described, to obtain the estimated relative concentrations. The variance of $\hat{\beta}$ is given by

$$\begin{aligned} var(\beta) &= \sigma^2(X'D^{-1}X)^{-1}(X'D^{-1}VD^{-1}X)(X'D^{-1}X)^{-1} \\ &= \sigma^2 S^{-1}CS^{-1} \end{aligned}$$

The matrix C is $r \times r$ and can be constructed without constructing and storing the large X, D, and V matrices. The variance σ^2 can be estimated by

$$\hat{\sigma}^2 = \frac{Y'D^{-1}Y - \hat{\beta}'(X'V^{-1}Y)}{n_\Delta - c_\Delta},$$

where n_Δ is the number of differences included in the fit and c_Δ is trace of the matrix $S^{-1}C$. (This analysis is a multivariate extension of that given in Appendix A, section 3 of Ref. 3.) Thus, the above covariance matrix of $\hat{\beta}$ can be estimated, and the square roots of the diagonal elements provide standard errors for the estimated relative concentrations of the references. For simplicity of programming, each peak must contain five or more data points. Peaks with fewer data are excluded from the fitting procedure by the program software.