Iterative Least-Squares Fit Procedures for the Identification of Organic Vapor Mixtures by Fourier Transform Infrared Spectrophotometry

Xiao Hong-kui,[1,2] Steven P. Levine,*[,1] and James B. D'Arcy[1,3]

School of Public Health, The University of Michigan, Ann Arbor, Michigan 48109-2029, School of Public Health, Beijing Medical University, Beijing 100083, Peoples' Republic of China, and Biomedical Science Department, Research Laboratories, General Motors Corporation, Warren, Michigan 48090

Least-squares fitting (LSF) was applied to the qualitative analysis of IR spectra based on comparing standard reference spectra with the sample mixture spectrum. Identification of compounds in the sample was made by judging the fit level of the spectrum of each compound with the sample spectrum. An iterative procedure was developed to eliminate compounds with the worst fit levels in order to approach an optimal fit for the sample spectrum. The qualitative analysis results obtained from the optimal fit were further used for quantitative analysis.

INTRODUCTION

Recently, much attention has been focused on computerized processing of infrared (IR) spectral information for qualitative and quantitative analysis of mixtures. A number of programs have been developed to assist the chemist in the identification and quantitation of unknown compounds. Many approaches to automated spectral interpretation have been developed such as pattern recognition, factor analysis, hierarchical clustering, and expert systems (*1–18*). Summaries of recent literature in this and related subjects are available in symposium proceedings (*19*, *20*).

Most of these systems make decisions principally based on information of peak position, uniqueness, and intensity for structural elucidation of compounds in the condensed phase. However, the IR spectra of compounds present in the vapor phase at trace concentrations are different from those in the condensed phase. This is due to a lack of effects from changes in hydrogen bonding, the dielectric constant of the medium, and nonpolar solvent–solute interactions.

Many of the systems for interpretation of condensed-phase IR spectra cannot be readily applied to gas or vapor systems. A successful application of an expert system to the interpretation of the IR spectra of pure compounds in the vapor phase was accomplished with PAIRS (*14*). A computerized interpretation system, MIXIR, has also been developed, which shows promise for application to the interpretation of the IR spectra of vapors (*21*). This system is in the early stages of development of testing on such spectra.

In the workplace, the concentration of most toxic organic compounds may be below one-tenth the occupational exposure limit, frequently expressed as the threshold limit value (TLV) or the permissible exposure limit (PEL) (*22*). When the concentrations of organic vapors are as low as 1 ppm, and the composition of the sample is complex, spectral peaks may be overlapped. In this circumstance, it is extremely difficult to interpret IR spectra by using computerized techniques that depend on picking peaks. This is due to low peak intensities, "chemical noise", and electronic noise (*23–27*).

* Author to whom correspondence should be addressed.
[1] The University of Michigan.
[2] Beijing Medical University.
[3] General Motors Corp.

In this context, "chemical noise" may be considered to consist of components of the vapor-phase mixture that are outside the reference library, and of little or no interest to the analyst. For example, in the workplace, this may include low concentrations of low molecular weight hydrocarbon gases, along with oxides of sulfur and nitrogen. Because of the problems associated with noise, overlap of spectral features, and low peak intensities, no computerized software is currently available for the qualitative interpretation of the IR spectra of mixtures of gases and vapors in low concentration.

In this study, least-squares fitting (LSF), sometimes called classical least squares (CLS), was applied to the qualitative analysis of IR spectra based on comparing standard reference spectra with the sample mixture spectrum. Identification of compounds in the samples was made by judging the fit level of the spectrum of each compound with the sample spectrum. An iterative procedure was developed, and manually applied, to eliminate compounds not meeting predetermined worst-fit criteria. The qualitative analysis results thereby obtained from the optimal fit were further used for quantitative analysis using LSF.

For traditional quantitative applications, where the identities of the components of the mixture are known, LSF approaches have been used with good success (*23–27*). As expected, Warner and co-workers found that the results of an analysis using LSF can be strongly affected by not including all components in samples. They reported the results obtained when using nonnegative least squares and linear programming for unknown samples (*28*). Haaland and Easterling approached the problem of quantitation of components in a mixture by selecting spectral regions that yield the best fit for reference spectra and then combining the results from all spectral regions in a statistically efficient manner. These methods were first reported in a seminal paper by Haaland (*29*) and then in later publications (*30*, *31*). Later, they applied partial-least-squares (PLS) methods for quantitative IR spectroscopy (*32*).

In this paper we present the results obtained when using a commercially available version of the LSF program developed by Haaland (*29–31*) for the qualitative and quantitative IR spectral analysis of trace organic vapors in impurity- and moisture-free air. Interpretation of data is presented that yields optimal approaches to the analysis of samples of unknown composition and concentration.

EXPERIMENTAL SECTION

Samples. Three air mixture samples containing five to 11 compounds in zero air were provided by T. Pritchett, U.S. Environmental Protection Agency (EPA). The concentration of compounds was verified by EPA using gas chromatography (GC). Forty-seven reference standard vapor IV spectra were acquired as a reference library with the use of GC-certified gas cylinders (Scott Specialty Gas Co.).

Hardware. A Nicolet 20 SXB Fourier transform infrared spectrometer was used with a liquid nitrogen cooled HgCdTe–InSb sandwich detector and 10-m Hanst gas cell. Interferograms were

0003-2700/89/0361-2708$01.50/0

collected to yield spectra of from 0.5 to 8.0 cm^{-1} resolution through deresolution of the 0.5-cm^{-1} interferograms by the method described in previous papers (*23–27*).

Software. For the detailed theory of LSF software, see papers by Haaland et al. (*29–31*). Detailed discussions of the application of LSF software to quantitative analysis of trace gases in ambient air have also been published (*23–27*).

METHODS

The least-squares fitting method III, which assumes a linear base line over the width of the peak, was adopted for this study. With method III, more than one peak, and therefore more than one base line, can be present within a window. From ref 29, the best fit in a least-squares sense is obtained by minimizing the expression $\sum_i[A_{si} - kA_{ri}]^2$, where A_{si} is the sample absorbance at frequency i, A_{ri} is the reference absorbance at frequency i, and k is the ratio of sample and reference concentrations.

The error variance, σ^2, is estimated from

$$\sigma^2 = \frac{\sum_i(A_{si} - \hat{k}A_{ri})^2}{n - 1}$$

where n is the total number of observations.

For large n (typically $n > 400$) a 95% confidence interval on the true concentration in the sample is $\hat{C}_s \pm 1.96[\mathrm{SE}(\hat{C}_s)]$, where

$$\mathrm{SE}(C_s) = \frac{\hat{\sigma}C_r}{(\sum A_{ri}^2)^{1/2}}$$

and C_s is the least-squares estimate of the sample concentration, C_r is the known concentration of the reference, and $\mathrm{SE}(C_s)$ is the standard error of the estimated concentration (*29*).

The reference spectra in a selected region were fitted with the sample spectra. The lack of spectral fit was expressed in terms of an estimated error variance. If the concentrations of the reference spectra are known, concentration errors could be expressed as the standard error of the estimated concentration, or as the 95% confidence interval of the estimated concentration.

The infrared spectral interpretation procedure described herein is designed to report the presence of each compound chosen from a reference library of compounds on the basis of a comparison of the reported concentration and 1.96 times the standard error, which was used to calculate the 95% confidence interval.

Two methods of data analysis were evaluated in this study. The first method attempted to eliminate compounds from the library of compounds by analyzing the spectra in one group of 47 compounds, or in groups of 16. The second method attempted to build a set of possible compounds by fitting the reference spectra one at a time to the sample spectrum.

Method I (Set Reduction Method (SRM)). First, in order to provide a point of comparison for all iterative procedures, the Nicolet LSF software was altered to allow the use of a reference set of up to 48 compounds, rather than the 16 compounds that are normally the maximum allowed by the software. This method is referred to as SRM-47, since 47 compounds were used in this study. After the first iteration, all positives are kept for the second iteration. At this point, this method becomes the same as the SRM detailed immediately below.

Secondly, in order to institute the iterative LSF set reduction approach using the commercially available version of the Nicolet LSF software, the library had to be divided into subsets of 16 compounds each from the library of 47 compounds. To accomplish this, the reference standard spectra were divided into three categories: aliphatic hydrocarbons, aromatic compounds, and oxygenated compounds. The IR window for each compound was then selected. The window can be either a narrow window chosen specifically for that compound or a "general" window covering the fingerprint and/or the C–H stretch region(s). If sufficiently wide windows are chosen, the results may, in some circumstances, be approximated by the use of a "general" window, since the overlapping windows merge into a single region for LSF analysis.

Sample spectra were then analyzed by the LSF program, which yielded the concentration and a value of 1.96 times the standard error, which was used to calculate the 95% confidence interval for each compound. Judgment was made by the manual comparison of each compound's concentration and 95% confidence interval. *Those compounds whose reported 95% confidence interval did not include a concentration of zero were kept* and combined to form a new data set for the next LSF analysis. The remaining compounds were eliminated.

This procedure was repeated until no further compounds were eliminated, and the standard error for each analyte stayed essentially constant. This took between three and four iterations. All compounds remaining were then reported as positive identifications.

Method II (Set Building Method (SBM)). A specific IR window was selected for each compound. Compounds were then entered, one reference standard spectrum at a time, into the LSF program file for analyzing the sample spectrum. This procedure was repeated with each compound in the library (a 47-compound library in this example). Compounds were reported as not present according to the same criteria as in method I. In the case in which only one compound would be reported as present, analysis of the data would have stopped at this point.

In cases in which more than one compound was reported as being present when the SBM was used, then the compounds selected formed a set of reference spectra for the ILSF analysis of the sample using the SRM. Compounds were then eliminated by iteration using method I.

Both of the ILSF methods were performed manually, with data from each cycle of the LSF software reentered by hand.

Quantitative Analysis. For both the SRM and the SBM, the iterative LSF (ILSF) analysis yielded results that are considered to be quantitative when the final iteration was completed. Data obtained by using the ILSF methods were compared to that obtained when the spectra of known compounds were entered as reference spectra into the LSF program file for quantitative analysis.

RESULTS AND DISCUSSION

For each step of the investigation, results are shown only for an illustrative fraction of the complete data set. This is done to conserve space, since a full reporting of the data would not add to the information content of the results or the generalizability of the conclusions. The full data set is available upon request.

Assumptions in the Use of ILSF Methods. When one is using ILSF methods, it is important to understand the assumptions implicit in the use of the estimated concentration and its relationship to the 95% confidence interval as a critical value on which to base decisions of the presence or absence of compounds in the spectrum of a mixture (*33*). (The relationship of the standard error to the 95% confidence interval is explained in the Methods section.) These assumptions are (1) spectral errors are normally distributed for the sample spectra; (2) there are no errors or noise in either the reference spectra or the concentrations of the reference spectra; (3) there are no model errors (Beer's law is followed, the method III linear base-line assumption across each peak is correct, and temperature, pressure, and pressure broadening must be

Table I. Summary of Qualitative Results Obtained for Samples I–III by Using the Set Reduction Methods (SRM) and the Set Building Method (SBM)

	sample I iteration				sample II iteration				sample III iteration				mean of
	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th	final iteratn[b]
						SRM-General Windows[a]							
true pos	6	6	6	*c*	5	5	5	4	8	7	7	*c*	
false pos	8	1	1		12	3	1	1	10	3	3		
true neg	33	40	40		30	39	41	41	26	33	33		
false neg	0	0	0		0	0	0	1	3	4	4		
sensitivity[d]	100	100	100		100	100	100	80	73	64	64		81
specificity[e]	80	98	98		71	91	98	98	72	92	92		96
						SRM-47							
true pos	5	5	5	*c*	5	5	5	5	10	10	10	10	
false pos	11	6	5		21	11	6	2	22	6	4	3	
true neg	30	35	36		21	31	36	40	14	30	32	33	
false neg	1	1	1		0	0	0	0	1	1	1	1	
sensitivity	83	83	83		100	100	100	100	91	91	91	91	91
specificity	73	85	88		50	74	86	95	61	83	89	92	92
						SRM-Specific Windows							
true pos	6	6	6	*c*	5	5	5	5	8	8	8	8	
false pos	8	1	1		16	4	2	2	9	7	3	3	
true neg	33	40	40		26	38	40	40	27	29	32	32	
false neg	0	0	0		0	0	0	0	3	3	3	3	
sensitivity	100	100	100		100	100	100	100	73	73	73	73	91
specificity	80	98	98		62	90	95	95	75	80	89	89	94
						SBM-Specific Windows							
true pos	6	6	6	6	5	5	5	5	9	9	9	9	
false pos	15	7	0	0	11	3	1	1	15	5	2	2	
true neg	26	34	41	41	31	39	41	41	21	31	34	34	
false neg	0	0	0	0	0	0	0	0	2	2	2	2	
sensitivity	100	100	100	100	100	100	100	100	82	82	82	82	94
specificity	63	83	100	100	74	93	98	98	58	86	94	94	97

[a] 650–1350 cm^{-1} for fingerprint, 2900–3200 cm^{-1} for C–H stretch region. [b] Mean values were based on the results of the final iteration. [c] When two iterations yield identical results, no further iterations were performed. [d] The percentage of compounds actually present in the mixtures and identified as such. [e] The percentage of compounds actually not present in the mixtures and identified as such.

constant); (4) if analysis is performed by using specific frequency windows for each analyte, and these windows, in the case of overlap, are pooled by LSF, spectral error variance is constant in all bands pooled; (5) there are no components of the mixture not initially included in the set of reference spectra. Thus, for actual samples of mixtures, the confidence interval will be underestimated (*33*).

Protocol Designation. Four separate protocols were performed for each of the three samples using the library of 47 spectra. The tests included (1) ILSF set reduction method starting with up to 48 compounds in a single iteration and with the use of broad, or "general", frequency windows (SRM-47), (2) ILSF set reduction method starting with three sets of 16 compounds and with the use of broad, or "general", frequency windows (SRM-G), (3) ILSF set reduction method starting with three sets of 16 compounds and with windows optimized, or "specific", for each component (SRM-S), and (4) ILSF set building method with windows optimized, or "specific", for each component (SBM-S).

In method SRM-47 and SRM-G, broad windows were used, encompassing both 2900–3200 cm^{-1} for the C–H stretch region and 650–1350 cm^{-1} for the fingerprint region. In the other two methods, all windows were in the fingerprint region between 650 and 1350 cm^{-1}, except for hexane and cyclopentane, for which peaks in the C–H stretch region (2900–3200 cm^{-1}) were used.

Sensitivity and Specificity. In this text and in Tables I and II, nomenclature is used in which the terms "sensitivity" and "specificity" are defined in a manner consistent with use in the health sciences and with use in a prior publication in this journal (*16, 17*). "Sensitivity" is a measure of true positive results and is defined as the percentage of compounds actually present in the mixture and identified as such. For example, in Table I, for the results obtained for sample III by using the SRM-general method

$$\text{sensitivity} = \frac{\text{true positives (7)}}{[\text{true positives (7)} + \text{false negatives (4)}]} = 64\%$$

Similarly, "specificity" is a measure of true negative results and is defined as the percentage of compounds actually not present in the mixture and identified as such. In the same example as above

$$\text{specificity} = \frac{\text{true negatives (33)}}{[\text{true negatives (33)} + \text{false positives (3)}]} = 92\%$$

Comparison of Results. The data show that, for the six-component mixture in sample I, results were approximately equivalent for the SRM-general and SRM-specific tests, but improved to the point that there were neither false positives nor false negatives with the SBM-specific approach. The SRM-47 method produced results that had lower sensitivity and specificity than that of the other methods.

For the five-component mixture in sample II, the SBM-specific results were equivalent in specificity to the SRM-general results, but better than those obtained with the SRM-specific and SRM-47 approaches. For sensitivity, the SBM, SRM-47, and SRM-specific methods were equivalent and were better than the SRM-general results.

For the 11-component mixture in sample III, the SBM-specific was better than either of the SRM results in both sensitivity and specificity, except for the high sensitivity achieved with the SRM-47 method.

Table II. Qualitative Analysis of Three Samples Using the Set Building Method with Specific Windows (SBM-S)

		sample I iteration				sample II iteration				sample III iteration			
	window, cm^{-1}	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
acetaldehyde	1029–1189	TN[a]				FP[b]	TN			FP	TN		
butyl acetate	990–1330	FP	TN			TN				FP	TN		
2-ethoxyethyl acetate	1011–1310	TN				TN				FP	TN		
ethyl acetate	1000–1145	FP	TN			TN				TP	TP	TP	TP
n-butyl alcohol	880–1150	TN				TN				TN			
ethoxyethanol	970–1310	FP	TN			TN				FP	TN		
1-propanol	924–1155	FP	FP	TN		TN				FP	TN		
2-propanol	1031–1200	FP	TN			TN				TP	TP	TP	TP
ethyl ether	976–1240	TN				TN				TP	TP	TP	TP
bis(chloroethyl) ether	1077–1273	TN				TN				FP	TN		
ethylene oxide	764– 954	TN				TN				TN			
1,4-dioxane	832–1177	TN				TN				TN			
tetrahydrofuran	840–1144	TP[c]	TP	TP	TP	FP	FP	TN		FP	FP	FP	FP
propylene oxide	700–1197	TN				TN				FP	FP	TN	
acetone	1160–1260	TN				TN				TN			
methyl ethyl ketone	872–1241	TN				TN				TN			
methyl propyl ketone	866–1257	TN				TN				TN			
methyl amyl ketone	1087–1255	TN				TN				TN			
methyl isobutyl ketone	896–1312	TN				TN				TN			
cyclopentane	2802–3052	FP	TN			TN				TP	TP	TP	TP
hexane	2799–3025	FP	TN			FP	FP	FP		FP	TN		
benzene	620–1075	TP	TP	TP	TP	FP	FP	TN		TN			
ethylbenzene	1152–1291	TP	TP	TP	TP	TN				TP	TP	TP	TP
styrene	878–1032	FP	TN			FP	FP	FP	FP	FN[d]			
toluene	996–1143	FP	FP	TN		TP	TP	TP	TP	FP	FP	FP	FP
m-xylene	958–1165	FP	FP	TN		TN				TN			
o-xylene	840–1190	TN				TN				TN			
p-xylene	961–1161	TN				FP	TN			TN			
carbon tetrachloride	712– 841	TN				TN				TP	TP	TP	TP
chloroform	1188–1250	TN				TN				TN			
methylene chloride	1219–1309	TP	TP	TP	TP	TN				TN			
vinyl chloride	819–1066	TN				TP	TP	TP	TP	TN			
3-chloropropene	850–1126	TN				FP	TN			FN			
1,1-dichloroethane	670–1324	TP	TP	TP	TP	TN				TP	TP	TP	TP
1,2-dichloroethane	1180–1340	FP	FP	TN		TN				FP	FP	TN	
1,1-dichloroethylene	1040–1173	TN				FP	TN			FP	FP	TN	
1,1,1-trichloroethane	981–1138	TP	TP	TP	TP	FP	TN			FP	TN		
1,1,2-trichloroethane	866– 983	FP	TN			FP	TN			TP	TP	TP	TP
trichloroethylene	754– 973	TN				TP	TP	TP	TP	FP	TN		
tetrachloroethylene	734– 942	FP	FP	TN		TP	TP	TP	TP	TN			
chlorobenzene	994–1148	FP	FP	TN		TP	TP	TP	TP	TN			
o-chlorotoluene	710–1161	FP	FP	TN		TN				TN			
Freon-11	809–1066	TN				FP	TN			TP	TP	TP	TP
acetonitrile	862–1120	TN				FP	TN			TN			
acrylonitrile	862–1046	TN				TN				FP	TN		
dimethyl disulfide	917–1347	TN				TN				TN			
pyridine	958–1245	TN				TN				TN			
sensitivity[e]		100	100	100	100	100	100	100	100	82	82	82	82
specificity		63	83	100	100	74	93	98	98	58	86	94	94

[a] TN: true negative. [b] FP: false positive. [c] TP: true positive. [d] FN: false negative. [e] See Table I.

In sample III, cyclopentane is present. Both cyclopentane and *n*-hexane are present in the library of reference spectra used in this study. Cyclopentane was detected by the SBM on the first iteration, as was *n*-hexane. By the second iteration, *n*-hexane was eliminated. (This is illustrated in Tables II and III.) This is because, on the first iteration using the SBM, the hexane spectrum is matched with the sample spectrum. Since the sample spectrum contained cyclopentane, the peaks of the sample spectrum in the C–H stretch region contained features similar to that of the spectrum of hexane. At this point, since the LSF file of the sample did not yet contain cyclopentane as a known constituent, the false positive result was given for hexane.

However, in the second iteration, cyclopentane had been added to the reference file set with hexane, and the LSF result showed that the peaks in the C–H stretching region were attributed to cyclopentane. Therefore, the remaining spectral features did not fit that of hexane. At this point, the 95% standard error for the fit of the cyclopentane was reduced from 0.29 to 0.12, thereby indicating an improved fit. This performance was paralleled by that of the two SRM tests in which *n*-hexane was rejected in the second iteration when general windows were used and in the third iteration when specific windows were used. (These data are not shown, but are available upon request.)

Note that in these methods, when the results of two successive iterations are identical, no further iterations are performed. This does not mean that the results are more accurate when the convergence rate is higher. While this may be the case, it has not been systematically investigated.

In most cases, the SRM-specific results were better than those obtained by using the SRM-general method. This is because the use of a wide frequency window for all compounds in many cases includes "no information" regions that will, at the concentrations studied here, contain significant noise and/or cause the base lines to be improperly drawn for peaks used for the identification and quantitation of target analytes. The linear base-line assumption will be more accurate over

Table III. Results of Analyses of Samples I–III Using the Set Building Method with Specific Windows

	1st iteration		2nd iteration		3rd iteration		4th iteration	
	concn[a]	SE[b]	concn	SE	concn	SE	concn	SE
			Sample I					
tetrahydrofuran	5.96	0.28	3.17	0.27	1.20	0.32	2.25	0.38
1,1-dichloroethane	4.28	0.65	3.01	0.41	2.34	0.40	2.45	0.40
benzene	4.47	0.92	6.32	0.42	6.49	0.35	6.31	0.32
ethylbenzene	5.70	0.27	2.63	0.54	2.89	0.11	2.65	0.07
methylene chloride	2.02	0.18	2.23	0.30	2.11	0.31	2.33	0.23
1,1,1-trichloroethane	3.52	1.96	2.57	0.10	2.79	0.26	2.68	0.12
			Sample II					
vinyl chloride	11.03	2.37	2.46	0.09	2.44	0.08	2.52	0.08
trichloroethylene	2.07	0.54	2.82	0.36	2.72	0.20	2.71	0.02
toluene	13.06	2.92	2.88	0.26	2.29	0.17	2.27	0.14
chlorobenzene	2.36	0.11	2.61	0.08	2.46	0.06	2.43	0.06
tetrachloroethylene	2.27	0.27	2.45	0.02	2.39	0.02	2.39	0.02
styrene (FP)[c]	6.91	2.48	2.29	0.07	0.17	0.07	0.19	0.07
			Sample III[f]					
Freon-11[d]	12.12	0.25	9.03	2.65	4.19	0.06		
cyclopentane	2.98	0.29	1.46	0.12	1.30	0.08		
1,1,2-trichloroethane	5.62	0.74	5.04	2.34	1.44	0.19		
ethylbenzene	11.35	0.62	4.57	0.83	5.67	0.97		
2-propanol	6.25	2.63	4.87	0.59	1.98	0.50		
styrene (FN)[e]	0.67	0.88						
3-chloropropene (FN)	−2.14	3.03						
carbon tetrachloride	1.34	0.23	1.59	0.31	1.66	0.05		
ethyl ether	1.86	0.39	3.30	0.29	2.76	0.15		
1,1-dichloroethane	1.19	1.12	2.05	1.07	0.98	0.32		
ethyl acetate	1.54	0.85	2.20	1.14	1.34	0.06		
tetrahydrofuran (FP)	4.20	2.23	8.92	2.19	0.42	0.27		
toluene (FP)	31.12	14.90	4.35	0.31	0.80	0.49		

[a] Concentration (ppm (v/v)) in zero air. [b] 1.96 × standard error. Calculate the confidence interval as shown in the Methods section. [c] False positive (indicated concentration > standard error). [d] The concentration of Freon-11 was not certified with a certified gas cylinder standard. [e] False negative (concentration is negative, or < standard error). [f] No improvement in results obtained after third iteration.

a narrower spectral range. However, in cases in which the frequency window is chosen poorly, results obtained with the specific window methods may degrade. (This will be discussed below.)

The data shown in Table I also demonstrates that the SBM was almost always better than or equivalent to the SRM methods in sensitivity and specificity. This is illustrated by reference to the mean value of the sensitivity and specificity, obtained from the final iteration for each method.

In some cases, the results obtained for specific analytes vary with each method. For example, as can be seen from Tables II and III, styrene is reported as a false negative in sample III by all methods except for the SRM-47 method. The SRM-47 method correctly identifies styrene, as well as 2-propanol and ethyl ether, but then Freon-11 is missed. These changes in results are, undoubtedly in large part, due to the interaction between changes in spectral windows, base-line assumptions, and the ILSF method chosen. These factors are the subject of continuing investigations.

The operation of the SBM is further illustrated in Table II, in which each of the iterations for each of the samples is shown. For example, for sample I, styrene is reported as a positive in the first iteration, but is correctly identified as a true negative by the second iteration. In contrast to that, for sample III, styrene is incorrrectly identified as a negative in the first iteration. Succeeding iterations do not contain styrene in the sample set.

As an illustration of the complexity of the spectra of the mixtures used for this investigation, Figure 1 shows the spectra of the mixture and its components, as well as of 2-ethoxyethyl acetate, which is the false positive result for sample I obtained by using SRM-G and SRM-S.

Confidence Interval and Number of Iterations. The 95% confidence interval of the concentration results obtained from LSF is the critical value in the ILSF qualitative analysis. Based on the comparison between the concentration and its confidence interval, expressed in the tables in terms of 1.96 times the standard error, the lack of fit of standard reference spectra to the sample spectrum was calculated. With this information, a judgment was made of whether or not each compound existed in the sample.

As shown in Table III, by reference to the results obtained for sample III, the value of the standard error strongly affects the positive or negative classification of each target analyte. When the SRM-specific method is used (results not shown), ethyl ether is classified as being a false negative. This is because the 95% confidence interval includes zero, and the reported result is therefore "not present" for ethyl ether. If a value of the confidence interval of less than 95% had been used, ethyl ether would have been accurately reported as present. However, as shown in Table II, ethyl ether is reported as a true positive when one uses the SBM.

Similarly, for sample III in Table II, the styrene is reported as "not present". Once again, had the standard error been set at, for example, 90% rather than 95%, the confidence interval would not have included zero, and the compound would have been reported as present.

However, in that situation, the false positive results might have increased as well. Therefore, these is always the risk of losing sensitivity while improving specificity.

Spectral Regions. In the mixtures used for this study, it is rare to find a compound that has spectral features that are completely resolved from those of other compounds. However, in the gas phase under conditions of ambient temperature and pressure, the peaks neither shift nor change shape. Therefore, the selection of optimal frequency windows becomes important.

Frequency regions chosen for each compound for use with the SBM are given in Table II. However, this may be somewhat misleading since the LSF program will combine

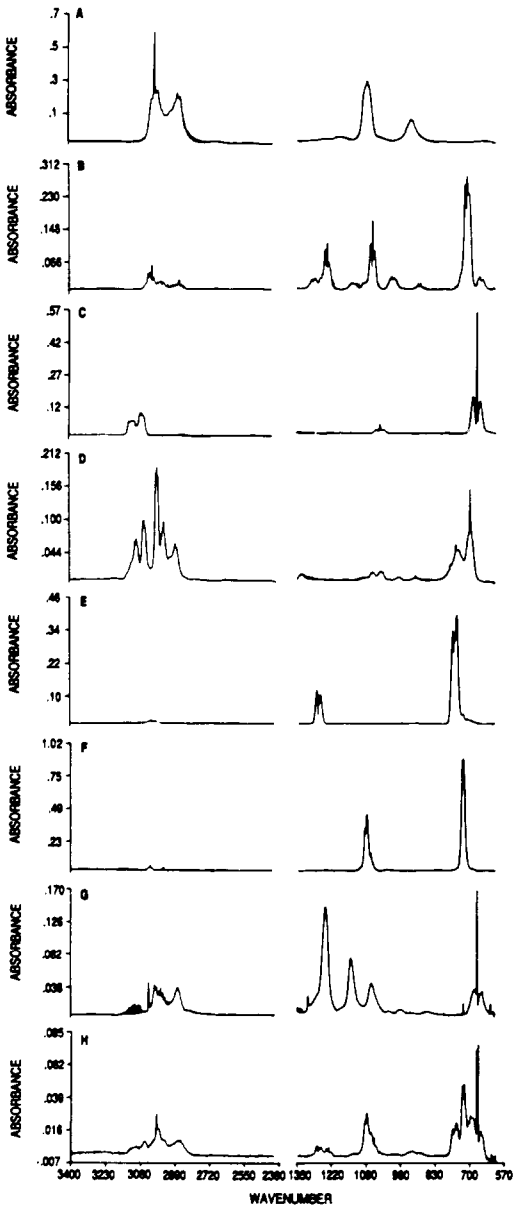

Figure 1. Portions of the infrared spectra of standard reference compounds and sample I (600–1350 cm^{-1}, 2400–3300 cm^{-1}). All the standard reference spectra selected are based on the results of iterative least-squares set reduction method-specific. Traces: (A) 100 ppm tetrahydrofuran, (B) 50 ppm 1,1-dichloroethane, (C) 49.1 ppm benzene, (D) 51.2 ppm ethylbenzene, (E) 49.1 ppm methylene chloride, (F) 49.6 ppm 1,1,1-trichloroethane, (G) 10.6 ppm 2-ethoxyethyl acetate, (H) sample mixture I.

overlapping windows in each iteration. The spectral region thus used will vary for each mixture depending on the degree of overlap of the individual windows.

Selection of appropriate IR frequency windows for LSF was discussed at length by previous authors (*23–27*). However, the most optimal IR window for each compound selected for one mixture may not be appropriate for other mixtures (*26*). As further evidence of this, sample I was analyzed by using a set of reference spectra consisting only of the compounds known to be present in the sample. No iteration was performed.

When sample I was analyzed by LSF using either specific windows within the fingerprint region or the entire fingerprint region, both benzene and ethyl benzene were reported at concentrations significantly greater than those reported by GC analysis (Tables IV and V).

These results were improved significantly when alternate windows in the C–H stretching region were used for these two

Table IV. Comparison of Results Obtained for the Analysis of Sample I Using FTIR and GC Methods

	concentration,[a] ppm					
	FTIR					
	SRM-G	SRM-S	SBM-S	SRM-47[b]	LSF	GC[c]
tetrahydrofuran	2.92	2.32	2.30	FN[d]	2.30	2.3
1,1-dichloroethane	1.51	2.33	2.57	2.00	2.57	3.46
benzene	4.04	2.34	2.27	1.14	2.27	2.30
ethylbenzene	4.35	2.60	2.65	2.87	2.65	2.11
methylene chloride	1.60	2.38	2.44	2.34	2.44	2.40
1,1,1-trichloroethane	0.30	2.70	2.70	2.89	2.70	2.47

[a] Data for SRM-G, SRM-S, and SBM-S were obtained from the final iteration of each method. Reference spectra for the LSF column were based on the results obtained by GC analysis. [b] The sensitivity of SRM-47 for all samples was 91% (Table I). This example should not be interpreted to imply that the use of SRM-47 results in a greater incidence of FN results. [c] GC results are not certified. These results are shown for comparison purposes only. [d] False negative. Compound was identified by the GC method, but not by this FTIR method, and was included in the reference set for LSF analysis.

Table V. Effect of Windows on the Results Obtained by FTIR for Sample I

	window, cm^{-1}, and concentration, ppm				GC
compound	window I	concn	window II	concn	results
tetrahydrofuran	840–1144	2.22	840–1144	2.30	2.3
1,1-dichloroethane	670–1324	1.71	670–1324	2.57	3.46
benzene	*620–1075*	6.11	*2650–3200*	2.27	2.30
ethylbenzene	*1152–1291*	3.76	*2650–3200*	2.65	2.11
methylene chloride	1219–1309	2.48	1219–1309	2.44	2.40
1,1,1-trichloroethane	981–1138	2.71	981–1138	2.70	2.47

analytes. This is shown in Table V. Illustration is also given in Figure 1, in which the spectra are shown in both the fingerprint and the C–H stretch regions for sample I and for the components of the sample.

Results may also have been improved if windows had been used for these analytes in both the fingerprint and C–H regions simultaneously. However, this was not tested.

For the 2-propanol in sample III, which has strong C–H stretching peaks, after its IR window was changed from the fingerprint region to the C–H region, the results were unchanged (actual data not shown). The results given for benzene, ethyl benzene, and 2-propanol indicate that, for those compounds which have strong peaks in the C–H stretching region and weak peaks in the fingerprint region, the results may be improved by selecting an IR window for LSF analysis in the C–H stretch region. This, in some cases, would also avoid the complicated overlapped peak envelopes in the fingerprint region of the spectra of certain mixtures.

Noniterative Procedure. Table IV summarizes the quantitative aspects of the LSF-based methods used for the analysis of these samples. Data is shown only for sample I. The data column labeled "LSF" shows the results of analysis where only those compounds known to be present from the GC results were included from the beginning in the LSF data set, and no iteration was performed.

These data indicate that results obtained by using the LSF, SRM-47, SRM-specific, and SBM-specific methods are in good

agreement with the results of GC analysis of the samples. The results obtained by using the SRM-general method are not as accurate. Note that no information is available certifying the absolute accuracy of the GC results, so these data are given for comparison purposes only. Therefore, it is clear that, for the compounds and mixtures, at the concentrations evaluated in this study, LSF-based methods are usable for the quantitation of complex mixtures of unknown composition of traces of organic vapor in zero air.

Resolution. Previous studies have shown that, for analysis of trace concentrations of vapors in ambient air, 2-cm^{-1} resolution was adequate (*23–27*). It is important that a resolution be chosen that can both provide the required information and not require the use of excessive computer time and/or higher performance instruments. This is because the ultimate application of this research is to real-time and/or portable Fourier transform infrared (FTIR) systems for occupational health use (*22*).

The effect of resolution on quantitative aspects of the analysis of this data set was studied for all three sample mixtures. There was no demonstrable difference in accuracy when the resolution was varied from 0.5 to 4 cm^{-1}. Thus, 2-cm^{-1} resolution was used in this study.

Analysis Time. Consideration of time of data analysis is important. For these samples, the average computation time, including printing of results, was 4 h for the SRM-47 method, but only 1 h for each of the other methods. Actual computational time for the SBM is less than for the SRM because a higher fraction of the total time is taken by the SBM for printing of intermediate results. Since this will not be a factor in the fully automated software, it is estimated that the SBM will be faster than the SRM by approximately a factor of 2.

It is recognized that the Nicolet 1280 computer is not ideal for this type of analysis. Indeed, a DEC VAX system is generally recommended for such applications. However, we are automating this approach with a 20-MHz PC-DOS 80386-387 system. With this system, we expect computational time to be decreased by a factor of almost 2 orders of magnitude when compared to that required by the Nicolet system.

CONCLUSIONS

Iterative least-squares fit (ILSF) methods based on a commercially available version of the LSF program written by Haaland were developed for qualitative and quantitative analysis of complex mixtures of vapors at trace concentrations. These programs were evaluated for three mixtures containing five to 11 components at approximately 2 ppm in zero air. The results indicate that the ILSF method is very promising as a tool for the interpretation of IR spectra of mixtures.

Generally, performance of the ILSF methods, as measured by "sensitivity" and "specificity", were in the following order: set building method-specific, better than set reduction method-47, approximately equal to set reduction method-specific, better than set reduction method-general. Both qualitative and quantitative results were improved by using specific frequency windows for each compound and by using the set building method rather than the set reduction method.

The standard error after LSF for each compound played an important role in qualitative analysis. The "sensitivity" and "specificity" obtained by each method for each component will be, in part, a function of the set point chosen for the percentage value of the standard error, and thereby for the size of the confidence interval. In general, for a given method, specificity can only be improved at the expense of sensitivity (and vice versa). Investigation is under way into the idea of a programmable confidence interval set point for each iteration.

Work is continuing in this study. Efforts are centered on fully automating the ILSF procedures, on testing the effects of false positive and false negative components on accuracy, on specifying the presence of certain compounds and compound classes, on investigating approaches to base-line problems, and on testing these procedures on sample mixtures in ambient air.

ACKNOWLEDGMENT

The authors thank Ying Li-shi, whose invaluable contributions to the SRM and use of LSF for this application are central to the success of this work. The efforts of Tom Pritchett, U.S. EPA-ERT, are greatly appreciated for supplying the three sample mixtures used in this study. In addition, we thank Dan Sparks and Bill Herget at Nicolet Analytical Instrument Company for valuable guidance, and Mary Weed for artwork.

LITERATURE CITED

(1) Antoon, M. K.; D'esposito, L.; Koenig, J. L. *Appl. Spectrosc.* **1979**, *33*, 351.
(2) Varmuza, K. *Anal. Chim. Acta* **1980**, *122*, 227.
(3) Frankel, D. S. *Anal. Chem.* **1984**, *56*, 1011.
(4) Brown, C. W.; Lynch, P. F.; Obremski, R. J.; Lavery, D. S. *Anal. Chem.* **1982**, *54*, 1472.
(5) Moldoveanu, S.; Rapson, C. A. *Anal. Chem.* **1987**, *59*, 1207–1212.
(6) Fuller, M. P.; Ritter, G. L.; Draper, C. S. *Appl. Spectrosc.* **1988**, *42*, 217.
(7) Zupan, J. *Anal. Chim. Acta* **1978**, *103*, 273–288.
(8) Zupan, J.; Munk, M. E. *Anal. Chem.* **1983**, *55*, 2137–2142.
(9) Zupan, J.; Munk, M. E. *Anal. Chem.* **1985**, *57*, 1609–1616.
(10) Zupan, J.; Munk, M. E. *Anal. Chem.* **1986**, *58*, 3219–3225.
(11) Woodruff, H. B.; Smith, G. M. *Anal. Chem.* **1980**, *52*, 2321.
(12) Woodruff, H. B.; Smith, G. M. *Anal. Chim. Acta* **1981**, *133*, 545–553.
(13) Tomellini, S. A.; Saperstein, D. D.; Stevenson, J. M.; Smith, G. M.; Woodruff, H. B. *Anal. Chem.* **1981**, *53*, 2367–2369.
(14) Tomellini, S. A.; Stevenson, J. M.; Woodruff, H. B. *Anal. Chem.* **1984**, *56*, 67–70.
(15) Tomellini, S. A.; Hartwick, R. A.; Stevenson, J. M.; Woodruff, H. B. *Anal. Chim. Acta* **1984**, *162*, 227–240.
(16) Puskar, M. A.; Levine, S. P.; Lowry, S. R. *Anal. Chem.* **1986**, *58*, 1156–1162.
(17) Puskar, M. A.; Levine, S. P.; Lowry, S. R. *Anal. Chem.* **1986**, *58*, 1981–1989.
(18) Ying, L. S.; Levine, S. P.; Tomellini, S. A.; Lowry, S. R. *Anal. Chem.* **1987**, *59*, 2197.
(19) Eighth International Conference on Computers in Chemical Research and Education, June 15–20, 1987, Beijing, Peoples Republic of China; *Anal. Chim. Acta* (Special Issue) **1988**, *210*, 1–224.
(20) Second Hidden Peak Symposium on Computer-Enhanced Analytical Spectroscopy, June 1–3, 1988, Snowbird, UT. Proceedings to be published by Plenum Press.
(21) Tomellini, S.; Wythoff, B. *Anal. Chim. Acta*, in press.
(22) Air Contaminants—Permissible Exposure Limits. 29 Code of Federal Regulations, Part 1910.1000, U.S. Dept. of Labor (OSHA) Document 3112, 1989.
(23) Strang, C. R.; Levine, S. P.; Herget, W. F. *Am. Ind. Hyg. Assoc. J.* **1989**, *50*, 70–77.
(24) Strang, C. R.; Levine, S. P. *Am. Ind. Hyg. Assoc. J.* **1989**, *50*, 78–83.
(25) Ying, L.-S.; Levine, S. P.; Strang, C. R. *Am. Ind. Hyg. Assoc. J.* **1989**, *50*, 354–359.
(26) Ying, L.-S.; Levine, S. P. *Am. Ind. Hyg. Assoc. J.* **1989**, *50*, 360–365.
(27) Ying, L.-S.; Levine, S. P. *Anal. Chem.* **1989**, *61*, 677–683.
(28) Warner, I. M.; Davidson, E. R.; Christian, G. D. *Anal. Chem.* **1977**, *49*, 2155–2159.
(29) Haaland, D. M.; Easterling, R. G. *Appl. Spectrosc.* **1980**, *34*, 539–548.
(30) Haaland, D. M.; Easterling, R. G. *Appl. Spectrosc.* **1982**, *36*, 665–673.
(31) Haaland, D. M., Easterling, R. G.; Vopicka, D. A. *Appl. Spectrosc.* **1985**, *39*, 73.
(32) Haaland, D. M.; Thomas, E. V. *Anal. Chem.* **1988**, *60*, 1193.
(33) Haaland, D. M., Sandia Laboratories, Albuquerque, NM, personal communication, March 1989.

RECEIVED for review March 27, 1989. Revised manuscript received August 14, 1989. Accepted September 29, 1989. The authors thank the U.S. EPA Environmental Response Team (EPA-ERT) (Research Contract 68-03-3255 and related contracts) and the Centers for Disease Control (CDC-NIOSH) (Research Grant 1-R01-02404) for their generous support.