

## **General Disclaimer**

### **One or more of the Following Statements may affect this Document**

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.
- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.
- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.
- This document is paginated as submitted by the original source.
- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

Least Squares Methods of  
Analyzing Spectroscopic Data\*



J. H. Shaw

The Ohio State University  
Columbus, Ohio 43210



DRA

\*This work was supported in part by

NASA Grant NSG-7469

(NASA-CR-173241) LEAST SQUARES METHODS OF  
ANALYZING SPECTROSCOPIC DATA (Ohio State  
Univ., Columbus.) 23 p HC A02/MP A01

N84-16858

CSSL 12A

Unclas

G3/64

00604

## **Abstract**

**An overview of the methods of line-by-line and whole-band analysis developed by The Ohio State University spectroscopy group is presented.**

# ORIGINAL PAPER OF POOR QUALITY

## Introduction

One of the objectives of our group is the collection and analysis of laboratory spectra of molecules of atmospheric interest. In recent years we have developed efficient techniques for extracting the maximum amount of information from these spectra with a minimum of observer bias. A description of the approach to the analysis of single lines and entire bands, taken by the members of our group listed in Table 1, is described here. More detailed descriptions are given elsewhere (see references).

## Single Line Analysis

A typical problem in spectral analysis is the determination of line positions. One method is to estimate the line center and to determine its position with respect to other calibration features by interpolation, as indicated in Fig. 1. This method is often adequate but it is tedious if there are many lines to be measured. It is also difficult to estimate the precision of the results since the judgments used to determine the line center and in the interpolation process cannot be quantified.

Less reliance is placed on the observer if the experimental signals  $I_{\text{exp } i}(\nu)$  are compared with the corresponding signals  $I_{\text{art } i}(\nu)$  of an artificial line centered at  $\nu_0$  (art) as shown in Figure 2a. The spectrum of the differences

$$\Delta(\nu) = I_{\text{exp } i}(\nu) - I_{\text{art } i}(\nu),$$

shown in Fig. 2b, indicates the sensitivity of the data to changes in the line position. This sensitivity can also be tested by calculating the partial derivative  $\partial I_{\text{exp } i}(\nu) / \partial \nu_0$ . A best estimate of the line position  $\nu_0$  is obtained by shifting the position  $\nu_0$  (art) until the sum of the squares of the differences

$$SS = \sum \Delta I_i^2(v)$$

ORIGINAL PAGE IS  
OF POOR QUALITY

is a minimum as indicated in Fig. 2c. The rate of change of the SS with respect to the difference  $\Delta v = v_0(\text{exp}) - v_0(\text{art})$  gives a measure of the precision with which the line center can be found. Unlike the first method, once the model for the artificial line and the criteria for establishing the uncertainty of the line position have been defined, no other observer judgment is required.

Other parameters can also be determined by this method. By generating artificial lines  $I(v, S_{i \text{ art}})$  as shown in Fig. 2d, corresponding to lines of different intensities, and minimizing

$$SS = \sum \Delta I_i^2(v, S_{i \text{ art}}),$$

where

$$I(v, S_{i \text{ art}}) = I_{\text{exp } i}(v) - I(v, S_{i \text{ art}}),$$

estimates of the line intensity and its precision can be obtained. In this method

- (1) The intensity and position estimates, and also estimates of other variables  $p_j$  which affect the signals  $I_i(v)$ , are obtained from the same set of experimental data.
- (2) Estimates of the goodness of fit to  $I_{\text{exp } i}(v)$  are obtained from the sums SS. These minima should not be significantly larger than those due to the noise in the data.
- (3) Good fits are only obtained if the models are accurate. The accuracy of the models as well as the parameter estimates  $p_j$  can be judged by examination of the residuals spectra  $\Delta I_i(v)$ .
- (4) All the information about each model and its parameters contained in the spectra can be extracted since all the data are examined.
- (5) The ultimate precision is dependent on the characteristics of the experimental data. Some of these characteristics are described by the partial derivatives

$$\frac{\partial I_{\text{exp}}(\nu)}{\partial p_j}$$

(6) All of the parameters can be estimated simultaneously by using appropriate least squares regression techniques.

Niple (8,9) and Tu (10) have considered some of the limitations of the experimental data by analyzing simulated spectra of single lines. They have shown that the precisions of the retrieved parameters depend on

- (a) the SNR of the spectrum,
- (b) the number  $N_o$  of data points analyzed,
- (c) the spacing  $\delta\nu$  of the data points,
- (d) the number of models required and parameters retrieved,
- (e) the accuracy of the models,
- (f) the magnitudes and shapes of the partial derivatives  $\frac{\partial I}{\partial p_j}$ , and
- (g) the correlations between the parameters  $p_j$ .

It is necessary to use a minimum of four models containing at least six adjustable parameters to create a set of simulated signals  $I_{\text{exp } i}(\nu)$  suitable for analysis. These are briefly described in Table 2. Tu (10) has investigated the information content in spectra similar to those in Fig. 3, calculated for a Lorentz line, a triangular ISRF with  $H = \alpha$ , a background and baseline independent of  $\nu$ , and a SNR of  $10^4$ . He has estimated the dependence of the fractional uncertainty  $\Delta S/S$  in the retrieved line intensity on  $S$  as shown in Fig. 4 for the case where all six parameters are retrieved. Because of increasing correlations between some of the parameters there is a limited range of  $S$  values for which satisfactory retrievals can be made. These correlations are related to the derivatives  $\frac{\partial I(\nu)}{\partial p_j}$  shown in Fig. 5.

ORIGINAL PAGE IS  
OF POOR QUALITY

Some improvement in these fractional uncertainties is obtained by fixing one or more of the parameters, although this may introduce systematic errors if these parameters are fixed to incorrect values. It is seen, from Fig. 4, that, at small S values, the greatest improvement is obtained by fixing the baseline.

From these studies quantitative evaluations of experimental designs can be made. These have yielded some interesting results in addition to confirming the intuitive conclusions obtained by less rigorous analyses of experimental spectra. In particular

- (1) The usable range can be determined. This is the range of experimental conditions for which all six parameters of these simple models can be estimated.
- (2) For the cases considered here this range is primarily limited by correlations between S, B,  $\alpha$ , and H.
- (3) The usable range can be extended by fixing one or more of the adjustable parameters.
- (4) The systematic errors introduced by fixing the parameters may be larger than the uncertainties of the parameters retrieved by the regression methods used here. This is especially likely if the fixed parameter was highly correlated with free parameter.

#### Whole Band Analysis

For many atmospheric problems it is sufficient to know the intensity, width, and position of a few lines. However, these data also serve as the raw material from which

2. the number of parameters retrieved is relatively small and this allows their precision to be increased,
3. the usable range is larger than that for single lines, because many of the highly correlated parameters previously measured are no longer estimated directly, and
4. the accuracy of the band models can be rigorously tested.

When the band systems in Fig. 6 were analyzed (5) by this method the residuals spectrum shown in the lower part of the figure was obtained. The weak Q branch near  $2615 \text{ cm}^{-1}$  and associated P and R branch lines are just above the noise level and are too weak to be modelled satisfactorily. Even so, the sum of the squares of the differences is within 50% of that expected from the noise and this is considered to be a good retrieval.

This agreement confirms that the line shape, the band models, the ISRF, and other instrumental effects have been modelled sufficiently well to describe the experimental data. These data have an estimated SNR of 200-300 and were obtained with a Fourier Transform Spectrometer (FTS) having a spectral resolution of about  $0.06 \text{ cm}^{-1}$ . The results of the analysis of this band system have been described by Hoke (5, 6). The center of the principal band was estimated to occur at  $2614.241224 \pm 0.000055 \text{ cm}^{-1}$ , this precision is at least an order of magnitude better than is usually claimed from measurements made with instruments of similar resolution and performance.

If the background, baseline, and ISRF parameters are retrieved, together with the band intensity and line width parameters, then the line intensities and widths can be estimated to about 2%. In all other reported measurements of similar quantities these instrumental parameters have been assumed known. When our analyses were repeated with these same assumptions the uncertainties in the line intensities and widths were

ORIGINAL PAGE IS  
OF POOR QUALITY

more fundamental molecular properties can be derived, provided a sufficiently large number of lines in a vibration-rotation band can be analyzed.

The strongest band of  $\text{CO}_2$  shown in Fig. 6 contains over 100 lines. If a line-by-line analysis on this band were performed by the methods just described up to 600 parameter values would be required. However, the spectrum in Fig. 6 contains three  $\text{CO}_2$  bands with lines of measurable intensity and thus more than 1000 parameters may be required to describe this spectrum. Many of these lines lie outside the usable range described above and most are overlapped and blended with other lines. Unless the investigator makes arbitrary decisions concerning some of the adjustable parameters in these line by line models, only a small fraction of the lines can be analyzed satisfactorily.

However, the parameters describing these lines are not independent and the models for the background and baseline can be assumed to be smoothly varying functions over the band. Thus the spectrum in Fig. 6 can be described by models similar to those in Table 2 and a series of additional models which relate the parameters associated with the individual lines (5, 6). The models describing these additional relationships and the typical numbers of adjustable parameters in these models are shown in Table 3. By extending the analysis of the previous section to the retrieval of the parameters in this table, the problem is reduced to estimating of the order of 40 parameters rather than over 1000. We have used this technique to analyze overlapping bands of  $\text{CO}$ ,  $\text{CO}_2$ , and  $\text{N}_2\text{O}$  which contain over 3000 data values and several hundred lines (see Refs. 3-7). The advantages of this approach are

1. the parameters of primary interest are found directly from the experimental data,

reduced to less than 0.1%.

Portions of the experimental spectrum, the corresponding calculated spectrum, and the residuals spectrum obtained during the course of this investigation are shown in Fig. 7. Although it is difficult to detect significant differences between the observed and calculated spectra the residuals spectrum shows systematic variations attributed to poor modelling of the ISRF. These systematic differences are not seen in the final residuals spectrum shown in Fig. 8.

In our analysis of CO, CO<sub>2</sub> and N<sub>2</sub>O bands we have consistently obtained stable, reproducible solutions, but the size of the data sets which can be handled and the number of parameters retrieved is limited by the computer resources available. As these resources expand it is expected that several spectra of the same band and/or several band systems will be analyzable simultaneously. The band systems of most molecules are described by more complicated models than those required in these investigations. In these cases the whole band methods of analysis can be powerful tools for testing the accuracies of these models and for identifying the significant parameters.

These analytical methods involve the application of standard statistical techniques to the analysis of typical experimental data. However they have advantages over the more conventional methods of spectral analysis which are similar to the advantages claimed for FTS over dispersive spectroscopic techniques. These are usually described as the throughput and multiplex advantages.

By analogy with these comparisons the multiplex advantage of whole band analysis over line-by-line analysis is the direct, simultaneous retrieval of the parameters of interest. There is also a corresponding throughput advantage in that the entire data set is searched for information about each parameter.

## Selected References

1. Y. S. Chang Ph.D. diss. The Ohio State Univ. (1976).
2. Y. S. Chang and J. H. Shaw, *Applied Spectroscopy*, **31**, 218 (1977).
3. R. L. Hawkins Ph.D. diss. The Ohio State Univ. (1982).
4. R. L. Hawkins and J. H. Shaw *JQSRT*, **29**, 543 (1983).
5. M. L. Hoke Ph.D. diss. The Ohio State Univ. (1982).
6. M. L. Hoke & J. H. Shaw, *Applied Optics* **21**, 929, 935 (1983).
7. C. L. Lin, J. H. Shaw and J. G. Calvert *JQSRT* **23**, 387 (1980).
8. E. Niple Ph. D. diss. The Ohio State Univ. (1980).
9. E. Niple and J. H. Shaw *Appl. Spectroscopy* **33**, 569 (1979).
10. N. Tu and J. H. Shaw, paper CBI, Fall Meeting, Ohio Section, American Physical Society (1983).

**Table 1 Contributors to these studies**

**Y. S. Chang**

**R. L. Hawkins**

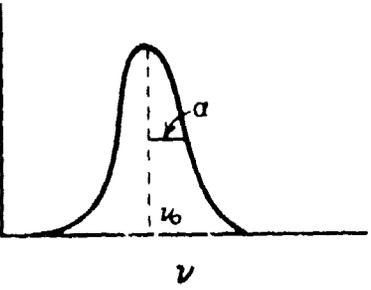
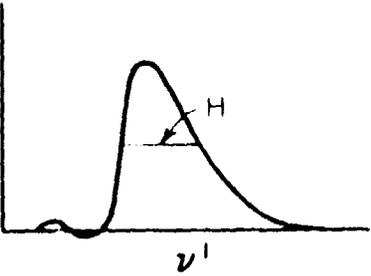
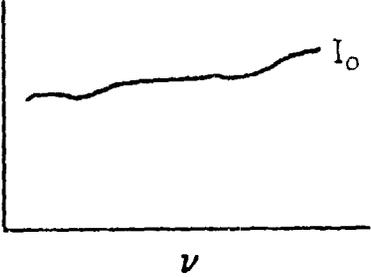
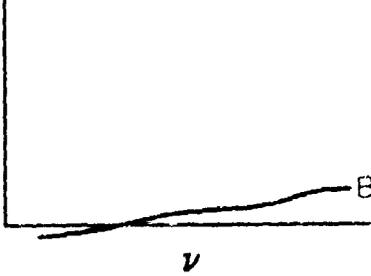
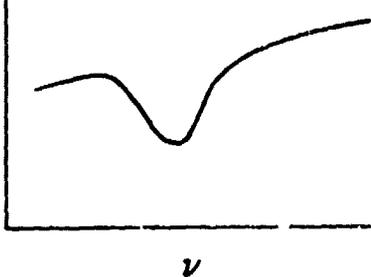
**M. L. Hoke**

**C. L. Lin**

**E. Niple**

**N. Tu**

Table 2 Single Line Models and Parameters

	Model	Minimum Parameters
<p>1 Line Shape (Lorentz, Voigt, ...)</p>		<p><math>\nu_0</math> <math>\alpha</math> <math>S = \int k d\nu</math></p>
<p>2 Instrument Spectral Response Function (ISRF) (sinc, triangle Gaussian, ...)</p>		<p>H</p>
<p>3 Background to line (polynomial in <math>\nu</math> and other terms)</p>		<p><math>I_0(\nu_1) = I_0</math></p>
<p>4 Baseline to line (polynomial in <math>\nu</math>)</p>		<p><math>B(\nu_1) = B</math></p>
<p>Synthetic spectrum</p>		

**Table 3 Models for Describing Entire Bands**

Quantity	Model	Number of Parameters
Line Intensities	Band Intensity $S_B = \sum S(\text{lines})$	~3 if interaction terms are included (the sample temperature and $E''$ line are also required)
Line Positions	$\nu = E' - E''$	For $\text{CO}_2$ the upper and lower state energies can each be modelled by expressions containing 3 or 4 parameters  <8
Line widths	The variation in the Lorentz widths over the band is modelled by empirical relations.	<10
ISRF	The models contain 3-4 parameters which are assumed constant over the band.	<4
$I_o(\nu_i)$	Several parameters may be required if channel spectra are present.	<10
$B(\nu_i)$	Usually 2 terms are adequate.	<3

Total number of parameters retrieved 30 - 40.

Figure Captions

Fig. 1. Line positions by interpolation. The position of the  $j$ th line is

$$[2622 + \Delta x_j / \Delta x_1] \text{ cm}^{-1}.$$

Fig. 2. a) The spectrum  $I_{\text{art } i}(\nu)$  of an artificial line is compared with the experimental data  $I_{\text{exp } i}(\nu)$ . b) the differences  $\Delta I_i(\nu)$  are similar in shape to the partial derivative  $\partial I_{\text{exp } i}(\nu) / \partial \nu$ . c) the best estimate of the line position occurs at the minimum value of the sum  $\sum \Delta I_i^2(\nu)$ ; d), e), f) corresponding steps required to determine the line intensity.

Fig. 3. A set of lines generated by using a Lorentz line shape, a triangular ISRF with  $H = \alpha = 0.1 \text{ cm}^{-1}$ , and constant  $I_0$  and B functions.

Fig. 4. The dependence of the fractional uncertainty in the retrieved line intensity on the line intensity for the case of 6 retrieved parameters and for the cases where B,  $I_0$ , and H are fixed to their correct values.

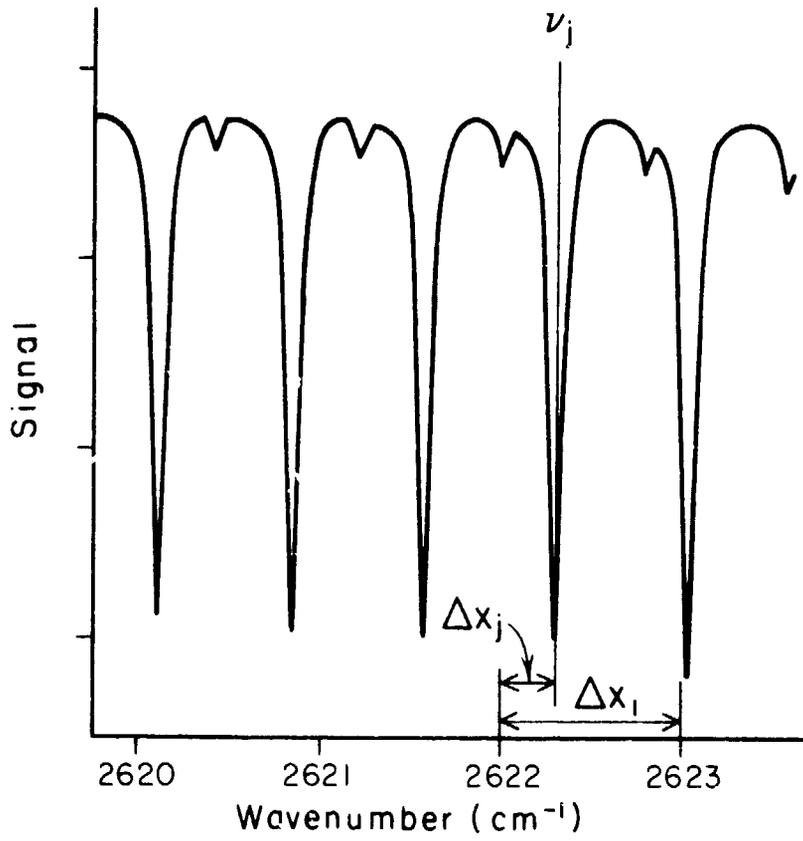
Fig. 5. The derivatives  $\partial I_i(\nu) / \partial p_j$  as functions of  $\nu$  for the parameters  $p_j = S, \alpha, \nu_0, H, I_0$  and B and the dependence of these derivatives on the line intensity S.

Fig. 6. Upper curve: a spectrum of  $\text{CO}_2$  near  $2600 \text{ cm}^{-1}$ . Lower curve: the residuals spectrum after modelling the entire spectrum and retrieving more than 30 adjustable parameters.

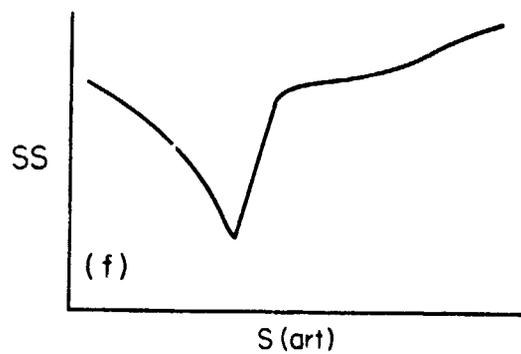
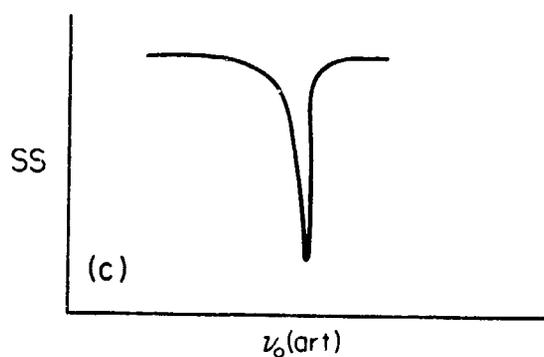
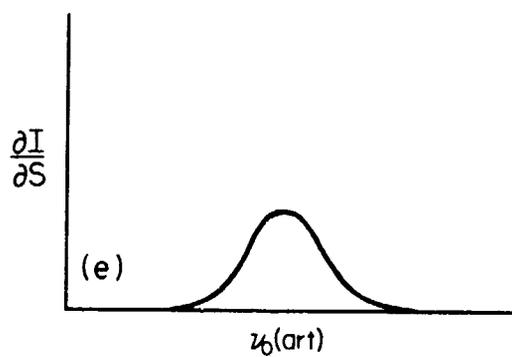
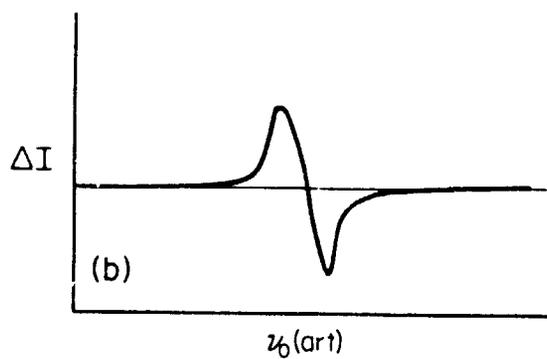
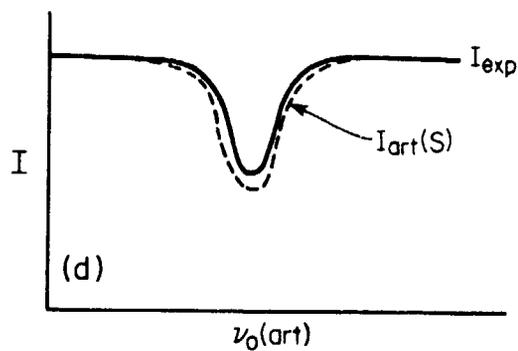
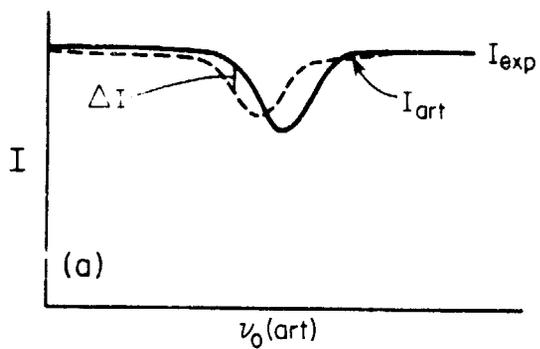
Fig. 7. The upper portion of this figure shows a comparison between a portion of the experimental spectrum in Fig. 6 and a calculated spectrum. The residuals spectrum in the lower part of this figure has systematic variations attributed to incorrect modelling.

Fig. 8. The upper part of this figure is the same portion of the experimental spectrum as that in Fig. 7. The lower part of this figure is the residuals spectrum obtained by using the best estimates for the parameters.

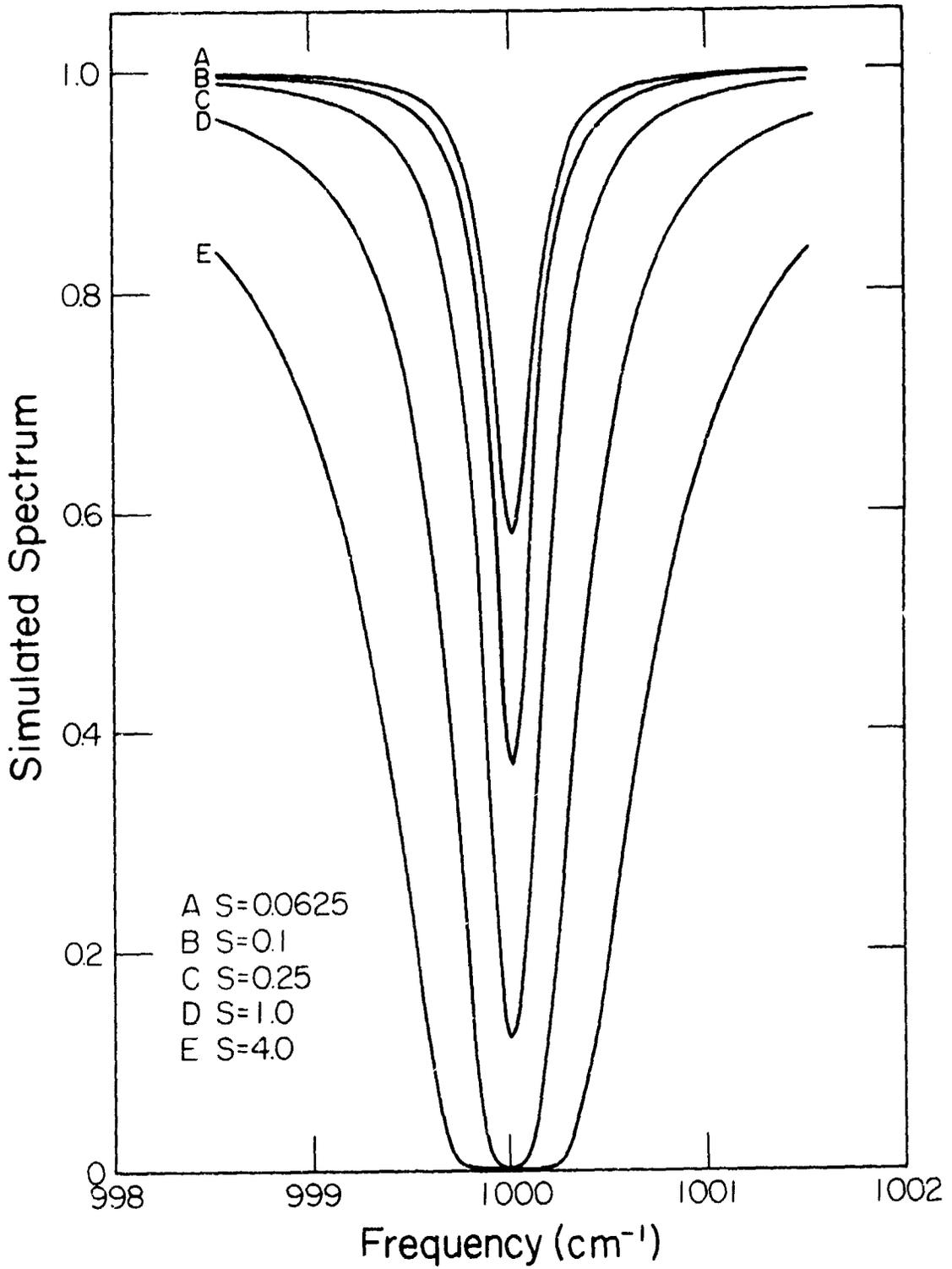
ORIGINAL PAGE IS  
OF POOR QUALITY

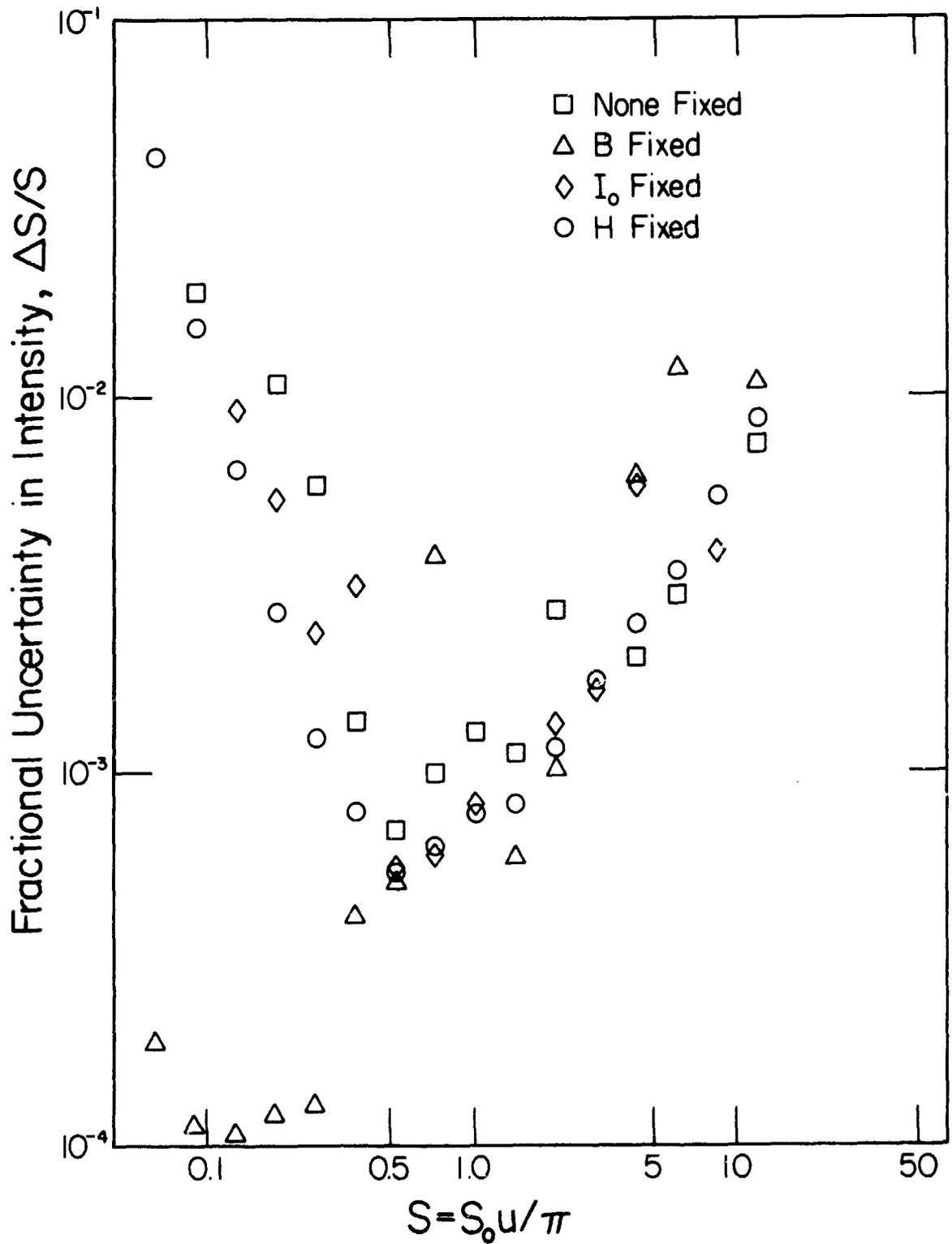


ORIGINAL PAGE  
OF POOR QUALITY

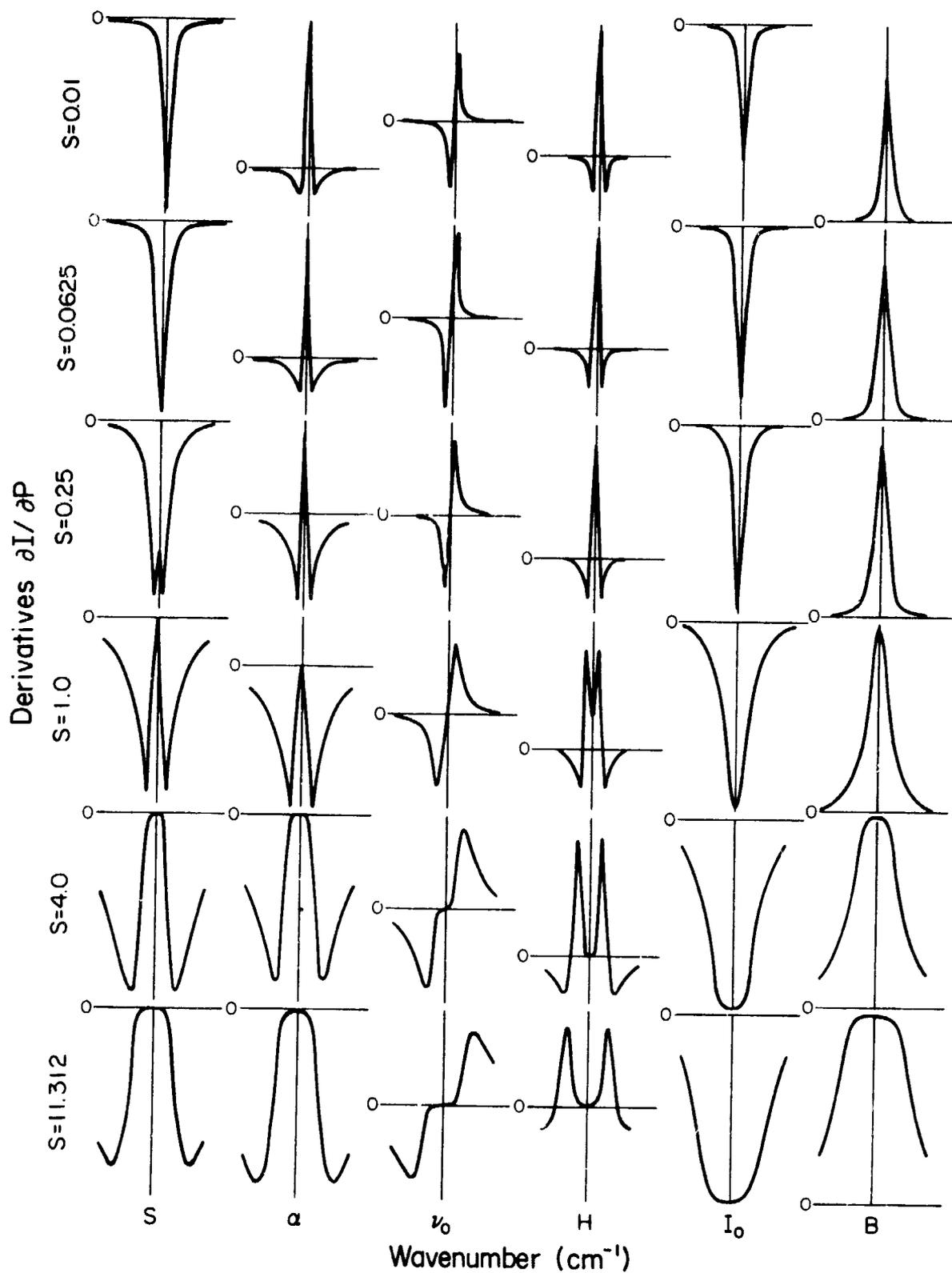


ORIGINAL PAGE IS  
OF POOR QUALITY

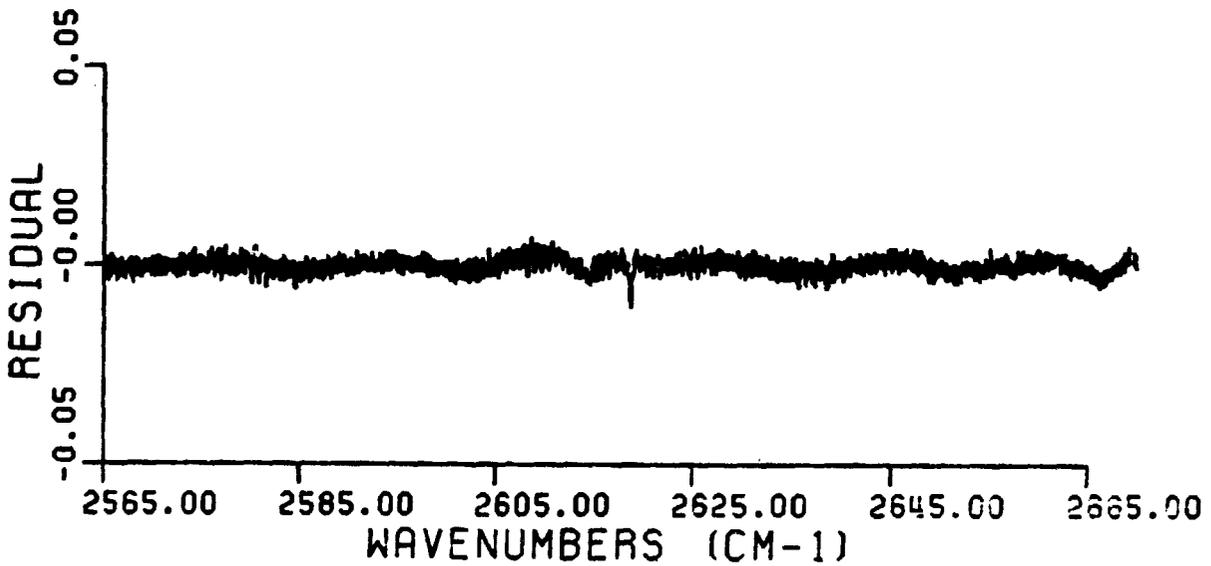
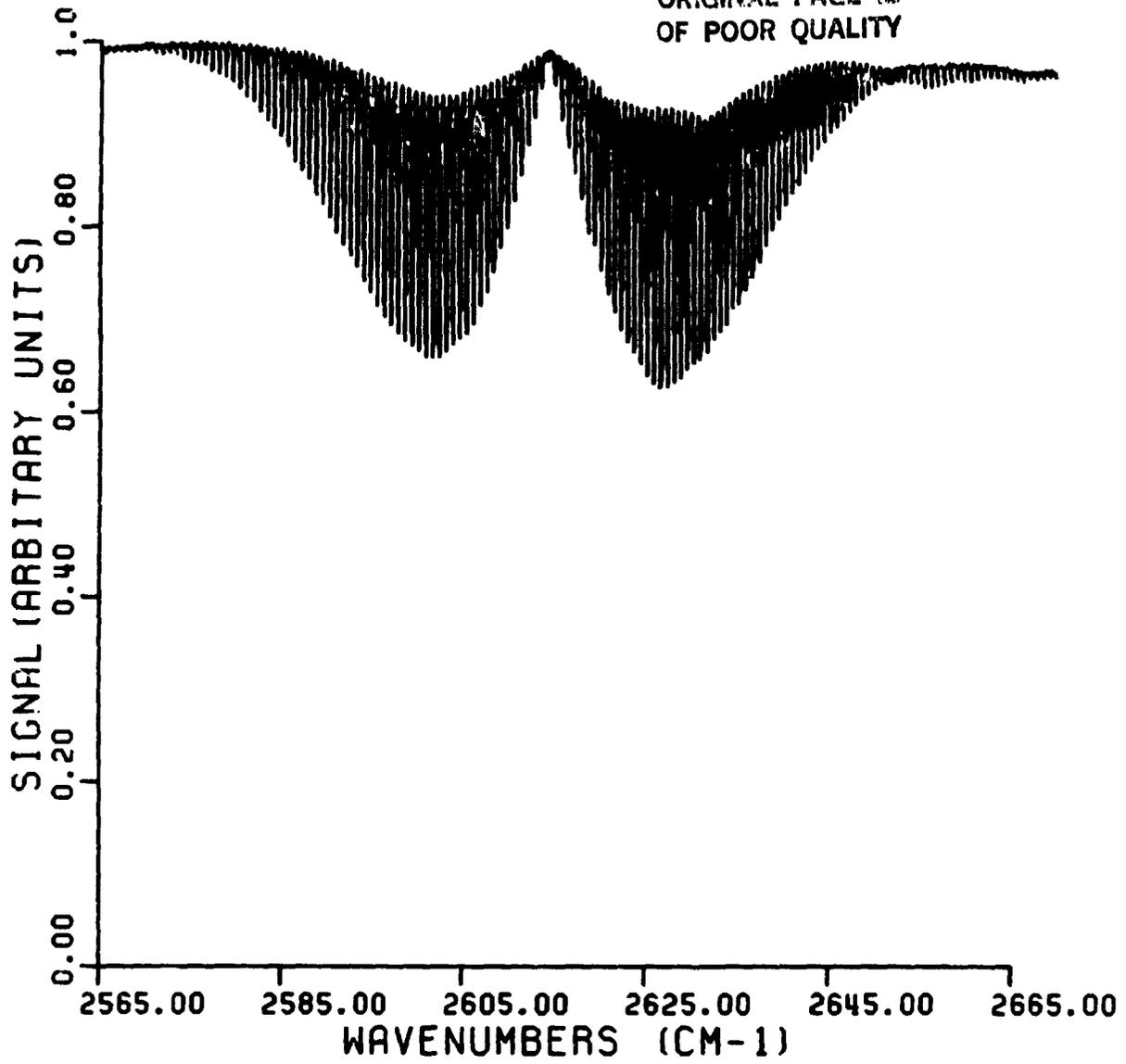




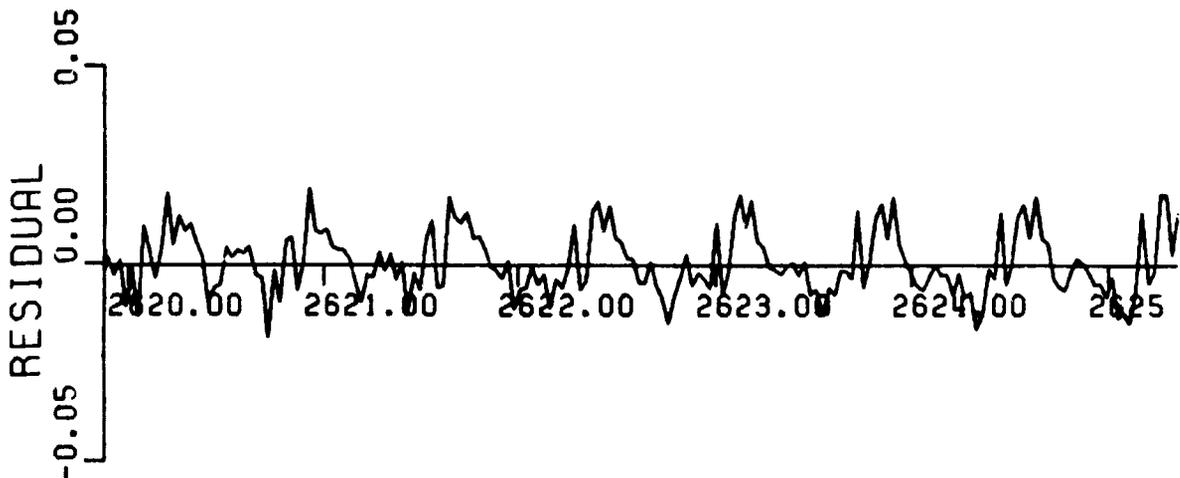
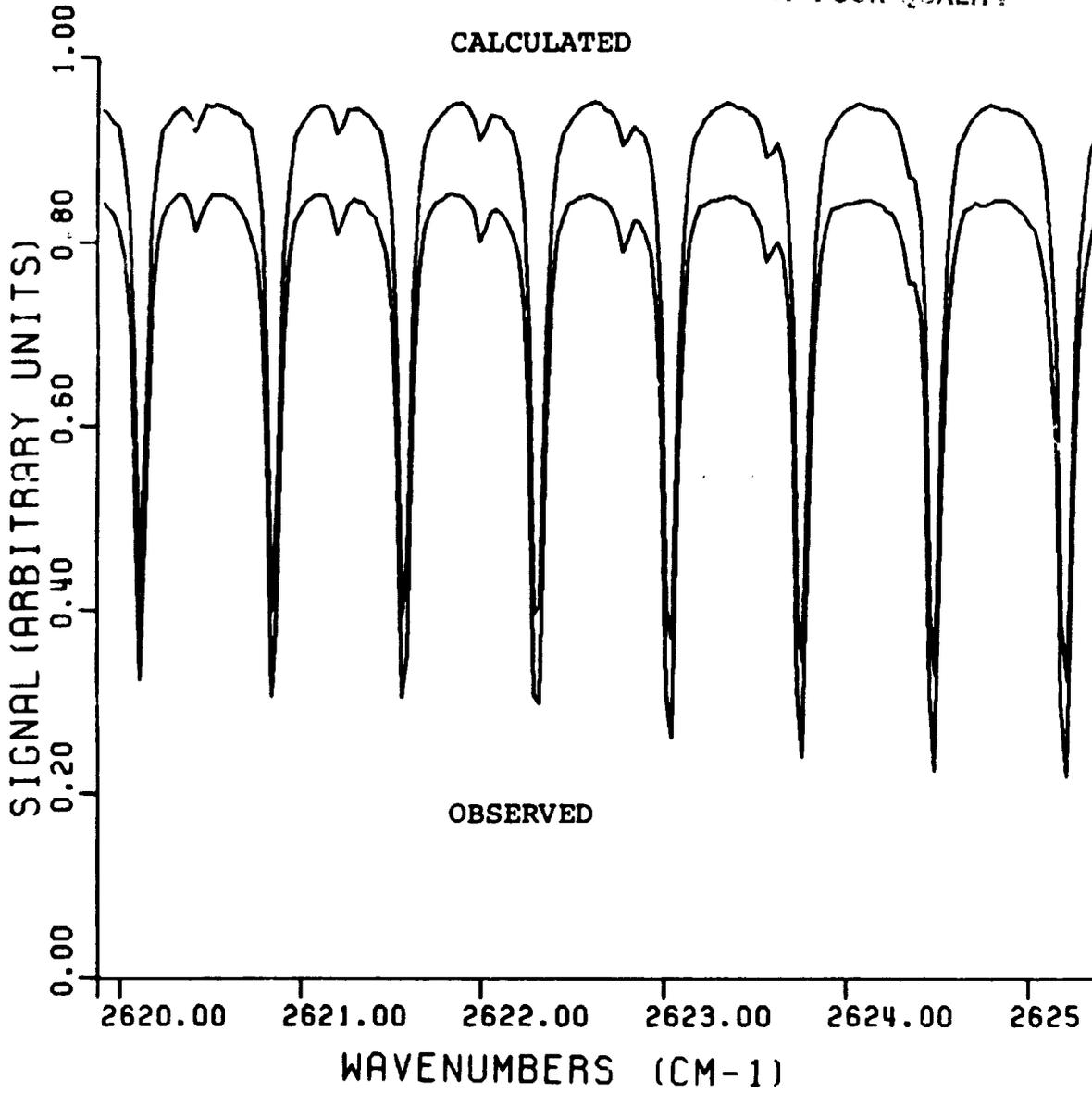
ORIGINAL FROM  
OF POOR QUALITY



ORIGINAL PAGE IS  
OF POOR QUALITY



ORIGINAL FILED  
OF POOR QUALITY



1  
OF POSITIVE

