Usage of robust methods to estimate spectroscopic parameters of vibrational bands of linear molecules from the experimental data

S.A. Tashkun

Institute of Atmospheric Optics,
Siberian Branch of the Russian Academy of Sciences, Tomsk

Received September 1, 2004

To obtain fitted spectroscopic parameters of vibrational bands of linear molecules, it is suggested to use the least median method instead of the widely used least squares method. The former is known to be robust with respect to presence of outliers in data and provides statistically justified estimates. On the contrary, the latter has zero resistance against outliers that may lead to distorted estimates and deficient models. Importance and usefulness of the least median method is illustrated with an example of deriving fitted spectroscopic parameters of the 40002–01101 band of the $^{12}C^{16}O_2$ molecule.

Introduction

The overwhelming majority of papers dealing with recording and modeling the frequencies of rotational-vibrational transitions of linear molecules employ the polynomial representation of the rotational dependence of energy levels $E(J)$ in a vibrational state v:

$$E_v(J) = G_v + B_v[J(J+1)] - D_v[J(J+1)]^2 + H_v[J(J+1)]^3, \quad (1)$$

where G_v, B_v, D_v, and H_v are the spectroscopic parameters of the state, which are determined from fitting to experimental values E_v. The sum of square deviations between the experimental and calculated energy values is used as a characteristic to be minimized. The linear least squares method (LLSM) is employed as a minimization technique, because the parameters G_v, B_v, D_v, and H_v enters into the equation for $E_v(J)$ linearly. Having the fitted parameters, it is possible to perform interpolations and extrapolations with respect to the rotational quantum number J. A disadvantage of this approach is that it fails to describe correctly the states involved in local (in terms of J) resonances, because Eq. (1) is valid only in the case of no resonance interactions. When using the LLSM, it is supposed indirectly that experimental data include no outliers, that is the model (1) can describe all data with acceptable discrepancies. Unfortunately, small discrepancies still do not guarantee the absence of outliers.

LLSM method and robust estimation methods for linear models

Theory of LLSM is well developed. A useful overview of the use of LLSM as applied to problems of molecular spectroscopy can be found in Ref. 1. The estimates of the parameters $\hat{G}_v$, $\hat{B}_v$, $\hat{D}_v$, and $\hat{H}_v$ of the model (1) are unbiased and have the smallest variance in the class of linear estimates (Gauss–Markov theorem). Statistical characteristics of the estimates (standard errors, correlation matrix, inflation variance factors) give theoretically justified information about properties of the model. Using this information, an investigator can try to construct a model, optimal in a certain sense. In addition, there are a number of high-quality computer programs, implementing this method.

A corner stone of the LLSM theory is the assumption that the sample of data has a sufficient length and errors in data are distributed according to the normal law. If this assumption is violated, most conclusions of this theory become invalid. Unfortunately, large samples with the normal distribution of measurement errors are the exception rather than the rule. For operation with short samples, in which the distribution of random errors deviates from the normal law, one should use robust methods for estimation of model parameters.[2,3] Several attempts to use such methods in molecular spectroscopy were undertaken (see, for example, Ref. 4). However, these methods are more time and memory expensive as compared with the LLSM, and their computer realization is more complicated. That is likely why such methods have not found wide utility in solving spectroscopic problems.

Upon fitting, sample data having large discrepancies call for further investigation. There are two causes for such outliers. On the one hand, they can be experimental defects (measurement errors, noise, incorrect assignment, etc.) and be meaningless from the viewpoint of physical phenomena described by the model. On the other hand, they can evidence the inadequacy of the model used. It is clear that to select one of these two versions, an investigator has to analyze the situation.

It is well known that the LLSM has zero resistance to outliers[2]: only one outlier is sufficient to

0235-6880/04/11 783-03 $02.00

change cardinally the estimates of the parameters and, consequently, the structure of the model. This outlier itself may be characterized by a minor discrepancy. To eliminate this disadvantage, robust methods are used. The difference between the LLSM and robust methods is illustrated by Figs. 1 and 2, borrowed from Ref. 3. Let we have data, consisting of three groups of points *A*, *B*, and *C*, which are to be fitted with the use of some linear model. If we use the LLSM, then the fitting line will be approximately such as shown in Fig. 1.

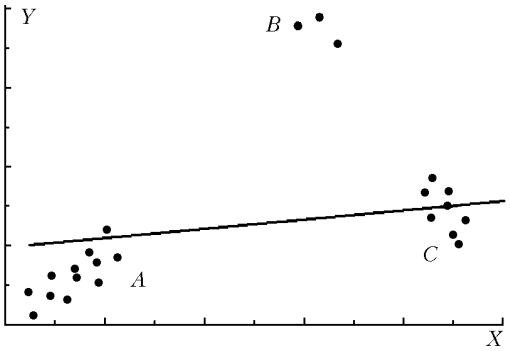

Fig. 1. Fitting by the linear least squares method.

In this case, the points of group B, having the largest discrepancies, are classified as outliers and are not described by the model. This is explained by the fact that the LLSM avoids appearance of large discrepancies and, in our case, gives the average opinion about all points. On the other hand, using the high-robust fitting, we obtain the pattern, shown in Fig. 2. It can be seen that this method gives the clear opinion about most points. The points of group B appear to be well described, but the points of group C are the obvious outliers. Evidently, the further analysis of groups B and C is needed to reveal the source of their anomalous behavior. Whether these points are noise measurements or they bear important information about the phenomena under study, which is not described by the linear model, — these questions remain open.

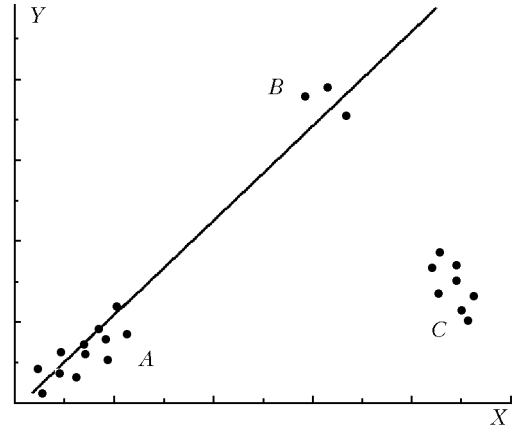

Fig. 2. Fitting by the least median method.

The least median method (LMM)[2] is the most resistant estimation method. Being developed for linear models, this method has a 50% resistance to outliers. This means that this method can find up to 49 outliers in a sample of 100 points. However, it is extremely time-expensive, because it deploys the combinatorial enumeration of all possible subsamples of a fixed length from the entire sample.

Application of LMM to analysis of CO_2 bands

As an example demonstrating the usefulness of LMM in spectroscopy, consider the problem on determination of G_v, B_v, and D_v parameters of the upper state of 40002–01101 band of the $^{12}C^{16}O_2$ molecule. Experimental wave numbers ν_{obs} were borrowed from Ref. 5. Calculated wave numbers were modeled as

$$\nu_{calc} = E_{v'}(m) - E_v(m-1), \qquad (2)$$

where $v' = 40002$, $v = 01101$; $m = -J$ for the *P*-branch and $m = J + 1$ for the *R*-branch. The parameters of the state 01101 were taken from Ref. 6.

The fitting with the use of LLSM gave the standard deviation (SD) of 0.0053 cm^{-1} and the discrepancies given in the third column of the Table.

Fitting of spectroscopic parameters of the upper state of 40002–01101 band with the use of LLSM and LMM

Line	ν_{obs}, cm^{-1}	LLSM, cm^{-1}	LMM, cm^{-1}	Weight
*P*5	4804.2693	−0.01151	−0.01481	0.0
*P*7	4802.7082	−0.00452	−0.00816	0.0
*P*9	4801.1437	0.00140	−0.00269	1.0
*P*11	4799.5747	0.00537	0.00083	1.0
*P*13	4798.0009	0.00737	0.00249	1.0
*P*15	4796.4086	−0.00600	−0.01097	0.0
*P*17	4794.8386	0.00637	0.00167	1.0
*P*19	4793.2507	0.00457	0.00057	1.0
*P*21	4791.6578	0.00176	−0.00107	1.0
*P*23	4790.0614	−0.00036	−0.00162	1.0
*P*25	4788.4613	−0.00191	−0.00135	1.0
*P*27	4786.8574	−0.00303	−0.00072	1.0
*P*29	4785.2504	−0.00323	0.00029	1.0
*P*31	4783.6415	−0.00172	0.00176	1.0
*P*33	4782.0302	0.00030	0.00155	1.0
*P*35	4780.4175	0.00287	−0.00158	1.0
*R*3	4811.3016	−0.00998	−0.01433	0.0
*R*5	4812.8643	−0.00389	−0.00890	0.0
*R*7	4814.4257	0.00333	−0.00231	1.0
*R*9	4815.9807	0.00684	0.00073	1.0
*R*13	4819.0752	0.00771	0.00173	1.0
*R*15	4820.6162	0.00728	0.00211	1.0
*R*17	4822.1483	0.00201	−0.00177	1.0
*R*19	4823.6814	0.00213	0.00027	1.0
*R*21	4825.2068	−0.00079	−0.00035	1.0
*R*23	4826.7270	−0.00409	−0.00123	1.0
*R*25	4828.2445	−0.00519	−0.00027	1.0
*R*27	4829.7579	−0.00561	0.00036	1.0
*R*29	4831.2690	−0.00385	0.00122	1.0
*R*31	4832.7766	−0.00165	−0.00064	1.0
*R*33	4834.2823	0.00177	−0.00602	0.0
*R*39	4838.7915	0.00911	−0.07542	0.0
*R*41	4840.2943	0.00622	−0.13047	0.0
*R*43	4841.7919	−0.00908	−0.21753	0.0
SD, cm^{-1}		**0.0053**	**0.0014**	

Column 1 shows the line assignment, and column 2 presents the experimental wave numbers. All the data had unity weight. The largest discrepancy is within 0.012 cm^{-1}, and the model (2) seems to be adequate to the experimental data.

Analyze the same data with the use of LMM. The program implementing this method was borrowed from Ref. 2. The LMM found 9 outliers among 34 points, and zero weights were assigned to them, as shown in column 5 of the Table. Then the rest 25 points were fitted again using the LLSM. The standard deviation obtained in this fitting became as low as 0.0014 cm^{-1}, and the discrepancies are given in column 4. The calculated discrepancies of the zero-weight points are underlined. The largest discrepancy for the points, included in fitting, is 0.0027 cm^{-1}. Among the zero-weight data, of greatest interest are $R39$–$R43$ lines, for which the discrepancy is as high as 0.21 cm^{-1}. Other lines are apparently noisy.

What is the cause for inadequacy of the model for the case of $J \sim 40$? The answer was found from the global fitting of $^{12}C^{16}O_2$ line positions with the model of effective Hamiltonian given in Ref. 7. It appeared that the state 40002 is in a strong resonance with the state 21113, and the maximum of the resonance interaction falls just on $J \sim 40$ [Ref. 8]. It is clear that the simple low-order polynomial model (2) cannot correctly describe the local (in terms of J) interaction. Thus, the robust fitting method proved to be capable of separating the data, which are inadequate to the model, whereas the ordinary LLSM fails to do this.

Conclusions

The aim of this paper is to attract attention of investigators dealing with assignment of spectra of linear molecules to the possibility of using the LMM, which is characterized by a higher resistance to outliers in the experimental data as compared with that of the widely used LLSM. The LMM performs combinatorial enumeration of subsamples of fixed length from the sample of fitted data, and therefore it is quite time-expensive.

In addition to the models considered in this paper, this method can also be used to determine experimental energy levels from observed transition frequencies using the basic quantum-mechanics Ritz principle [see, for example, Ref. 9].

References

1. D.L. Albritton, A.L. Schmeltekopf, and R.H. Zare, in: *Molecular Spectroscopy: Modern Research*, ed. by K.N. Rao (Academic Press, New York, 1976), Vol. 2, pp. 1–67.
2. P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection* (Wiley, New York, 1987), 328 pp.
3. F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel, *Robust Statistics: The Approach Based on Influence Functions* (Wiley-Interscience, New York, 1986).
4. A.F. Ruckstuhl, W.A. Stahel, and K. Dressler, J. Mol. Spectrosc. **160**, No. 2, 434–445 (1993).
5. Ph. Arcas, E. Arie, M. Cuisenier, and J.P. Maillard, Can. J. Phys. **61**, 857–866 (1983).
6. L.S. Rothman, R.L. Hawkins, R.B. Wattson, and R.R. Gamache, J. Quant. Spectrosc. Radiat. Transfer **48**, Nos. 5/6, 537–566 (1992).
7. S.A. Tashkun, V.I. Perevalov, J.-L. Teffo, L.S. Rothman, and Vl.G. Tyuterev, J. Quant. Spectrosc. Radiat. Transfer **60**, No. 5, 785–801 (1998).
8. D. Bailly, S.A. Tashkun, V.I. Perevalov, J.-L. Teffo, and Ph. Arcas, J. Mol. Spectrosc. **190**, No. 1, 1–6 (1998).
9. S.A. Tashkun, V.I. Perevalov, J.-L. Teffo, A.D. Bykov, and N.N. Lavrentieva, J. Quant. Spectrosc. Radiat. Transfer **82**, Nos. 6/8, 165–196 (2003).