Error Estimates for Fitted Parameters: Application to HCl/DCl Vibrational–Rotational Spectroscopy

Scott E. Feller*
Department of Chemistry, Wabash College, Crawfordsville, IN 47933-0352; *fellers@wabash.edu*

Charles F. Blaich
Department of Psychology, Wabash College, Crawfordsville, IN 47933

A standard experiment in physical chemistry laboratory courses is to analyze vibrational–rotational spectra of isotopic mixtures of hydrogen chloride. The spectroscopic parameters extracted from these data can then be related to molecular properties such as bond strength and bond length (*1*). A recent article in this *Journal* described the determination of spectroscopic parameters via a nonlinear least-squares fit of the spectral data to an anharmonic oscillator/distorted rotator model (*2*). This method has several advantages over the analysis traditionally employed for this experiment (*1*) and exposes students to the technique of nonlinear regression, a method made accessible by tools available in standard spreadsheet packages (*3*). Most importantly, the connection between experimental result and theoretical model is much clearer when students fit their data directly to the equations of the model, rather than transforming the equations into a form amenable to linear regression.

Unfortunately, as noted by the authors of ref *2*, the nonlinear least-squares approach does not directly provide error estimates for the fitted parameters. A second recent article in this *Journal*, however, described a method for the determination of parameter uncertainties in nonlinear fits that relies on numerically finding the partial derivatives of the fitting function with respect to the parameters using finite differences (*4*). Although they did not demonstrate their method on the analysis of infrared diatomic spectra, it could well have been applied to this problem. A drawback to their approach is the mathematical sophistication required of students if they are to use the algorithm as anything other than a "black box".

Finally, it should be noted that within many popular models, including the example presented here, analysis of infrared spectra does not require nonlinear regression. Schwenz and Polik have shown that the multiple linear regression method can be applied successfully to this problem (*5*) (multiple linear regression techniques are also available in many standard spreadsheet packages). They pointed out that the nonlinear least square method should give the same result as their multiple linear regression technique. They noted the disadvantages of the nonlinear regression, namely, there is no guarantee that the nonlinear regression converges to the global best fit and the calculation of errors in the fitting parameters is difficult. The nonlinear regression method, however, is attractive in that it can be easily extended to more complicated (i.e., truly nonlinear) spectroscopic models.

To summarize, the advantage of the nonlinear fitting procedure is its simplicity to set up as a spreadsheet and its direct connection between experimental data and theoretical model, which students can readily comprehend. The greatest weakness of nonlinear regression is that the mathematics of the parameter uncertainty calculation can be complicated.

In the following we describe an alternative approach to the analysis of the HCl infrared spectrum experiment. This method uses the nonlinear Solver approach described in ref *2* and extends it in a straightforward manner to the calculation of uncertainties using the technique of statistical Monte Carlo sampling. The approach offers several opportunities for students to learn topics from statistics and model fitting, and it emphasizes the connection between uncertainty in the measured data and the uncertainty in the parameters being fit to a theoretical model.

First, we present a short review of a model for vibrational and rotational energy levels and how this can be fit to experimental data. We then describe a Monte Carlo sampling technique that provides rigorous error estimates for each of the fitted parameters. In addition to providing error estimates for the present case of a nonlinear least squares analysis, this exercise more clearly demonstrates some of the assumptions commonly used in software packages for linear least squares fitting. The Microsoft Excel spreadsheet and Visual Basic macro used to perform these calculations are available as supplemental material[W] and could easily be altered for a variety of model-fitting applications.

Theoretical Model

The vibrational energy levels for a diatomic molecule, taking into account the lowest order anharmonic term in the potential energy, are

$$E(v) = \omega_e(v + \tfrac{1}{2}) - \omega_e\chi_e(v + \tfrac{1}{2})^2 \qquad (1)$$

where the energy, E, and fundamental vibrational frequency, ω_e, are given in wavenumbers (cm^{-1}), and the anharmonicity constant, χ_e, is unitless. The rotational energy levels, also given in units of cm^{-1}, are

$$E(J) = BJ(J + 1) - DJ^2(J + 1)^2 \qquad (2)$$

where B is the rotational constant and D is the centrifugal distortion constant. Combining eqs 1 and 2 and writing the rotational constant as an explicit function of the vibrational quantum number yields

$$E(v,J) = \omega_e(v+\tfrac{1}{2}) - \omega_e\chi_e(v+\tfrac{1}{2})^2 + B(v)J(J+1) - DJ^2(J+1)^2 \qquad (3)$$

The dependence of the rotational constant on vibrational quantum state is often represented by the relation

$$B(v) = B_e - \alpha_e(v + \tfrac{1}{2}) \qquad (4)$$

resulting in a model that contains five parameters: ω_e, χ_e, B_e,

Table 1. Spectroscopic Constants and Associated Uncertainties from Student Infrared Spectra and Monte Carlo Sampling

Molecule	ω_e /cm^{-1}	χ_e	B_e /cm^{-1}	α_e /cm^{-1}	D/cm^{-1}	χ^2	σ_{fit} /cm^{-1}	k/N m^{-1}	r_e /pm
H^{35}Cl	2989.533 (0.012)	0.0173546 (0.0000013)	10.5910 (0.0005)	0.30265 (0.00006)	0.000522 (0.000004)	36.3	0.027	515.825 (0.004)	12.7470 (0.0003)
H^{37}Cl	2987.279 (0.013)	0.0173425 (0.0000013)	10.5739 (0.0005)	0.30209 (0.00008)	0.000511 (0.000004)	33.1	0.044	515.848 (0.004)	12.7476 (0.0003)
D^{35}Cl	2144.523 (0.006)	0.0124970 (0.0000008)	5.4485 (0.0002)	0.11214 (0.00003)	0.000142 (0.000002)	31.7	0.012	516.028 (0.003)	12.7462 (0.0003)
D^{37}Cl	2141.395 (0.012)	0.0124838 (0.0000020)	5.4321 (0.0006)	0.11175 (0.00008)	0.000141 (0.000005)	31.0	0.030	516.042 (0.006)	12.7466 (0.0007)

NOTE: Uncertainty values are in parentheses.

α_e, and D. Note that other models for vibrational–rotational energy levels are possible, having a greater or fewer number of parameters. The model chosen here is typical of those presented in undergraduate physical chemistry texts. The procedure to be described is equally applicable to other models.

Data Acquisition

Students acquired infrared spectra of the fundamental and first overtone bands for gas-phase HCl and DCl using a Mattson 5000 series FTIR with 0.04 cm^{-1} resolution, allowing study of four isotopic combinations: H^{35}Cl, H^{37}Cl, D^{35}Cl, D^{37}Cl. For each molecule, spectral transitions were assigned to changes in vibrational ($v_i \rightarrow v_f$) and rotational ($J_i \rightarrow J_f$) quantum numbers. Class data were pooled to find the mean value and variance for the location of each spectral peak. (An important benefit of this exercise is that in setting up the spreadsheet calculation students often discover errors they made in the assignment of quantum numbers or in recording the peak frequencies.)

Least-Squares Fitting

A spreadsheet is constructed containing a column of experimentally measured mean transition frequencies (ω_i^{DATA}) and four columns containing the initial and final vibrational and rotational quantum numbers assigned to each transition. Then, a column of predicted frequencies (ω_i^{MODEL}) is constructed using the differences in energy levels given by eq 3. The reader is referred to ref *2* for explicit spreadsheet formulas that can be used with only slight modification. The best values of the model parameters are then obtained by minimizing the weighted sum of the squares of the difference between experimental data and theoretical model

$$\chi^2 = \sum_i \left(\frac{\omega_i^{MODEL} - \omega_i^{DATA}}{\sigma_i} \right)^2 \qquad (5)$$

where the summation includes all the spectral peaks for a given isotope (~10 peaks in both the R and P branches of the fundamental and first overtone bands in the present case), and the σ_i give the standard deviation associated with each of the data points. In this work the Solver function available in Microsoft Excel was employed for the minimization procedure, but similar functionality is provided in other spreadsheet packages (*3*).

As described in refs *1* and *6*, the assignment of uncertainties in the data is crucial for rigorous model fitting. However, this fact is unappreciated by many practitioners of least-squares fitting. Ideally, the σ_i should be calculated from the variance of a large number of repeated measurements for each data point. If repeated measurements are available for only a few points, or if a small number of measurements have been made for each point, it may be best to assume a uniform standard deviation for the data. In the case where no repeated measures are obtained, the σ_i could be assigned uniformly on the basis of additional considerations such as the resolution of the spectrophotometer. In the present case the second approach was used because each data point was measured only four times (once by each student team), resulting in assumed σ_i between 0.013 and 0.077 cm^{-1} for the four isotopic combinations.

Reasonable initial guesses for the spectroscopic parameters (e.g., literature values from any physical chemistry textbook) result in good convergence of χ^2 to its minimum value. Students should be encouraged, however, to investigate the effect of convergence criteria, minimization algorithm, and initial parameter values on the χ^2 value obtained. Table 1 gives the values of the spectroscopic parameters and χ^2 for each isotope.

Assuming the assignment of uncertainty in the data is correct, χ^2 provides a measure of the validity of the model. An approximate guideline is that χ^2 should be approximately equal to the number of degrees of freedom, df = $m - n$, where m is the number of data points and n is the number of parameters being determined. In the present case, 36 spectral peaks were used to determine 5 parameters, giving df = 31, approximately equal to the χ^2 values given in Table 1. A more quantitative criterion comes from the fact that the probability distribution for the minimized χ^2 is given by the chi-square distribution if the model is correct. At the 90% confidence level the upper and lower limits of the chi-square distribution are 19.3 and 45.0, respectively. This means that χ^2 greater or less than these values would occur by chance only 10% of the time if the assumed model is correct, providing a guide for rejecting incorrect models. A second measure of the quality of the fit is the root mean square difference between the observed frequencies and those predicted by the model. This quantity, labeled σ_{fit} in Table 1, is approximately equal to the standard deviation of the data. In other words, the difference between data and model are comparable to the random measurement errors in the data.

Monte Carlo Sampling

Having obtained best values for each fitted parameter and having tested the validity of the model, the remaining task is to determine error estimates for each parameter. For linear least-squares fitting the uncertainty in each parameter is determined at the minimized χ^2 value, from the uncertainty in each data point and the value of each data point, using matrix algebra. For the simple case of fitting data to a straight line, this approach yields relatively simple analytic formulas. For more complicated functional forms the calculation is tedious but can be carried out by most software packages. It is important to emphasize that the distinction between linear and nonlinear least squares fitting is made on the basis of how the model is a function of the parameters, not on how the model is a function of the independent variable(s). For example, fitting data to a polynomial expansion

$$f(x,\alpha_0,\alpha_1,\alpha_2,\alpha_3) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 \qquad (6)$$

to determine the parameters $\{\alpha_i\}$ is a linear least squares fitting problem. Examples of functions that are nonlinear in their parameters include

$$f(x,\alpha_0) = \sin(\alpha_0 x); \quad f(x,y,\alpha_0,\alpha_1) = \alpha_0 x + \alpha_0\alpha_1 y \qquad (7)$$

For nonlinear least squares fitting problems, parameter uncertainty is more difficult to determine. Fortunately, there is a general technique for estimating these parameters based on generating a large number of synthetic data sets so that many χ^2 minimizations can be carried out and the distribution of fitted parameters studied. This method has been successfully applied to many problems of nonlinear model fitting in the physical sciences (see for example refs *7* and *8*). The central idea behind this Monte Carlo sampling is that the finite-sized data sets collected in the laboratory are drawn from a distribution of possible experimental results (due to random measurement errors). Monte Carlo sampling allows a quantitative determination of precision by generating a large number of data sets consistent with the experimentally measured data (i.e., drawn from a normal distribution with the same mean observed in the laboratory and a standard deviation equal to the error estimate for each data point). After carrying out the fitting procedure on each member of the synthetic data set, a distribution of parameter values is obtained whose standard deviations are the uncertainty in the fit parameters. The text *Numerical Recipes* by Press et al. (*6*) (to which the reader is referred for a thorough discussion of this method) describes the power in this way: "Offered the choice between mastery of a five-foot shelf of analytical statistics books and middling ability at performing statistical Monte Carlo simulations, we would surely choose to have the latter skill."

Fifty synthetic data sets were constructed by performing a random draw from a normal distribution for each of the 36 peak spectral frequencies. The mean and standard deviation of each normal distribution were set to the mean and the error estimate of the students' measurements for each spectral peak. We performed these random draws using a combination of Excel's RAND and NORMINV functions. The NORMINV function takes a cumulative probability value and returns a value for a normal distribution with a given mean and standard deviation. The RAND function generates a random number from a uniform distribution between 0 and 1. By inserting the RAND function into the NORMINV function and including the appropriate information from the student measurements into the NORMINV function, we could create a synthetic data set each time the spreadsheet was recalculated. Each synthetic data set then undergoes the least-squares fitting procedure. From the parameter distribution the standard error is obtained for each fit parameter. The entire procedure is run as a Microsoft Excel macro using Microsoft Visual Basic and requires a few minutes of computer time.

Table 1 presents results for the spectroscopic constants using this method. The confidence limits are very small with relative errors on the order of 10^{-5} and 10^{-6}. The precision of these students' results is largely attributable to the high instrumental resolution available in our laboratory. Greater uncertainty in the location of spectral peaks will naturally lead to larger uncertainty in the parameters. An attractive feature of this approach is that students can easily construct the distribution of any molecular property related to the fit parameters. As examples, we determined the uncertainties in equilibrium bond length and force constant and included these in Table 1. This allows the student to decide if there are statistically significant differences in bond length or strength between the isotopic combinations. With this set of student data, there are significant differences between the hydrogen- and deuterium-containing isotopes. Presumably, this difference arises because the deuterium compounds fit the model better (the vibrational states studied are in a region of the potential curve that is more reasonably well represented by a single anharmonic term).

Summary

This exercise provides several benefits to students in a modern physical chemistry course. The uncertainties in both spectroscopic and molecular properties are very small, illustrating the power of spectroscopic methods to determine molecular structure. The fitting procedure is relatively straightforward, allowing a strong connection between theory and experiment. If multiple spectra are not available to explicitly calculate the standard deviation in peak location, students could assume uncertainties in their data based on the resolution of their instrument and then proceed to obtain approximate error estimates for their fit parameters via the Monte Carlo sampling method. Repeating this procedure with varying levels of assumed instrumental resolution would explicitly demonstrate the relationship between instrumental resolution and the precision of molecular properties.

This analysis of the data brings together a number of important topics in statistics and model building, with numerous possible extensions. For example, different models for fitting the data (e.g., harmonic vs anharmonic oscillator, or rigid vs nonrigid rotor) could be compared. This analysis, which is covered in many physical and analytical chemistry texts, involves application of the *F* test and shares many spreadsheet calculations with the nonlinear fitting and Monte Carlo sampling described here. Advanced students may wish to try reducing the number of parameters. For example, by writing ω_e as a function of the force constant, B_e as a function of the bond length, and D as a function of both k and r_e, the

analysis can be done solely in terms of k and r_e. Alternatively, parameters could be added to allow the centrifugal distortion constant to be a function of vibrational quantum state (in a form analogous to the rotational constant, B). An interesting project would be to compare the uncertainty in force constant and bond length obtained from infrared spectroscopy with that from quantum chemical calculations (e.g., by investigating different basis sets).

The method of error analysis described here is general and can be employed as an alternative to linear least-squares fitting. For functions other than a straight line or simple polynomial, the use of a general minimization routine such as Solver is often easier than setting up a spreadsheet for multiple linear regression. This is especially true for multivariate problems such as the analysis of the HCl/DCl.

Acknowledgment

This work was supported by the National Science Foundation through grant MCB-9896211.

[W]Supplemental Material

The Microsoft Excel spreadsheet and Visual Basic macro used to perform these calculations are available as supplemental material in this issue of *JCE Online*.

Literature Cited

1. Shoemaker, D. P.; Garland, C. W.; Nibler, J. W. *Experiments in Physical Chemistry*; 5th ed.; McGraw-Hill: New York, 1989; Experiment 38.
2. Iannone, M. *J. Chem. Educ.* **1998,** *75*, 1188–1189.
3. Clark, R. W. *Chem. Educator* **1999,** *4*.
4. de Levie, R. *J. Chem. Educ*. **1999,** *76*, 1594–1598.
5. Schwenz, R. W.; Polik, W. F. *J. Chem. Educ.* **1999,** *76*, 1302–1307.
6. Press, W. H.; Teukolsky, S. A.; Vetterling, W. T; Flannery, B. P. *Numerical Recipes in Fortran: The Art of Scientific Computing*; Cambridge University Press: Cambridge, 1992; pp 650–694.
7. Wiener, M. C.; White, S. H. *Biophys. J.* **1992,** *61*, 434–447.
8. Armen, R. S.; Uitto, O. D.; Feller, S. E. *Biophys. J.* **1998,** *75*, 734–744.