Automatic assignment and fitting of spectra with PGOPHER.

Colin M Western[1], Brant E Billinghurst[2]

Abstract

An initial implementation of a tool for automatic assignment of spectra within the PGOPHER program is presented, together with its application to rotational analysis of the ν_{11} band of *cis*-1,2-dichloroethene. It is based on the AUTOFIT algorithm presented by N. A. Seifert et al. (*J. Mol. Spectrosc.*, 2015, **312**, 13) but implemented in a more efficient and general way, allowing it to be applied to a much wider variety of spectra.

Introduction

High resolution spectroscopy, particularly where rotational structure is resolved, is an excellent source of precise and unambiguous information on the structure and bonding in molecules. Unfortunately the richness of the information available can be accompanied by significant problems in extracting the information present, specifically in assigning the spectra. Traditionally a trial and error approach is used based on looking for patterns in spectra and comparing and fitting to models. Computers can assist with this; one of the first such tools involves Loomis-Wood plots [1], of which there have been several implementations; see for example [2] and [3]. Genetic algorithms have also been used to automate the trial and error process associated with assignment[4], though these seem to have had relatively few applications. Recent developments in assignment methods include the Automated Spectra Assignment Procedure (ASAP) [5] which is based on cross-correlations within a spectrum to pick out common differences, and automated microwave double resonance spectroscopy[6] which provides a more direct experimental method for finding linked transitions.

This paper is inspired by another recent development, the AUTOFIT algorithm[7] which is used to assign microwave spectra of asymmetric top molecules. The essential method is to try all possible assignments of three rotational transitions to observed lines within a given window, and then check each assignment by looking for other transitions calculated from the *A*, *B* and *C* values determined from the three assigned lines. This paper presents a much more general and efficient implementation of this algorithm, integrated directly into the PGOPHER program[8], which can be used on any spectrum with reasonably resolved rotational structure. It can be applied to any type of spectrum handled by the PGOPHER program, so includes linear molecules and symmetric tops, hyperfine structure and open shell molecules, and vibrational and electronic structure. In addition, the direct integration allows significant speed ups in the search process. The AUTOFIT program was chosen as the basis for the add-on to PGOPHER as it is relatively simple to program, is generally applicable and is tolerant of the presence of other species or transitions in the spectra.

The new automated assignment and fitting process is best understood with the aid of an illustrative example. For this purpose a high resolution infra-red spectrum of *cis*-1,2-

[1] School of Chemistry, University of Bristol, UK, C.M.Western@bristol.ac.uk

[2] Canadian Light Source, University of Saskatchewan, Canada, Brant.Billinghurst@lightsource.ca

dichloroethene was used which is dense, but resolvable. For this molecule the vibrational frequencies are known[9] but the only high resolution spectrum available is in the microwave region[10]. In this paper new spectra of *cis*-1,2-dichloroethene are recorded using the Canadian Light Source and the algorithm is applied to give the first rotational analysis of the ν_{11} band at 570 cm^{-1}; it is an isolated band with a large isotope shift [11] that makes a good test case.

The Assignment and Fitting Algorithm

The algorithm requires the following inputs:

1. A list of experimental line positions and (optionally) intensities
2. A set of starting molecular parameters sufficient for an approximate simulation of the spectra of interest.
3. The n_{par} molecular parameters to float and (optionally) a possible range for each.
4. A small number, n_{fit} of "fit" transitions and a search window (± w_{s}) associated with these.
5. A (possibly larger) number, n_{check}, of "check" transitions and an acceptance window (± w_{a}) associated with these.
6. Optionally, any lines already assigned.

The algorithm starts by calculating the starting positions of the fit transitions from the initial set of constants and then works through each possible set of assignments of the fit transitions to the observed line positions, limited only by the requirement that each assigned line is no more than w_{s} from the initial calculated position. For each of these trial assignments a standard least squares fit to the "fit" transitions (and any already existing assignments) is done to determine the n_{par} floated parameters, and these are used to calculate positions of the "check" transitions. A count, n_{OK} is then made of the number of check transitions whose calculated positions that are within ± w_{a} of an observed line, and an average residual (σ_{check}) is also calculated for these lines. At the end of the process the best fits are presented, ranked by n_{OK} and σ_{check}. The current algorithm does not make use of intensity information, other than to present, for each fit, the sum of the intensities, I_{sum}, of the assigned lines. However the user interface allows an easy visual check of each of the fits presented, and as seen in the example below, typically allows the correct fit to be identified easily.

Experience with this algorithm indicates it can be very effective, but the limiting factor is the number of trial assignments that can be tested. The program has some important optimisations to speed this up. Firstly, the calculations associated with each trial are independent, so are done in parallel on a multi CPU system. Secondly the fit associated with each trial can be significantly speeded up by pre-calculating the initial derivatives required for the fitting process. As each fit starts from the same point the first fit cycle can then be performed with only a few matrix operations on a small (n_{fit} × n_{fit}) matrix. The second fit cycle only recalculates line positions and re-uses the initial derivatives, and only if three or more cycles are required are the derivatives recalculated. Depending on the starting position chosen, two cycles are often sufficient. This is further speeded up by rejecting trial assignments at an early a stage as possible. Rejection occurs if the predicted mean error from the fit is more than the acceptance window, w_a, or if the parameters are moved outside the given possible range. Such rejections typically occur on the first cycle.

Two simple heuristics can optionally be applied to increase the selectivity. The first is to specify a maximum permitted number of blends (`MaxBlends`), i.e. assignments of multiple transitions to the same observed line. This prevents degenerate assignments with everything assigned to the same line (which may be achievable by setting all constants to zero or similar trivial values), and is likely to be a sensible constraint in many cases. Secondly the assigned frequencies can be required (with the `KeepOrder` flag) to be in the same order as in the starting spectrum; this covers the common case where a comb of lines (such as in a P or R branch) is to be assigned, but the numbering is uncertain.

However, even with all the optimizations discussed above, it is also important for the user to make a sensible choice of starting position, particularly of the parameters to float. As the number of trials increases as a high power of n_{fit} this suggests choosing $n_{\mathrm{fit}} = n_{\mathrm{par}}$ and limiting both by choosing a subset of transitions that are determined by a small number of parameters. This is illustrated below for an asymmetric top where the initial searches are limited to two or three parameters out of the four (or more) needed for a full fit. Other parameters are then determined by further related searches, typically over a reduced range. Once the essential set of parameters is obtained the assignment can be rapidly extended to a large number of lines by drastically narrowing w_s, making searches involving many lines very rapid. Alternatively a search involving “check” transitions only (with a small w_a) will simply assign to the nearest line.

Experimental

Spectra of *cis*-1,2-dichloroethene were collected by making use of the high brightness[12] provided by the Far-Infrared beamline at the Canadian Light source using a Bruker IFS125HR spectrometer equipped with a KBr beamsplitter, and an Infrared Laboratories Ge:Cu detector. A 2-meter White type gas cell with a path length of 72 meters was used with a *cis*-1,2-dichloroethylene pressure of 0.035 Torr. A notch filter for the 490-1190 cm^{-1} range was used to limit the spectral window. The spectrum was collected at room temperature, using a scanner velocity of 80 KHz and averaging over 400 scans. The raw experimental spectrum was converted to a list of frequencies and intensities by a tool added to PGOPHER for this purpose, though other methods, such as peak finders provided with spectrometer software could also be used.

Application to *cis*-1,2-dichloroethene

An initial simulation could be generated easily using rotational constants from the microwave spectrum[10] and manually adjusting the upper state constants to give reasonable agreement with a published rotationally unresolved spectrum[11]. This allows the region around the origin of the $^{35}Cl_2$ species to be identified as reasonably clear of interference from the $^{35}Cl^{37}Cl$ species, so this is the region initially worked on.

While the initial spectrum looks qualitatively correct, in that a sequence of sub-bands with approximately the right spacing is visible, the assignment is not clear. In the low *J* region four constants (the band origin and the three rotational constants) are required for a simulation, but the search space can be reduced by only looking for selected transitions. The molecule is a near prolate asymmetric top, so the energy levels for a given K_a will approximately follow $\nu_{\mathrm{eff}}(K_a) + \bar{B}J(J+1)$ where $\bar{B} = ½(B+C)$ provided K_a is not too small. This suggests an initial search using only 2 parameters for states with a specific K_a. Selecting $\Delta J = \pm1$ transitions with $K'_a = 6$, O− symmetry gives ~16 lines in the region of interest, which show a simple P and R branch type pattern reminiscent of linear molecules when plotted. Performing a search with, for example, $^{q}P_{6,6}(11)$ and $^{q}P_{6,9}(14)$ as fitted transitions and the remaining 14 as check transitions using $w_s = 0.3$ cm^{-1} (approximately twice the spacing between the selected lines within a given branch) and $w_a = 0.001$ cm^{-1} (the instrumental linewidth) covers 18,392 possible assignments. If the range of possible movement of the band origin is limited to ±0.1 cm^{-1} (the accuracy with which the origin can be estimated from the spectrum without assignment) the search takes about a second with 15,556 assignments rejected. The top four trial fits are shown in figure 1.

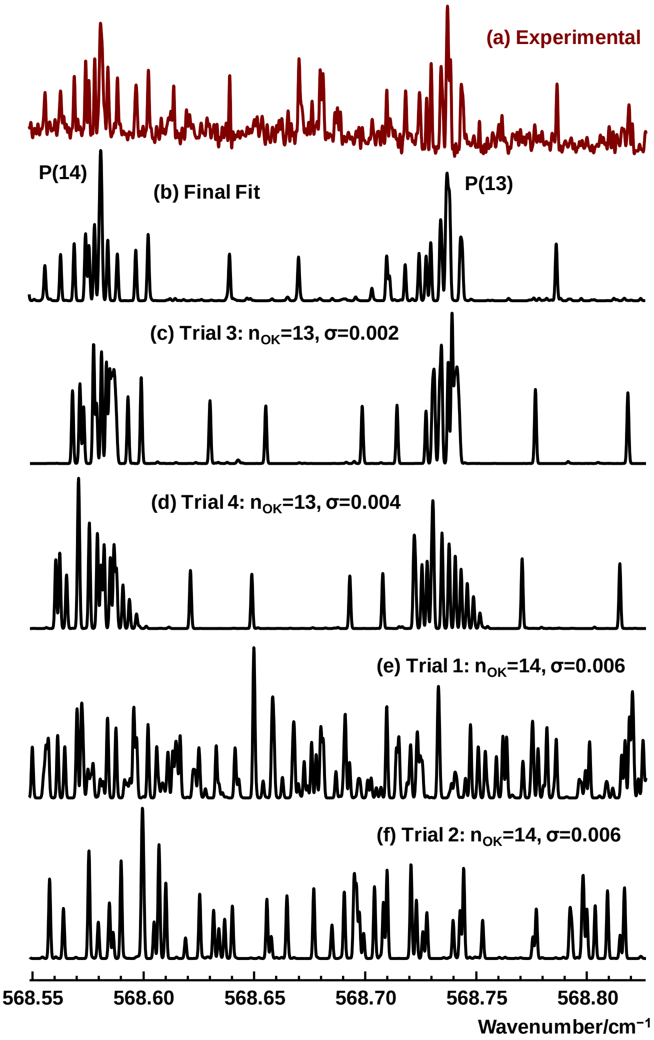

Figure 1. Short section of the experimental spectrum (a), together with the final fit (b, Table 1 below) and the initial 4 trial fits, (c)-(f). For clarity, the two better trial fits, (c) and (d), are shown at the top though they are ranked third and fourth by the program. This region only shows transitions from the $^{35}Cl_2$ species.

The two highest ranked trial fits (1(e) and 1(f)) are clearly wrong, but the next two (1(c) and 1(d)) are promising, showing two clusters of lines that approximately match the experimental spectrum. These clusters consist of $\Delta J = -1$ transitions with $J'' = 13$ and 14 (marked P(13) and P(14) in the figure), with individual lines having different K_a values. In fact both 1(c) and 1(d) have the same J assignments, and only differ in the specific line picked from each cluster. This suggests the J numbering has been correctly determined, and a reasonable value obtained for B' but the assignment of the specific line within each cluster to $K_a = 6$ may not be correct.

The next step is to discard the initial set of assignments and fit one of the clusters, which will approximately follow $\nu_{\text{eff}}(J) + AK_a^2$, suggesting a second two parameter search, this time to determine A'. The P(13) group looks reasonably clear; here we choose a search using $K_a = 5$ and 6 as fit transitions and the remaining P(13) transitions with $K_a \geq 4$ as check transitions. The lower K_a transitions are excluded, as the asymmetry splitting means they do not follow the simple formula. The search range can be much reduced, so only 169 trial assignments are required, though both the `MaxBlends` and `KeepOrder` speed-ups must be turned off as the K_a structure displays a band head. Of the possible assignments presented, the first gives a good fit to 15 transitions. This determines A', and $\delta' = B'-C'$ can then be determined with a further search using a single fit transition – the strongest low K_a line is the obvious choice – giving a further rapid search.

These successive searches are enough to give a good determination of the origin and all three rotational constants and the assigned line list can be rapidly extended as the predicted lines (at least for transitions with similar J values) are typically within the linewidth of the observed transitions, and can simply be assigned to the nearest transition. A tool is available for walking up (or down) in J in steps of 1, automatically assigning and fitting transitions with the new value of J by assigning to the nearest line. A related tool makes it easy to spot near misses or blends that should be excluded from the final fit. In the current case this allows the assignment to be rapidly extended to $J \sim 38$. At this point the overlap between the isotopologues (and possibly hot bands) means the assignment becomes more difficult and there is also some evidence of localised perturbations, so we halt the process at this point. This yields more than sufficient observations (1024, about 2/3 of the strong P branch lines in the region) to determine the rotational constants (table 1), including quartic and sextic centrifugal distortion terms. The average error is 0.00016 cm^{-1} compared with a linewidth of 0.001 cm^{-1}. A section of the final simulation is shown in figure 1(b).

Table 1. Rotational constants[a] (/cm^{-1}) of the ν_{11} band of *cis*-1,2-dichloroethene

	$CH_2{}^{35}Cl_2$	$CH_2{}^{35}Cl^{37}Cl$
Origin	570.766752(21)	568.386933(24)
A[b]	0.38383926(27)	0.38111002(19)
$\bar{B} = ½(B+C)$	0.07716251(14)	0.07531819(11)
$\Delta = B-C$	0.01554532(55)	0.01491999(40)
Δ_K	1.38195(145) × 10^{-6}	1.34870(56) × 10^{-6}
Δ_{JK}	-3.6893(146) × 10^{-7}	-3.5034(60) × 10^{-7}
Δ_J	4.9550(288) × 10^{-8}	4.505(14) × 10^{-8}
δ_K	1.625(140) × 10^{-7}	-4.5(68) × 10^{-9}
δ_J	1.3323(281) × 10^{-8}	1.085(13) × 10^{-8}
Φ_K	1.2358(626) × 10^{-10}	-[c]
Φ_{KJ}	-1.3692(1097) × 10^{-10}	-[c]
Φ_{JK}	3.1632(4075) × 10^{-11}	-[c]
Φ_J	2.321(173) × 10^{-12}	-[c]
ϕ_K	1.0629(924) × 10^{-9}	-[c]
ϕ_{JK}	1.764(111) × 10^{-10}	-[c]
ϕ_J	1.09(14) × 10^{-12}	-[c]
n_{obs}	1024	685
σ	0.00016	0.00017
J'_{max}	38	31

[a] Figures in brackets are the standard deviation in units of the least significant figure.

[b] Centrifugal distortion constants are in terms of Watson's A reduced Hamiltonian[13], using a *Ir* representation. Ground state constants are taken from the microwave data[10].

[c] Fixed at ground state value[10].

The $^{35}Cl^{37}Cl$ species is a little trickier to assign, as it is more difficult to identify regions clear of $^{35}Cl_2$ lines, though fortunately the region immediately to higher frequency of the central Q branch of the $^{35}Cl_2$ species is suitable. In this case a three parameter search was required to start with, with three transitions (specifically $^{q}R_{6,10}(16)$, $^{q}R_{6,11}(17)$ and $^{q}P_{7,11}(17)$) chosen to determine the band origin, A and $\bar{B}$. The basis for this is as above – the higher K_a lines follow the symmetric top formula $AK_a^2 + \bar{B}J(J+1)$ to a reasonable approximation. The first of the resulting suggested fits was good, and the same process as for the other isotopologue could then be followed with very quick searches, determining $B-C$ by a search on a low K_a transition and then stepping up in J to cover the entire spectrum. In this case the process was stopped at J' = 31 again due to increasing numbers of overlapping transitions and small perturbations reducing confidence in additional assignments. Table 1 gives the final constants determined.

The initial search for the $^{35}Cl^{37}Cl$ case was more time-consuming as three parameters were involved, but still practical with 6.3×10^{6} trials taking about 11 minutes on a standard desktop machine. It illustrates the importance of restricting searches where possible; an otherwise similar four parameter search for might require ~10^{9} trials or a day on the same computer. This is however still feasible, particularly if a high performance computer is used.

An important question is the quality of the automatic assignments produced. Even on a clear spectrum with no other species present some assignments to blends are to be expected, and assignments to accidentally co-incident lines (possibly from other species) are entirely possible. This is partly mitigated by only including the stronger lines in the automatic assignment process, but a manual check of the simulated and observed spectrum provides an important independent verification, as the observed intensities are not otherwise used. In this respect, it is no better or worse than traditional assignment methods. For the current spectrum there are sufficient strong unblended lines that any line with a significant residual can simply be excluded from the fit, and simulation of blended lines used as an independent confirmation of the assignment. It is also important to note that the process is tolerant of additional lines in the experimental spectrum; the requirement is only that a reasonable number are clear. The biggest limitation is that the initially selected "fit" transitions must be clear and unperturbed in the experimental spectrum, and experience suggests this is where manual trial and error to select these transitions is required, though the graphical user interface makes this easy.

The above provides an outline of the process; a detailed walk through of the assignment process used here, including screen shots, the spectrum used, intermediate and final data files and a line list has been added to the set of files included with the version 10.0 PGOPHER distribution. This is freely available from the website[14] and has been deposited in the University of Bristol data repository[15]. The walk through also illustrates the additional tools provided by the graphical user interface to support the search, including plots of selected subsets of lines to help identify possible searches, quick generation of line lists with selected *J,* symmetry or K_a, and identification and removal of mis-assignments. A walk through of a simple linear molecule example is also included.

Conclusions

The above introduces a very promising tool for assigning complex spectra. The tool as presented is general purpose, and should be applicable to any well-resolved spectrum. Initial trials on other systems, including quadrupole structure in microwave spectra and electronic spectra of open shell linear molecules, indicates similarly promising results. The method given here is more general than other methods; for example the ASAP method[5] is not suitable here as any given upper state is typically only involved in two transitions, which does not give sufficient selectivity for the ASAP method. A genetic algorithm[4] may also be problematic as only a partial model for the spectrum is available here; there are other lines present, presumably from hot bands and possibly the $^{37}Cl_2$ species. This is likely to be a general issue; broadband microwave spectra, the focus of the original AUTOFIT paper[7], typically show multiple overlapped spectra from different conformers and isotopologues. With the current implementation this is a positive feature, as can be seen in table S1 in the supplementary information, which shows the results of applying the techniques described here to selected microwave data from reference [7]. Six different isotopologues are seen, and automated searches in general are most likely to be successful for microwave spectra as the effective line density (compared to the resolution) is more favourable for microwave spectra than FTIR spectra.

The tool is currently being tested on a wider variety of systems to indicate optimizations and refinements, and to determine the best strategies to use for different types of molecules. An obvious enhancement is a more formal way of making use of intensity information, as predicted intensities are easily available. Any assignment involving a strong simulated line to a weak experimental line is suspicious and there may also be scope for automatically detecting blended lines. Ideally the program should be able to assign spectra with little knowledge of spectroscopy on the part of the user; the program is not at that stage, but is suitable for an experienced spectroscopist to rapidly assign spectra. The interactive nature of the program does, however, provide tools to assist in gaining the required understanding of the spectrum.

Acknowledgements

I would like to acknowledge helpful discussions with Nick Walker and his group. The spectra shown here were taken at the Canadian Light Source, which is supported by the Canada Foundation for Innovation, Natural Sciences and Engineering Research Council of Canada, the University of Saskatchewan, the Government of Saskatchewan, Western Economic Diversification Canada, the National Research Council Canada, and the Canadian Institutes of Health Research.

References

1. F. W. Loomis and R. W. Wood, *Physical Review*, 1928, **32**, 223-236.
2. B. P. Winnewisser, J. Reinstädtler, K. M. T. Yamada and J. Behrend, *J. Mol. Spectrosc.*, 1989, **136**, 12-16.
3. W. Łodyga, M. Kręglewski, P. Pracna and Š. Urban, *J. Mol. Spectrosc.*, 2007, **243**, 182-188.
4. W. Leo Meerts and M. Schmitt, *International Reviews in Physical Chemistry*, 2006, **25**, 353-406.
5. M. A. Martin-Drumel, C. P. Endres, O. Zingsheim, T. Salomon, J. van Wijngaarden, O. Pirali, S. Gruet, F. Lewen, S. Schlemmer, M. C. McCarthy and S. Thorwirth, *J. Mol. Spectrosc.*, 2015, **315**, 72-79.
6. M.-A. Martin-Drumel, M. C. McCarthy, D. Patterson, B. A. McGuire and K. N. Crabtree, *The Journal of Chemical Physics*, 2016, **144**, 124202.
7. N. A. Seifert, I. A. Finneran, C. Perez, D. P. Zaleski, J. L. Neill, A. L. Steber, R. D. Suenram, A. Lesarri, S. T. Shipman and B. H. Pate, *J. Mol. Spectrosc.*, 2015, **312**, 13-21.
8. C. M. Western, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 2016, **186**, 221-242.
9. H. J. Bernstein and D. A. Ramsay, *The Journal of Chemical Physics*, 1949, **17**, 556-565.
10. L. A. Leal, J. L. Alonso and A. G. Lesarri, *J. Mol. Spectrosc.*, 1994, **165**, 368-376.
11. S. W. Sharpe, T. J. Johnson, R. L. Sams, P. M. Chu, G. C. Rhoderick and P. A. Johnson, *Appl. Spectrosc.*, 2004, **58**, 1452-1461.
12. M. Winnewisser, B. P. Winnewisser, F. C. De Lucia, D. W. Tokaryk, S. C. Ross and B. E. Billinghurst, *Physical Chemistry Chemical Physics*, 2014, **16**, 17373-17407.
13. J. K. G. Watson, *J Chem Phys*, 1967, **46**, 1935-1949.
14. C. Western, http://pgopher.chm.bris.ac.uk PGOPHER, a Program for Simulating Rotational, Vibrational and Electronic Spectra, http://pgopher.chm.bris.ac.uk.

15. C. Western, *PGOPHER version 10.0*, University of Bristol Research Data Repository 10.5523/bris.160i6ixoo4kir1jxvawfws047m, 2016.