Robust Estimation in the Analysis of Complex Molecular Spectra

by

Werner A. Stahel, Andreas F. Ruckstuhl,
Peter Senn and Kurt Dressler,

Seminar für Statistik and Department of Chemistry,
Swiss Federal Institute of Technology (ETH), Zurich,
Switzerland

Research Report No. 65
1993, 5, 14

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

Robust Estimation in the Analysis of Complex Molecular Spectra

Werner A. Stahel, Andreas F. Ruckstuhl,
Peter Senn and Kurt Dressler,

Seminar für Statistik and Department of Chemistry,
Swiss Federal Institute of Technology (ETH), Zurich, Switzerland
Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

1993, 5, 14

Abstract

In atomic and molecular spectroscopy, measured transition energies represent differences between the energy levels of the emitting or absorbing system. In complex spectra of such transitions the identification of the pair of states corresponding to an observed transition energy is a difficult process. The assignment of transition energies to energy levels proceeds in stages. At each stage, some identifications will be erroneous. Robust fitting methods have proven extremely helpful for finding errors and for obtaining reliable identifications and estimates of energy levels.

Keywords: Robust estimation, L1 estimation, Analysis of Variance.

1 Introduction

Insight into the physical structure of atoms and molecules can be obtained by assessing the pattern of energy levels associated with their quantum mechanical states. Experiments associated with this topic form the scope of *atomic and molecular spectroscopy.* Accurate direct measurements of energy levels are not possible in general. Instead, transitions among the different states of the molecule can be observed. Thus, only separations among energy levels are available for analysis. The data we were dealing with stem from electromagnetic radiation induced by electric discharges in gaseous tritium. In the discharge the tritium molecules T_2 are excited through bombardment with electrons passing from the cathode to the anode. The excited molecules "fall back" towards the ground state in cascades of deexcitation steps.

For some such *transitions* a photon is emitted whose energy can be observed. This leads to the simple model

$$Y_{\ell u} = -\theta_\ell + \theta'_u + \varepsilon_{\ell u}, \tag{1}$$

where $Y_{\ell u}$ is the measured transition energy from the *upper state* u to the *lower state* ℓ, and $\varepsilon_{\ell u}$ is the random measurement error. The quantities θ_ℓ and θ'_u are the unknown energy levels of the two states. Since they can be determined only up to an additive constant, the energy level of the ground state – that is, the lowest θ_ℓ – is defined to be zero. The θ values are called *term values* in atomic and molecular spectroscopy.

This model can be written in the form of a linear regression as

$$Y = X\theta + \varepsilon, \tag{2}$$

where Y and ε are the vectors of the $Y_{\ell u}$'s and the $\varepsilon_{\ell u}$'s, respectively, $\theta = \begin{pmatrix} \theta \\ \theta' \end{pmatrix}$ is the vector of the unknown energy levels, and X is the design matrix, $X = [X^0 \quad X']$ with

$$X^0_{(\ell u),k} = \begin{cases} -1 & \text{if} \quad \ell = k \\ 0 & \text{if} \quad \ell \neq k \end{cases} \qquad X'_{(\ell u),k} = \begin{cases} 1 & \text{if} \quad u = k \\ 0 & \text{if} \quad u \neq k \end{cases}$$

The problem is the one-dimensional analogue of the basic problem in landscape triangularization. It also occurs in other fields of science.

The familiar method for estimating the parameters in (1) or (2) is least squares fitting. For larger problems including a few hundred or a few thousand possible states, a storage problem can appear in computation, and methods for avoiding it have been recently proposed by J.K.G. Watson (1989).

In complex spectra, there is a fundamental difficulty, however: If non-selective modes of excitation are used – such as in the electric discharges from which our data originate – then for each one of the huge number of observed transition energy values both the initial upper and the final lower states, u and ℓ, are in question. A wealth of skill and specialized methods is needed to associate each observed energy with a pair of states. In larger studies, this identification task is extremely difficult. Usually one starts from some tentative assignments, fits a theoretical model and uses the fit to find more identifications in the manner discussed below (Section 5). At all stages, some identifications will typically be erroneous.

The data set which will be partly analyzed in this paper consists of some 20.000 transition energies measured in the emission spectrum of T_2. The spectrum was obtained together with other molecular hydrogen spectra at Johns Hopkins University by G. H. Dieke and coworkers over a time period spanning fourty years (1924-1965). For the stable molecules H_2 and D_2 analyses of the data have been published (Dieke, 1958; Crosswhite, 1972; Freund et al., 1985), but for the radioactive molecule T_2 very little has appeared in print (Dieke and Cunningham, 1965; Dieke and Tomkins, 1951). Nevertheless, when Prof. Dieke died unexpectedly in 1965, he had started to make assignments in the T_2 spectrum. Upon closer examination, some of his roughly 4000 assignments turned out to be self-contradictory or incompatible with the H_2 and D_2 data. They had certainly been considered as tentative by Dieke and had not been finalised for publication.

In Dieke's T_2 data the measured transition energies, expressed as wave numbers, range from the near infrared ($\simeq 8500cm^{-1}$) to the ultraviolet ($\simeq 32.200cm^{-1}$). The standard deviation of the measurement errors is known to be $\simeq 0.02\ cm^{-1}$. If no gross errors were present, the standard errors of the estimated energy levels should be even smaller. However, when energy levels and transitions are associated incorrectly, the resulting gross errors will distort the least squares fit. Robust fitting methods are thus called for. The goal of this paper is to compare such procedures with least squares (Section 3).

The stationary states of a diatomic molecule such as T_2 are characterized by a quantum number for rotation, j, a quantum number for nuclear vibration, v, and an electronic state e. (More precisely, j is the angular momentum apart from spin.)

Within each electronic state the energy levels follow approximately a polynomial law of low order in v and $j(j+1)$. More precisely,

$$\theta_\ell = \sum_{w,k} \alpha_{e_\ell}^{(w,k)} (v_\ell + \frac{1}{2})^w [j_\ell(j_\ell+1)]^k + \eta_\ell \tag{3}$$

where v_ℓ and j_ℓ are the quantum numbers associated with state ℓ of the molecule, the summation is over a set $\{w+k \le m\}$ with $m \le 2$, say, and η_ℓ is the (deterministic) remainder term. Using an analogous equation for the upper state, with electronic state e, the observed quantities are modeled by

$$\begin{aligned} Y_{\ell u} \quad = \quad & - \sum_{w+k \le m_{e_\ell}} \alpha_{e_\ell}^{(w,k)} (v_\ell + \frac{1}{2})^w [j_\ell(j_\ell+1)]^k \\ & + \sum_{w+k \le m_{e_u}} \alpha_{e_u}^{(w,k)} (v_u + \frac{1}{2})^w [j_u(j_u+1)]^k + \tilde{\varepsilon}_{\ell_u}, \end{aligned} \tag{4}$$

where $\tilde{\varepsilon}_{\ell_u} = -\eta_\ell + \eta_u + \varepsilon_{\ell_u}$ contains a systematic and a random error.

The coefficients $\alpha_e^{(w,k)}$ lead to spectroscopic constants which are of interest in themselves. Again, they can be estimated by least squares or robust methods (Section 4).

Relation (3) is an approximation. There are localized deviations from such smooth behavior in v and $j(j+1)$. These perturbations are of special interest as they lead to new basic research problems. In order to measure such distortions, it is important to obtain extremely precise estimates for the term values from (1) and for the coefficients in model (4). We give some more details on applications of the models and some conclusions in Sections 5 and 6.

For more details about the relevant theory of molecular spectroscopy the reader is referred to Herzberg (1950). In order to facilitate cross-references, the notation used in that theory and in this paper is compared in Table 1.

[Table 1 about here]

2 The Data

The basic data set described above has been made available by H. M. Crosswhite. In this study, we examine a subset of 1429 transition energies for which tentative iden-

This paper	Molecular spectroscopy	
θ	T	term value
Y	ν	transition energy
$\varepsilon,\ \eta$	$\Delta\nu$	error
v	v	vibrational quantum number
j	N	rotational quantum number
e		electronic state
$\alpha^{(w,k)}$	Y_{wk}	Dunham coefficients

Table 1: Notations in this paper compared to usual notation in molecular spectroscopy

tifications with so-called triplet states were available. Each of these “spectral lines” thus consists of a value $Y_{\ell u}$, an upper state u characterized by the electronic state e_u and the quantum numbers v_u and j_u, and a lower state ℓ characterized by e_ℓ, v_ℓ and j_ℓ of the molecule, following Dieke. The upper electronic states e_u for which transitions are analyzed in this paper are denoted by $3b$, $3c^+$, $3c^-$, $4b$, $4c^+$, $4c^-$, and $5c^-$. The lower electronic state is always the same and is called $2a$.

The laws of quantum mechanics imply that only selected pairs of ℓ and u stand for possible transitions. The rule which applies to our data says that for transitions

- from b and c^+ electronic states to $2a$, the difference $j_u - j_\ell$ is either +1 or -1;
- from c^- to $2a$, j_u and j_ℓ must be identical.

This rule implies that the states of the molecule fall into one of two subsets between which transitions are virtually impossible and are therefore not observed. One subset contains all (lower) $2a$ and (upper) c^- states with even rotational quantum number j and all (upper) b and c^+ states with odd j – and vice versa for the other subset. For the T_2 molecule these subsets correspond to the so-called para and ortho modification, respectively.

The data set consisted of 662 transitions for the para and 767 transitions for the ortho modifications, involving 253 and 283 term values, respectively. The data set can be obtained from the authors.

Remember that the assignments were tentative. It turned out that some identical assignments were given to more than one transition. At most one of these assignments can be correct. In addition, some transitions were given more than one assignment. Such multiple assignments can be correct, as two nearly identical spectral lines can amalgamate into a so-called unresolved blend due to the limited resolution of the measurements and to physical reasons. This gives rise to an additional complication in the identification process.

3 Estimation of Term Values

Model (1) was fitted by least squares (LS) and by least absolute deviations (L1). For brevity, we only show the results for the para modification.

The residuals show a very long-tailed distribution, even more so for the L1 than for the LS fit (Figure 1).

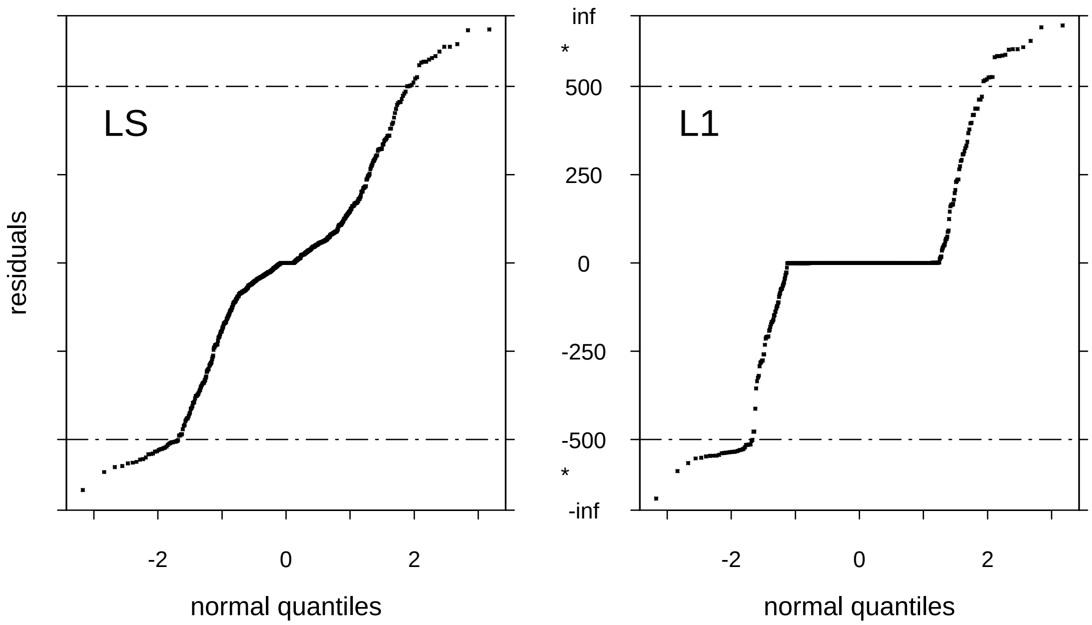

Figure 1: Normal plots of residuals from an LS and an L1 fit of model (1). * A nonlinear scale is used for $|r| > 500$.

Note. In this and most of the remaining figures, a special mode of plotting is used. Since the outliers would dominate the visual impression, the range of the plot in the vertical direction is restricted to $[a, b]$, with $-a = b = 500$ in most cases. Rather than clipping points outside this range, they are plotted in a nonlinear scale in the top and bottom margins. More precisely, the y coordinate is given by $\tilde{y} = y$ for $a \leq y \leq b$, by $\tilde{y} = b + (c - b)/[1 + (b - a)/(y - b)]$ for $y > b$, and analogously for $y < a$.

The observer should keep in mind that the L1 fitting method always produces as many zero residuals as there are parameters in the model, 252 in our case. There are 53 exact zeros even for the LS fit. They are mostly caused by states for which only one observed transition is available, which thus is fit exactly by the estimated term value associated with the state. For the least squares case, it would be more adequate to plot standardized residuals, dropping the afore mentioned cases for which they are indefinite. Adequate standardizations for L1 are not available (for approximations, see McKean et al., 1991). As an alternative, we simulated three sets of standard normal residuals and used them as the energy values $Y_{\ell u}$ together with the (ℓ, u) pairs of our dataset to apply LS and L1 fitting as before. For a graphical comparison (Figure 2), the resulting residuals were scaled to have the same median deviation as those obtained from the actual data. The right hand side of the figure also shows a more detailed view of the flat part in the Q-Q-plot of L1 residuals.

The distribution of residuals is clearly long-tailed, even more so in the L1 case. The median absolute value of the actual residuals was 80 for the LS and 0.020 for the L1 fit. The latter corresponds to the known scale of measurement errors.

Continuing with the usual analysis of residuals, we next examine the plot of residuals against the fit. Figure 3 shows a highly structured pattern for the L1 fit, whereas least squares produces a rather diffuse picture. The pattern suggests a classification into "good" and "bad" observations. However, this identification might flag the wrong values. For 49 states, there are only two transitions from or to the state. The L1 method usually does not have a unique solution in such cases. (It has a unique solution if one of the transitions leads to a state with no other transition and can therefore be fit exactly by adjusting the corresponding term value. More complicated cases can also be constructed.) Specific algorithms can fit either one of the two transition energies exactly in such a non-unique case, or give a term value

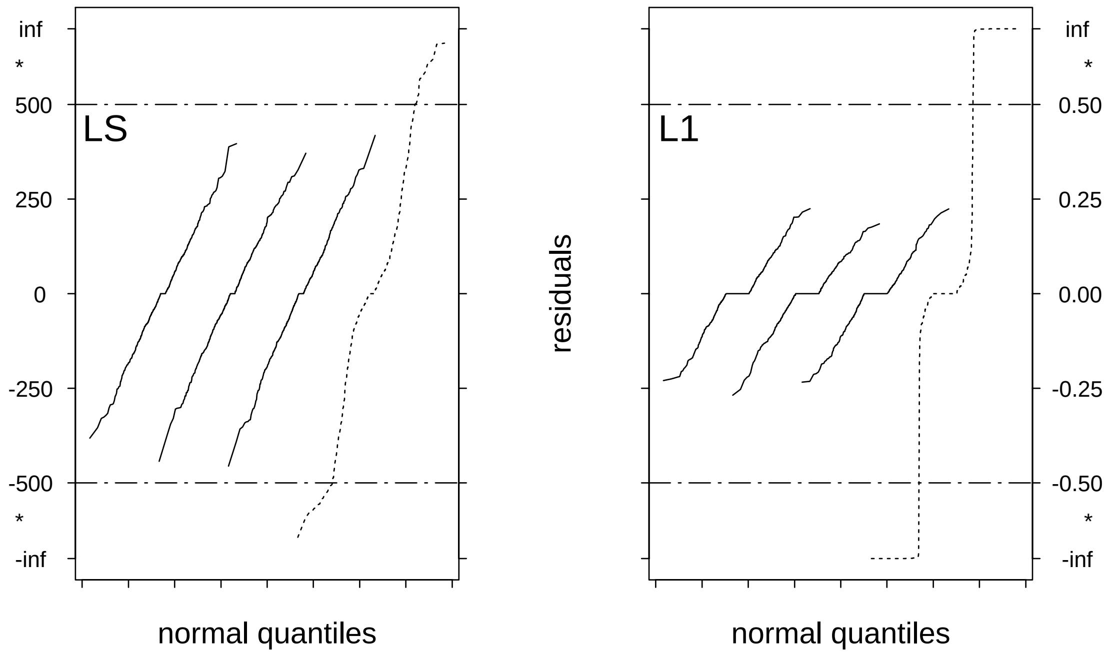

Figure 2: Normal plot for 3 simulated sets of residuals (solid lines), compared to those obtained for actual data (dotted line). The normal quantiles are shifted by multiples of 2 in order to separate the samples. * A nonlinear scale is used for $|r| > 500$ in the LS and for $|r| > 0.5$ in the L1 case.

in between the two corresponding estimates.

Going one step further, outliers can reinforce each other's effect on any fit if they occur in special combinations of design points, to the extent that they are well fit, while adequate observations will appear to be gross errors. This problem is a version of local breakdown, see, e. g., Hampel (1991).

More insight into the special design structure is obtained by displaying the transitions graphically. States can be ordered with respect to their electronic state and their vibration and rotation quantum number as primary, secondary and tertiary sorting criteria. This order was used to plot the upper state against the lower state for all 662 para transitions in Figure 4. The diagonal stripes which appear clearly in the diagram reflect the selection rules (Section 2). Since there can be only one energy value for a certain transition, the few positions carrying two symbols stem from two trial identifications, of which (at least) one is erroneous. The columns (lower states) or rows (upper state) which contain only 1 or 2 symbols lead to poorly determined

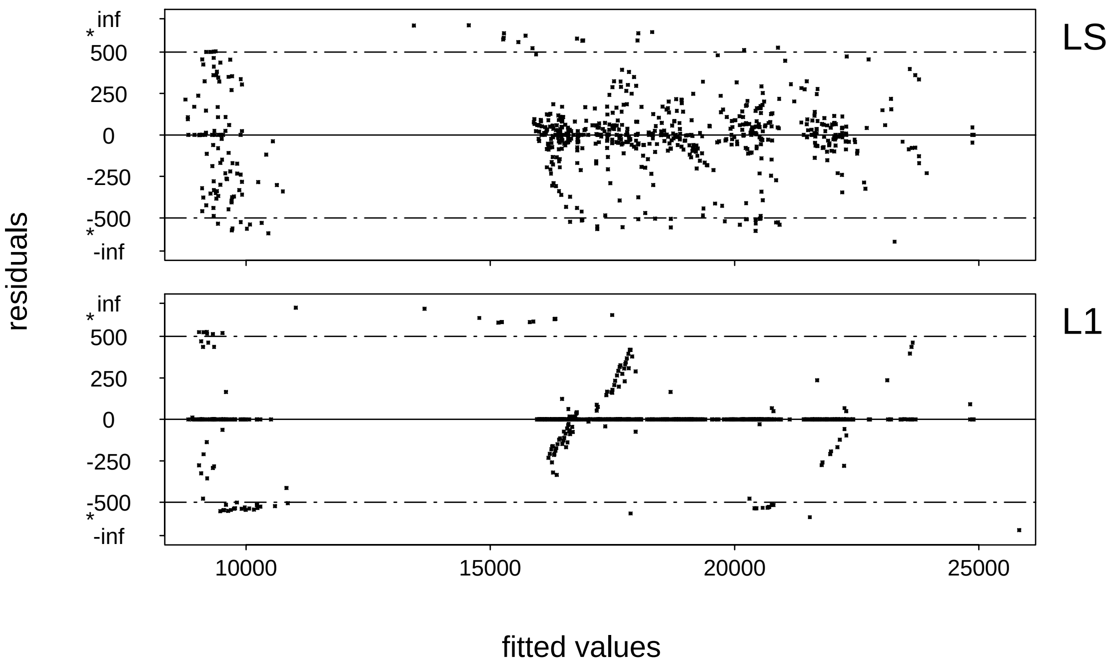

Figure 3: Residuals versus fit. * A nonlinear scale is used for $|r| > 500$.

term values.

A glance at the relative sizes of residuals shown by the size of the symbols in Figure 4 reveals that electronic states $3b$ and $4c^-$ have been identified less successfully than others, and for $5c^-$ there are very few transitions. In order to obtain more precise term values for the rest, all transitions from these states were dropped and model (1) was fitted again. As a kind of internal check, the differences δ_j between estimated term values for the full and the reduced data set were calculated (as far as the values existed for the latter case). The median deviations of these differences δ_j(defined as $\text{MAD}_j(\delta_j) = \text{med}_j[\delta_j - \text{med}_k(\delta_k)]$) were 147 and 0.022 for the LS and L1 cases, respectively. (The standard deviations were 307 and 50.9.) Clearly, the L1 values were more stable.

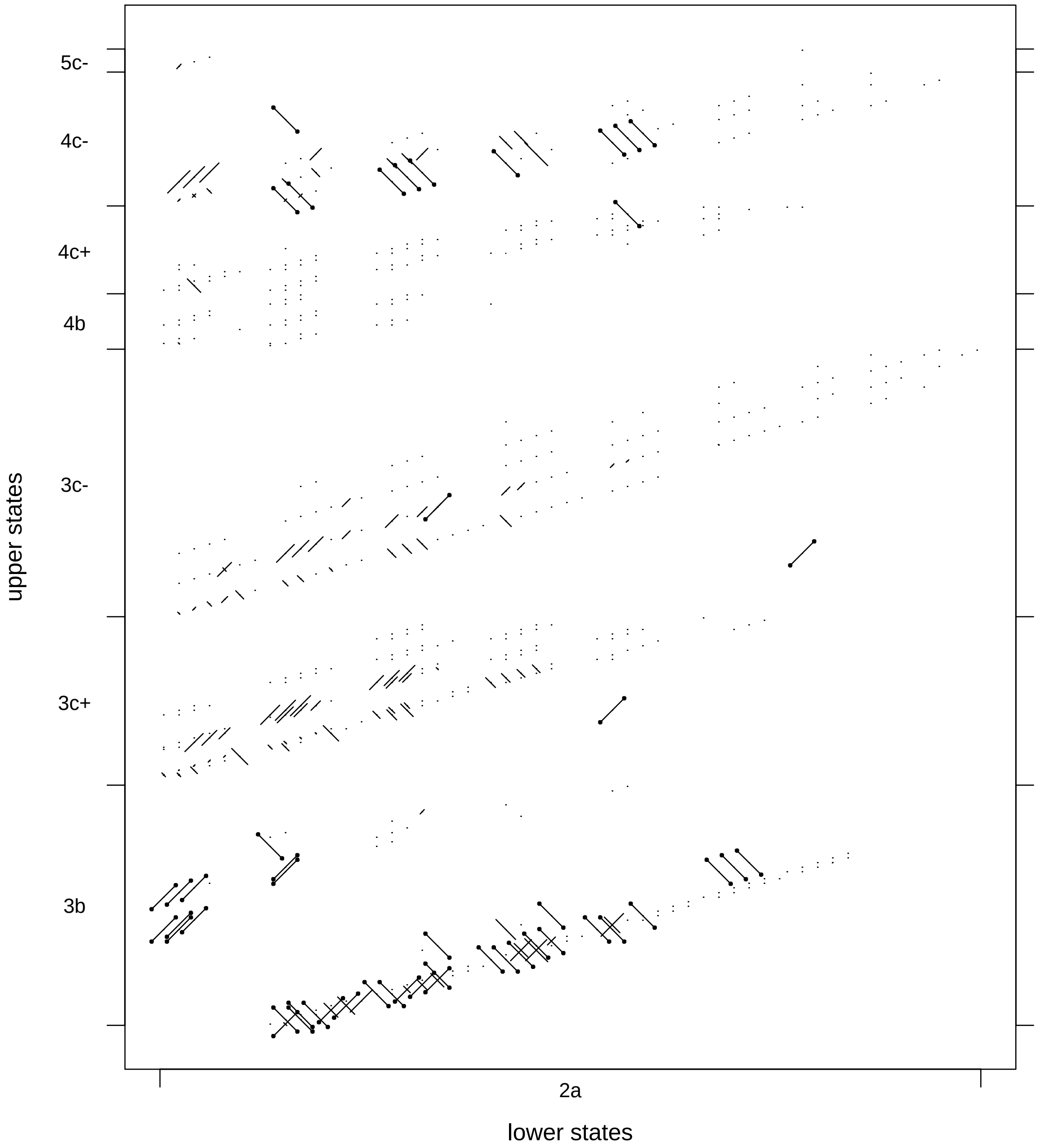

Figure 4: Upper and lower states (sorted according to e and the quantum numbers v and j) for the observed transitions. The size of the symbol is proportional to the absolute value of the residual, and its slope shows the sign. Size is truncated at 500 and truncation is marked by dots at both ends of the line.

4 Fitting the Model with Spectroscopic Constants

Model (4) with polynomial order $m = 2$ was fitted by LS and L1, using both para and ortho transitions. The LS residuals showed more clearly a very long-tailed distribution (Figure 5) than in the previous case. Again, the median absolute L1 residual (0.0728) was smaller by orders of magnitude than the LS one (140) and was of a similar order as the known size of measurement errors.

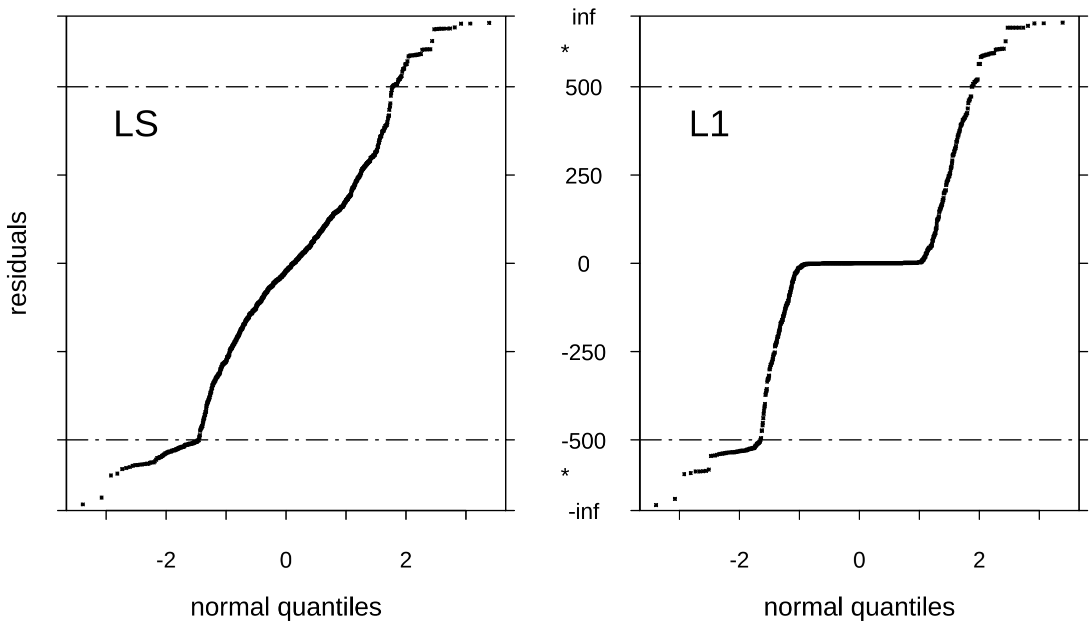

Figure 5: Normal plots of residuals from an LS and an L1 fit of model (4)

Term values $\tilde{\theta}_j$ can be calculated from the fitted model by (3) and compared with the $\hat{\theta}_j$ obtained in the previous section. Figure 6 shows a comparison of the two estimates for the "lower" electronic state 2a. Again, the L1 results show a much more satisfactory agreement of estimated term values with the theoretical model (3). Nevertheless, some systematic deviations show up for higher values of j. (Similar pictures were obtained for the upper states.)

For a quantitative comparison, the median deviations of diffferences of the two estimates were calculated and turned out to be 278 for LS and 0.96 for L1.

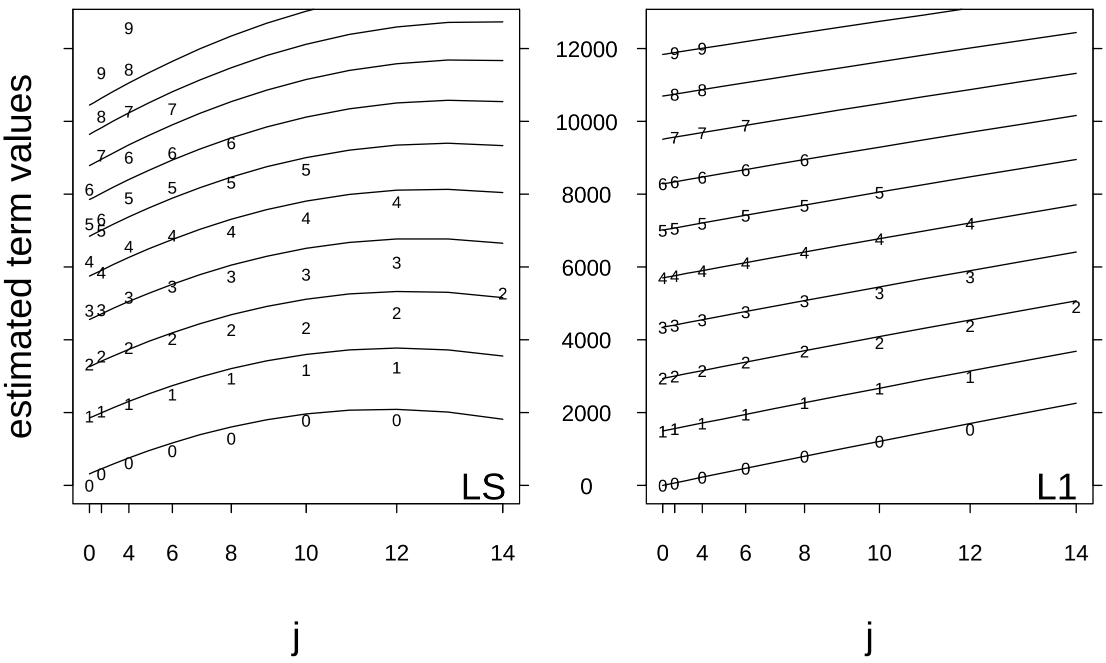

Figure 6: Term values estimated directly from model (1), $\hat{\theta}_\ell$ (marked with quantum number v_ℓ) and term values $\tilde{\theta}$ obtained from (3) with estimated coefficients (connected by lines) for 2a, plotted versus $j_\ell(j_\ell+1)$.

5 Fine Tuning

Although the L1 fit gives a satisfactory overall picture, it is desirable to improve it by a more sophisticated analysis. A specific reason relates to the subsequent use of estimated term values in the specific context. If very precise term values are found, they can be used to obtain identifications for further observed energy values. Consider the set of all states for which estimated term values exist. The simplest way to obtain new identifications consists in determining all transitions between these states which are possible according to the selection rules, and select those which do not appear in the data set of observed energy values which have already been identified. For this selection, estimated transition energies can be calculated, and the nearest observed energies can be found. This procedure only extends the set of identified observed energies, whereas the extrapolations discussed below also extend the set of states with known energy levels.

Clearly, some term values can be estimated more reliably than others on the basis of the identified transitions. Estimates are less reliable with states for which only one or two transitions are observed, and if they are fit badly by model (3). These cases can be found on the basis of the analyses discussed before. In order to improve the final term value estimates, the observations involving a discrepant estimated term value or leading to an excessive residual are eliminated, and the estimates are re-calculated in the reduced data set.

Similar results can be obtained more directly by using so-called redescending robust estimates, obtained by solving (for θ)

$$\sum_i \psi([y_i - x_i^T\theta]/\sigma)x_i = 0 \tag{5}$$

where ψ is, for example, of the "three-part redescending" type,

$$\psi(r) = \mathrm{sign}(r)\ \min\left(|r|, a, \max\left(\frac{d-r}{d-b}, 0\right)\right)$$

with $0 < a \leq b < d < \infty$, x_i^T is the row of X in (2), corresponding to the energy value y_i, and σ is an estimate of the scale parameter of the error term (see, for example, Hampel, 1986, Chapter 6 and Section 2.6a). The idea behind redescending estimates is that they can effectively "reject" distant outliers automatically in the sense that exactly the same result is obtained if clear outliers are dropped as if they are included in the data set. On the other hand, the solution of (5) is often not unique, and the estimate is well defined only if an iterative algorithm and a starting θ vector for it are given. Loosely speaking, the algorithm will find the solution of (5) which is "nearest" to the initial vector.

In the present problem, a reweighting algorithm can be used, starting from the term values obtained by robustly fitting model (4). The iterative reweighting algorithm consists in obtaining the weighted least squares estimate given by weights $w([y_i - x_i^T\theta]/\sigma)$ with $w(r) = \psi(r)/r$. A straightforward application of this scheme will fail if there is a state for which all transitions get zero weight since they show a large residual in the L1 fit of (4). Earlier in this section, the term value estimates of such states were called unreliable, and it was decided that they should be dropped. If the algorithm is modified accordingly, then it produces the refined results mentioned before directly, without the need to explicitly drop observations from the data set.

It is plausible that the term values estimated by a redescending estimator in this way are more precise than those obtained by fitting (1) by L1. Since "the truth" is not known, we calculated the "internal" measures of stability as before. The measure comparing estimated term values $\widehat{\theta}_j$ with those calculated from (3), $\widetilde{\theta}_j$, using L1 estimates for the coefficients was $\mathrm{MAD}_j(\widehat{\theta}_j - \widetilde{\theta}_j) = 0.70$. This compares favorably to the measure obtained when using L1 also for fitting term values $\widehat{\theta}_j$, which resulted in 0.96 (as mentioned in Section 4). Note, however, that this comparison is somewhat unfair since the refined algorithm is explicitly designed to find estimates which differ little from the $\tilde{\theta}_j$ values. A second measure of stability is obtained by comparing results between the full data set and the subset described in Section 3. The median deviations of these differences of estimated term values were 0.015 for the refined algorithm, as compared to 0.022 using direct L1 fitting for the same term values. A similar comment as above applies to this comparison. Nevertheless, the refined algorithm appears to be superior to straight L1 fitting according to these stability measures.

As mentioned before, assignments of spectral lines to additional states which are not contained in the previous identifications can be obtained from extrapolations based on (3). However, due to the limited accuracy of that approximation, there are systematic errors in such extrapolations. Figure 7 shows the differences between term values $\widehat{\theta}_\ell$ estimated directly as just described and term values $(\widetilde{\theta}_\ell)$ calculated from (3) with coefficients estimated as in Section 3, for the lower electronic state $2a$, plotted versus the state in the usual order. The clear appearent structure calls for higher terms $(m > 2)$ in the expansion (3).

There is, however, the usual trade-off between the higher flexibility obtained by including higher order terms and the additional statistical variability introduced by their estimation. Furthermore, the estimated model should yield reliable extrapolations to higher values of j and v when searching for further identifications of observed transition energies. Higher order terms of polynomials often lead to deteriorated extrapolations even in the vicinity of the data used to estimate the model.

This problem is aggravated by straightforward robust estimation. The latter allows for a poor fit for a minority of data points by definition. If the class of approximating functions is not quite adequate, this minority will tend to include the

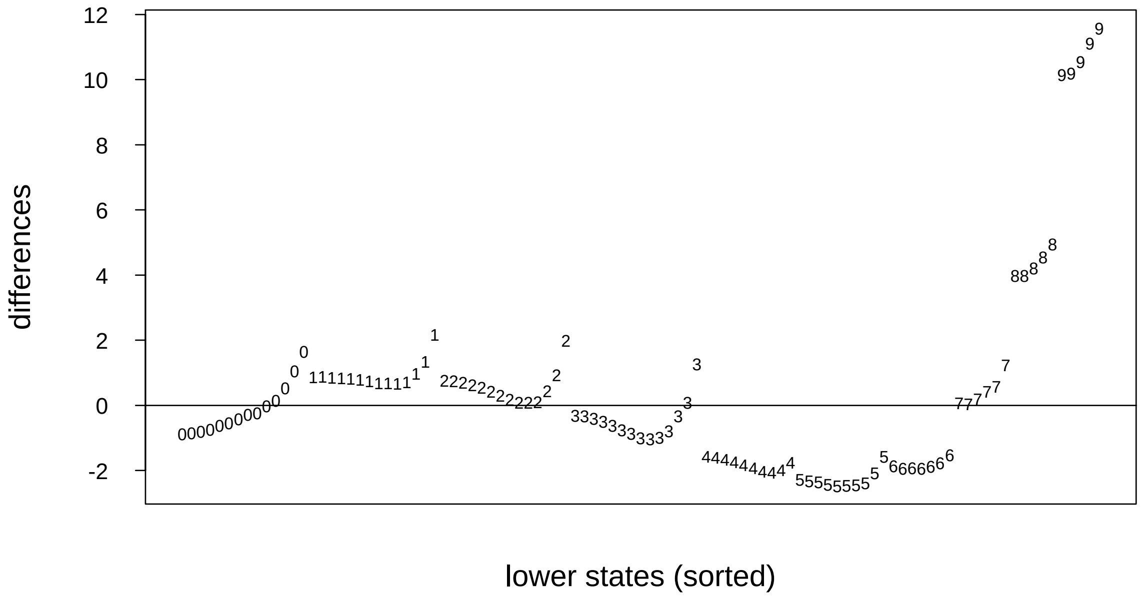

Figure 7: Difference between estimated term values ($\widehat{\theta}_\ell$) and term values ($\widetilde{\theta}_\ell$) obtained from estimated spectroscopic constants versus sorted (according to the quantum numbers) states marked with quantum number v_ℓ for the lower electronic state 2a.

outskirt of the range of design points (extreme j and v values), and extrapolation will be inappropriate. The problem is a fundamental one. The polynomial model is meant to serve two purposes: approximation and smoothing. Whereas the latter intends to eliminate the random

measurement errors, the former should keep the systematic error uniformly low over the range of validity of the approximation. In approximation theory, there is therefore a tendency to minimize the maximal residual (the supremum norm L_∞) – the opposite of robust procedures. It is not difficult to suggest ad-hoc procedures to overcome this difficulty. More research is needed to obtain satisfactory results, however.

6 Conclusions

The analyses presented in this paper show clearly that a robust fitting method produces results which are in agreement with the established theory and remain stable when parts of the data set are dropped, whereas traditional least squares methodology leads to confusing results. The number of gross errors is quite high in our data set, but this is very natural in the exploratory phase when identifications of states are still tentative.

One could ask if the LS results could have been improved by eliminating outliers, estimating again, and iterating these steps until no outlier can be detected any more. The answer is affirmative since such a procedure constitutes an algorithm for calculating a robust estimator. A second question is how the fitting methods would compare if only correctly identified observations were present. This could be anwered if the error distribution was known. If this was the case (for the "clean" or the "contaminated" data) one could also propose using the corresponding maximum likelihood estimator instead of L1 or LS. The error distribution can be estimated by the (possibly smoothed) distribution of residuals. Because of the small number of observations per parameter, it is unclear how well this estimation step would work, and if the asymptotically optimal maximum likelihood estimator would be nearly optimal for such a finite sample case. Further research is needed in these problems.

References

Crosswhite, H. M. (ed.) (1972). *The Hydrogen Molecule Wavelength Tables of Gerhard Heinrich Dieke.* Wiley, N.Y.

Dieke, G. H. (1958). The Molecular Spectrum of Hydrogen and Its Isotopes. J. Molec. Spectrosc. **2**, 494-517.

Dieke, G. H., and Cunningham, S. P. (1965). Bands of D_2 and T_2 Originating from the Lowest Exited $^1\Sigma_g$ States $(1s\sigma)$ $(2s\sigma)$ $^1\Sigma_g$ and $(2p\sigma)^2$ $^1\Sigma_g$. J. Molec. Spectrosc. **18**, 288-320.

Dieke, G. H., and Tomkins, F. S. (1951). The $3p\ {}^3\Sigma \to 2s\ {}^3\Sigma$ -Bands of TH and T_2. Phys. Rev. **82**, 796-807.

Freund, R. S., Schiavone, J. A., and Crosswhite, H. M. (1985). The Electronic Spectrum and Energy Levels of the Deuterium Molecule. J. Phys. Chem. Ref. Data **14**, 235-383.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

Hampel, F. R. (1991). Some Mixed Questions and Comments About Robustness. *In:* Stahel, W. A. and Weisberg, S. (eds.) . *Directions in Robust Statistics and Diagnostics.* Springer, New York.

Herzberg, G. (1950). *Molecular Spectra and Molecular Structure. I. Spectra of Diatomic Molecules.* Van Nostrand Reinhold, New York.

McKean, J. W., Sheather, S. J., and Hettmansperger, T. P. (1991). Standardizing Residuals from Robust Regression. Manuscript.

Watson, J. K. G. (1989). On the Use of Term Values in the Least-Squares Fitting of Spectra. J. Molec. Spectrosc. **138**, 302-308.