Exploring the limits of cryospectroscopy : least-squares based approaches for analyzing the self-association of HCl

Exploring the limits of cryospectroscopy: least-squares based approaches for analyzing the self-association of HCl

Liene I. De Beuckeleer, Wouter A. Herrebout*

Department of Chemistry, University of Antwerp, Groenenborgerlaan 171, 2020 Antwerp, Belgium

Keywords

hydrogen chloride, cryosolutions, self-association, infrared spectroscopy, least-squares fitting, information criteria

Abstract

To rationalize the concentration dependent behavior observed for a large spectral dataset of HCl recorded in liquid argon, least-squares based numerical methods are developed and validated. In these methods, for each wavenumber a polynomial is used to mimic the relation between monomer concentrations and measured absorbances. Least-squares fitting of higher degree polynomials tends to overfit and thus lead to compensation effects where a contribution due to one species is compensated for by a negative contribution of another. The compensation effects are corrected for by carefully analyzing, using *AIC* and *BIC* information criteria, the differences observed between consecutive fittings when the degree of the polynomial model is systematically increased, and by introducing constraints prohibiting negative absorbances to occur for the monomer or for one of the oligomers. The method developed should allow other, more complicated self-associating systems to be analyzed with a much higher accuracy than before.

* wouter.herrebout@uantwerpen.be

1. INTRODUCTION

Hydrogen chloride is a textbook example of a simple hydrogen-bonding molecule. The self-association of HCl into molecular clusters is used as a model in the study of intermolecular interactions. Using different spectroscopic techniques, HCl dimers up to hexamers have been studied in the gas phase through jet-cooled spectroscopy[1-5], in helium nanodroplets [6] and in solid matrices[7, 8]. In 1992, van der Veken and De Munck [9] reported a systematic study of HCl dissolved in liquefied noble gases and this work was expanded in 2001 by Herrebout, Van Gils and van der Veken [10]. Inspection of the data available showed that at sufficiently low concentrations, the spectrum is dominated by contributions of the monomeric species, while at higher concentrations, additional bands due to oligomeric species of HCl can be observed at wavenumbers well below that of the monomer.

To be able to rationalize the features observed, in the original paper [9] the phenomenon was firstly analyzed by subjecting it to a factor analysis [11], but because of unsatisfying results this methodology was abandoned and a band profile analysis was performed instead. In this method, the contributions of the monomer and the oligomeric species were approximated by subtracting the spectra of solutions containing larger amounts of HCl and a rescaled spectrum of a highly diluted solution recorded under similar circumstances. Subsequently, the difference spectra were least-squares fitted using a series of Gauss-Lorentz sum profiles. The analysis of the data obtained in isothermal concentration studies allowed assigning the different band features to dimers, trimers, and tetramers. The study of the temperature behavior yielded approximate values for the enthalpies of complexation in the cryosolutions [9, 12].

Based on the results obtained during ongoing research projects involving, amongst others, the studies of more complex systems such as C-H$\cdots$Y hydrogen [13-21], C-X$\cdots$Y halogen [13, 18, 22-28] and lone pair$\cdots\pi$ [29] interactions, and based on new technological developments made in recent years, we started to realize that the analysis performed [9, 12] had several drawbacks and could thus be improved. These ideas originated from the current availability of liquid cells with a smaller optical path allowing more concentrated solutions to be studied without saturating the detector used, and the integration of new, proportional-integral-derivative (PID) controlled setups allowing spectra to be recorded with a much higher temperature stability than that used in the original studies [9]. Apart from these technological developments, we were also triggered by the observation that results of least-squares band

fitting procedures depend strongly on the initial parameters chosen and thus can be severely biased by the end-user.

In this paper, we report on the development of more robust numerical methods in which the concentration dependent behavior observed for a large spectral dataset recorded a constant temperature is scrutinized. The new method allows the contributions due to monomers and due to different types of self-association to be separated directly, thereby avoiding the requirement that at the lowest concentrations studied the contribution of complex species should be negligibly small. It will be shown that due to overfitting simpler approaches such as regular least-squares fitting of absorbances versus monomer concentrations fail to accurately determine the different contributions. These drawbacks are corrected for by carefully analyzing, using the statistically acknowledged selection *AIC* and *BIC* criteria, the differences between consecutive fittings when the degree of the polynomial is systematically increased, and by introducing constraints prohibiting negative absorbances to occur for the monomer or for one of the complexes. The models and approaches developed, and the Matlab based software packages used for their implementation should allow other, more complex systems [30, 31] to be analyzed with a much higher accuracy than before, thereby avoiding the bias originating from the empirical selection of the initial parameters to be used in traditional least-squares band profile analyses.

2. EXPERIMENTAL SECTION

HCl (99%) was purchased from Sigma-Aldrich and was used without further purification. The argon used as a cryosolvent had a stated purity of 99.9995% and was supplied by Air Liquide.

Infrared spectra were recorded on a Bruker IFS 66v Fourier transform spectrometer. For the mid-infrared spectra, a Globar source was used in combination with a Ge/KBr beamsplitter and a LN_2-cooled broad band MCT detector. All interferograms were averaged over 500 scans, Blackman-Harris 3-term apodized and Fourier transformed with a zero filling factor of 4 to yield spectra with a resolution of 0.5 cm^{-1}. The experimental set-up used to investigate the solutions in liquid noble gases has been described before [32]. In the actual cryostat, a liquid cell with 1 cm path length and equipped with wedged Si windows was mounted below a LN_2 Dewar. The temperature of the cell body is measured using a Pt-100 thermoresistor. The SunRod electric minicartridge heater is controlled using a Eurotherm 3504 PID controller. The temperature variation during a typical experiment is less than 0.05 K.

Spectra were obtained and pre-analyzed using OPUS 6.5. Further calculations were performed using Matlab [33].

3. RESULTS AND DISCUSSION

In the following paragraphs, the general methodology used in this study will be described. Subsequently, results obtained using a fixed-degree polynomial approximation, and results based on the A and B information criteria used to select the appropriate polynomial degree and to avoid overfitting and/or negative absorbances are discussed in detail.

3.1 General Concept

The general concept of the method used is based on the fact that, with some exceptions [34, 35], cryosolutions are known to be in thermodynamical equilibrium. The spectra for cryosolutions of self-associating species therefore are a superposition of monomer spectra and spectra of the different complexes, with the relative absorbances determined by the equilibrium concentrations C and the molar attenuation coefficients ε of the monomeric species and of the associations formed. The latter, are determined by the equilibrium constants K involved.

2 monomer ⇌ dimer $$K_2 = \frac{C_{dimer}}{C_{monomer}^{\ 2}} = \frac{C_{di}}{C_{mono}^{\ 2}} \quad (1)$$

3 monomer ⇌ trimer $$K_3 = \frac{C_{trimer}}{C_{monomer}^{\ 3}} = \frac{C_{tri}}{C_{mono}^{\ 3}} \quad (2)$$

4 monomer ⇌ tetramer $$K_4 = \frac{C_{tetramer}}{C_{monomer}^{\ 4}} = \frac{C_{tetra}}{C_{mono}^{\ 4}} \quad (3)$$

Starting from these assumptions, each arbitrary wavenumber $\tilde{\nu}_i$ the measured absorbance A_{exp} can be written as a sum of contributions;

$$A_{exp}(\tilde{\nu}_i) = A_{mono}(\tilde{\nu}_i) + A_{di}(\tilde{\nu}_i) + A_{tri}(\tilde{\nu}_i) + A_{tetra}(\tilde{\nu}_i) \quad (4)$$

Depending on the wavenumber chosen and on the spectral features of the species present, the actual form of the expression and the number of contributions can differ, typical possibilities being

$$A_{exp}(\tilde{\nu}_i) = A_{mono}(\tilde{\nu}_i) \quad (5)$$

for a monomer contribution only,

$$A_{exp}(\tilde{\nu}_i) = A_{mono}(\tilde{\nu}_i) + A_{di}(\tilde{\nu}_i) \tag{6}$$

for a monomer and dimer contribution,

$$A_{exp}(\tilde{\nu}_i) = A_{mono}(\tilde{\nu}_i) + A_{di}(\tilde{\nu}_i) + A_{tri}(\tilde{\nu}_i) \tag{7}$$

for a monomer, dimer and trimer contribution, and eq (4) for a monomer, dimer, trimer and tetramer contribution.

By choosing an appropriate wavenumber $\tilde{\nu}_m$ for which the absorbance is due to monomers only, i.e.

$$A_{exp}(\tilde{\nu}_m) = A_{mono}(\tilde{\nu}_m) \tag{8}$$

$$\text{and } A_{di}(\tilde{\nu}_m) = A_{tri}(\tilde{\nu}_m) = A_{tetra}(\tilde{\nu}_m) = 0$$

and by using Lambert-Beer's law

$$A_j(\tilde{\nu}_i) = \varepsilon_j(\tilde{\nu}_i) C_j d \tag{9}$$

with ε the molar attenuation coefficient, C the concentration in the solution, j the type of species and d the path length of the cell used, the different contributions in equation 4 can be rewritten in terms of the absorbance of the monomer wavenumber $\tilde{\nu}_m$;

$$A_{mono}(\tilde{\nu}_i) = a_1(\tilde{\nu}_i, \tilde{\nu}_m) A_{mono}(\tilde{\nu}_m) \tag{10}$$

$$A_{di}(\tilde{\nu}_i) = a_2(\tilde{\nu}_i, \tilde{\nu}_m) A_{mono}(\tilde{\nu}_m)^2 \tag{11}$$

$$A_{tri}(\tilde{\nu}_i) = a_3(\tilde{\nu}_i, \tilde{\nu}_m) A_{mono}(\tilde{\nu}_m)^3 \tag{12}$$

$$A_{tetra}(\tilde{\nu}_i) = a_4(\tilde{\nu}_i, \tilde{\nu}_m) A_{mono}(\tilde{\nu}_m)^4 \tag{13}$$

The coefficients used in these expressions are defined as

$$a_1(\tilde{\nu}_i, \tilde{\nu}_m) = \frac{A_{mono}(\tilde{\nu}_i)}{A_{mono}(\tilde{\nu}_m)} = \frac{\varepsilon_{mono}(\tilde{\nu}_i)}{\varepsilon_{mono}(\tilde{\nu}_m)} \tag{14}$$

$$a_2(\tilde{\nu}_i, \tilde{\nu}_m) = \frac{A_{di}(\tilde{\nu}_i)}{A_{mono}(\tilde{\nu}_m)^2} = a_1(\tilde{\nu}_i, \tilde{\nu}_m)^2 \frac{\varepsilon_{di}(\tilde{\nu}_i)}{\varepsilon_{mono}(\tilde{\nu}_m)^2} \frac{1}{d} K_2 \tag{15}$$

$$a_3(\tilde{\nu}_i, \tilde{\nu}_m) = \frac{A_{tri}(\tilde{\nu}_i)}{A_{mono}(\tilde{\nu}_m)^3} = a_1(\tilde{\nu}_i, \tilde{\nu}_m)^3 \frac{\varepsilon_{tri}(\tilde{\nu}_i)}{\varepsilon_{mono}(\tilde{\nu}_m)^3} \frac{1}{d^2} K_3 \tag{16}$$

$$a_4(\tilde{\nu}_i, \tilde{\nu}_m) = \frac{A_{tetra}(\tilde{\nu}_i)}{A_{mono}(\tilde{\nu}_m)^4} = a_1(\tilde{\nu}_i, \tilde{\nu}_m)^4 \frac{\varepsilon_{tetra}(\tilde{\nu}_i)}{\varepsilon_{mono}(\tilde{\nu}_m)^4} \frac{1}{d^3} K_4 \tag{17}$$

Substituting the above equations 12 to 15 in equation 4 results in

$$A_{exp}(\tilde{\nu}_i) = a_1(\tilde{\nu}_i, \tilde{\nu}_m) A_{mono}(\tilde{\nu}_m) + a_2(\tilde{\nu}_i, \tilde{\nu}_m) A_{mono}(\tilde{\nu}_m)^2 + a_3(\tilde{\nu}_i, \tilde{\nu}_m) A_{mono}(\tilde{\nu}_m)^3 + a_4(\tilde{\nu}_i, \tilde{\nu}_m) A_{mono}(\tilde{\nu}_m)^4 \tag{18}$$

The above equation shows that for any arbitrary wavenumber, the contributions due to the different species present can in principle be derived by plotting the measured absorbances versus the monomer absorbance at the given reference wavenumber, and using least-squares to fit a polynomial to the resulting data.

It is of interest to note that in eqs. (15)-(18), the optical path length is introduced. To ensure that changes in the spectra are due to changes in the concentration only, it is therefore necessary that spectra belonging to one dataset have been recorded under identical circumstances using the same liquid cell. Data derived from different liquid cells or obtained using different set-ups should thus not be mixed during numerical analysis, unless corrections for the differences on path length are introduced explicitly.

3.2 Measurements and Baseline corrections

To assess the general principle of the method described above, and to further optimize the methods used, a dataset of 346 spectra of solutions in liquid argon was constructed. The mole fractions of HCl used for the dataset are difficult to accurately quantify [9, 12], but are estimated to vary between 5.0×10^{-3} and 2.0×10^{-5}. The temperature of the solutions was stabilized at 103 K, the temperature variation during a typical run being less than 0.05 K. These settings were chosen to allow a large fraction of HCl molecules to be involved in self-association and at the same time avoid additional features assigned to solid HCl particles suspended in the solution[36] to appear in the region 2750-2700 cm^{-1}. The concentrations used are chosen so that the region between minimum and maximum absorbance is uniformly covered.

As the outcome of the fitting procedures can largely depend on baseline artefacts, during all experiments, baseline corrections were performed using spectra of pure liquid argon recorded at exactly the same conditions. Moreover, special attention was paid to the removal of remaining water traces caused by small water particles suspended in the solution, or condensation of water vapors onto the cold windows of the cryostat and the detector.

To account for remaining drifts of the baseline, which we believe are caused by small changes in the temperature inside the spectrometer due to the cold cryostat present, an additional straight line baseline correction was applied to all data, the spectral limits used to define the background being 3100 and 2500 cm^{-1}.

3.3 Subtraction Procedures and Least-Squares Band Profile Analysis

As the results obtained from the newly developed numerical algorithms will be compared and rationalized with the help of the data reported earlier, it is of interest to summarize the most important data derived from the subtraction and the least-squares band profile analysis. Because the use of factor analysis has proven to be unsatisfying in the original paper [9], this method was not pursued in this study. However, recent insights on the use of factor analysis could lead to complementary results [37, 38].

In Figure 1A, the infrared spectrum of HCl in liquid argon recorded at 103 K using a solution with mole fraction of 5.0×10^{-3} is compared to the rescaled spectrum of a diluted solution recorded under the same conditions. In addition, the resulting spectrum of the oligomeric species, obtained by subtracting the upper and middle trace, is illustrated in the lower line. The result of the least-squares band profile analysis in which the complex spectrum is fitted using a sum of Gauss/Lorentz sum profiles, and the calculated contributions due to dimer, trimer and tetramer, obtained by summing the respective Gauss/Lorentz sum profiles, are given in Figure 1B. As the outcome of the least-squares band profile analysis is often biased by the choice of the initial parameters, in the current study these parameters were optimized to achieve the best agreement with the outcome of the original studies [9, 12]. The characteristic wavenumbers derived for the dimer are 2855, 2847 and 2830 cm^{-1}, while those for the trimer and tetramer are 2807 and 2797 cm^{-1}, and 2778 and 2855 cm^{-1}, respectively. Whereas the individual contributions might be seriously affected by the choice of the initial parameters, the summed band profiles for dimer, trimer and tetramer are considered more reliable and can be used to assess the outcome of the newly developed methods.

We would like to explicitly point out that in the spectrum of the highly diluted solution used for the subtraction procedures, a weak feature illustrating the formation of dimers can be observed near 2830 cm^{-1}. The appearance of this feature necessarily leads to an underestimation of the dimer contribution in the subtracted spectrum. The data also illustrates that even for small concentrations, additional features do to oligomeric species are present and the general requirement of a monomer only contribution introduced in the analysis is difficult to achieve.

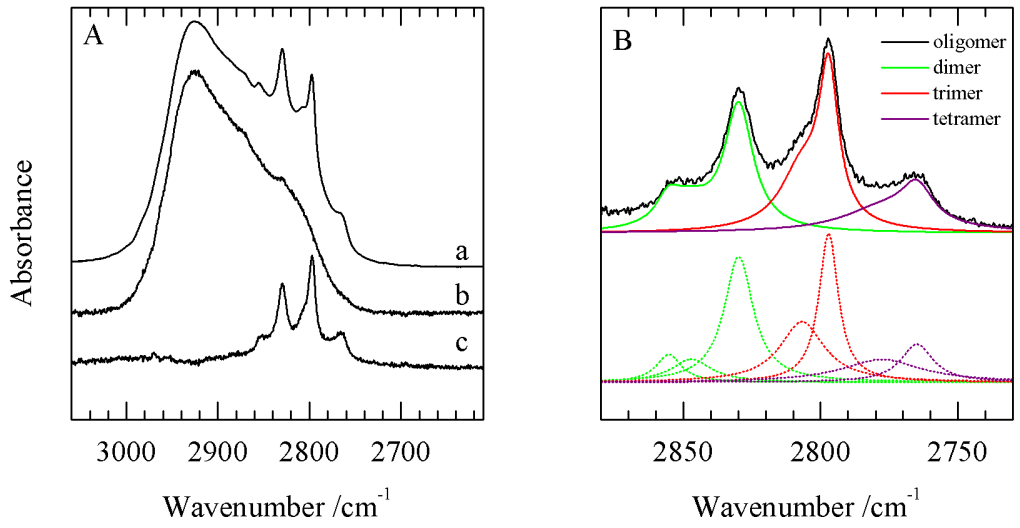

Fig. 1 (A) Subtraction procedure involving an original spectrum of a concentrated solution of HCl in liquid argon (a) and a rescaled spectrum of a highly diluted solution recorded under the same conditions (b). The result of the subtraction showing the summed contributions of the different oligomers present is given in the bottom trace c. (B) Results from a least-squares band profile analysis in which the spectrum of the complex species obtained in panel (A) is fitted using Gauss-Lorentz sum profiles (dotted line). The calculated contributions for the dimer, trimer and tetramer, obtained by summing the different Gauss/Lorentz sum profiles involved, are also given (solid line). As the outcome of the least-squares band profile analysis is often biased by the choice of the initial parameters, in the current study these parameters were optimized to achieve the best agreement with the outcome of the original studies [9, 12].

3.4 Least-squares fitting of polynomials

In Figure 2, typical spectra from the database are shown. Upon increasing the solute concentration, new bands due to self-association can easily be observed to emerge in the 2880-2740 cm^{-1} spectral region, i.e. at wavenumbers red shifted from the monomer wavenumber of 2869 cm^{-1}. The dashed line in Figure 2 represents the reference wavenumber $\tilde{\nu}_m$ used to determine the HCl monomer concentration. This wavenumber was chosen as internal standard because no features due to self-association are expected in this region and because its absorbance can be accurately determined for all concentrations studied. The solid lines A to H refer to the wavenumbers for which results obtained from least-squares fitted polynomials will be discussed.

In the following paragraphs the recorded dataset will be analyzed by fitting a polynomial through the measured absorbance values for every wavenumber individually. The data for the different wavenumbers thus are completely independent. In the final step, these contributions

for every wavenumber are combined and the isolated spectra of the species present in the solution are obtained without further smoothing. The noise observed therefore is a direct measure for the accuracy of the different contributions determined in the independent least-squares fits.

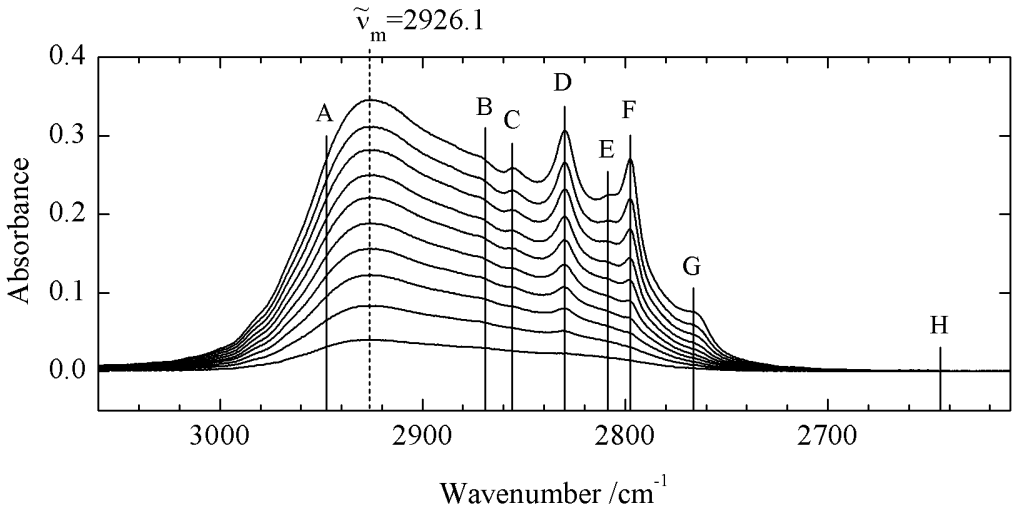

Fig. 2 Infrared spectra of solutions in liquid argon, at 103 K. From top to bottom, the mole fraction of HCl decreases from approximately 5.0 x 10^{-3} to 2.0 x 10^{-5}. The 10 spectra shown represent a small fraction of the spectral database used in the fitting procedures.

In Figure 3, results obtained by fitting, for each wavenumber, the experimental data with fixed 3rd, 4th and 5th degree polynomials are summarized. Panels a, b, and c refer to the results from a regular polynomial regression in which no constraints are introduced. Panels d, e, and f refer to the solutions from least-squares procedures in which nonnegative constraints are added to avoid negative intensities of the monomer or oligomer species.

The data in panel 3a, obtained by using a third polynomial for all wavenumbers to account for dimers and trimers, reveals a strong negative feature for the calculated dimer spectrum near 2765 cm^{-1}. This feature is observed at a wavenumber close to that previously assigned to the tetramer, and is believed to be caused by the underfitting of the data that do not allow contributions due to this species to appear. The artefact due to the underfitting is, at least partially, compensated for by additional intensity for the trimer.

The effect of adding an extra term allowing the contributions of the tetramer to be calculated is shown in panel 3b. It can be seen that, apart from some regions where little or no intensity is observed, negative intensities have largely disappeared. The drawback of adding an extra term, is a general increase in noise level. Also, for regions where no or little intensity

is observed e.g. for regions with an almost flat baseline, compensation effects are observed where negative contributions calculated for one or more species are balanced by positive contributions in the spectra of the other components.

The effect of increasing noise level and increasing compensation effects is further illustrated in panel 3c, where the results of a fitting procedure involving contributions due to dimer, trimer, tetramer and pentamer are shown. It is clear that under these conditions, overfitting of the experimental data leads to severe blurring of the interesting data present. As a result, it becomes more and more difficult to correctly distinguish the different oligomer contributions.

Comparison of the data in panels 3a and 3d shows that by introducing nonnegative constraints, the signal-to-noise ratio of the calculated spectra is significantly increased. Preventing the spectrum of the dimer to become negative near 2760 cm^{-1}, drastically reduces the overcompensation in the corresponding trimer spectrum. The reduction of the noise level and the underlying compensation effects upon the use of nonnegative constraints is also observed for the 4^{th} and 5^{th} degree polynomial fitting, shown in panels 4e and 4f, respectively. It can be seen that, in contrast to the results reported for the 3^{rd} degree polynomial fitting, the constrained fitting procedure hardly influences the calculated band profiles of dimer, trimer and tetramer. The result obtained in both approaches also resembles the data derived from the band fitting procedures using Gauss-Lorentz sum profiles, given in Figure 1B.

It should be noted that in panel 3f some contribution due to pentamer is also predicted to appear near 2780 cm^{-1}. We believe that the appearance of this feature is due to artefacts related to numerical instabilities of the procedures used. Additional evidence for this statement, illustrating that overfitting indeed limits the resolving power required to accurately separate the contributions of trimer and tetramer near 2780 cm^{-1}, will be reported below.

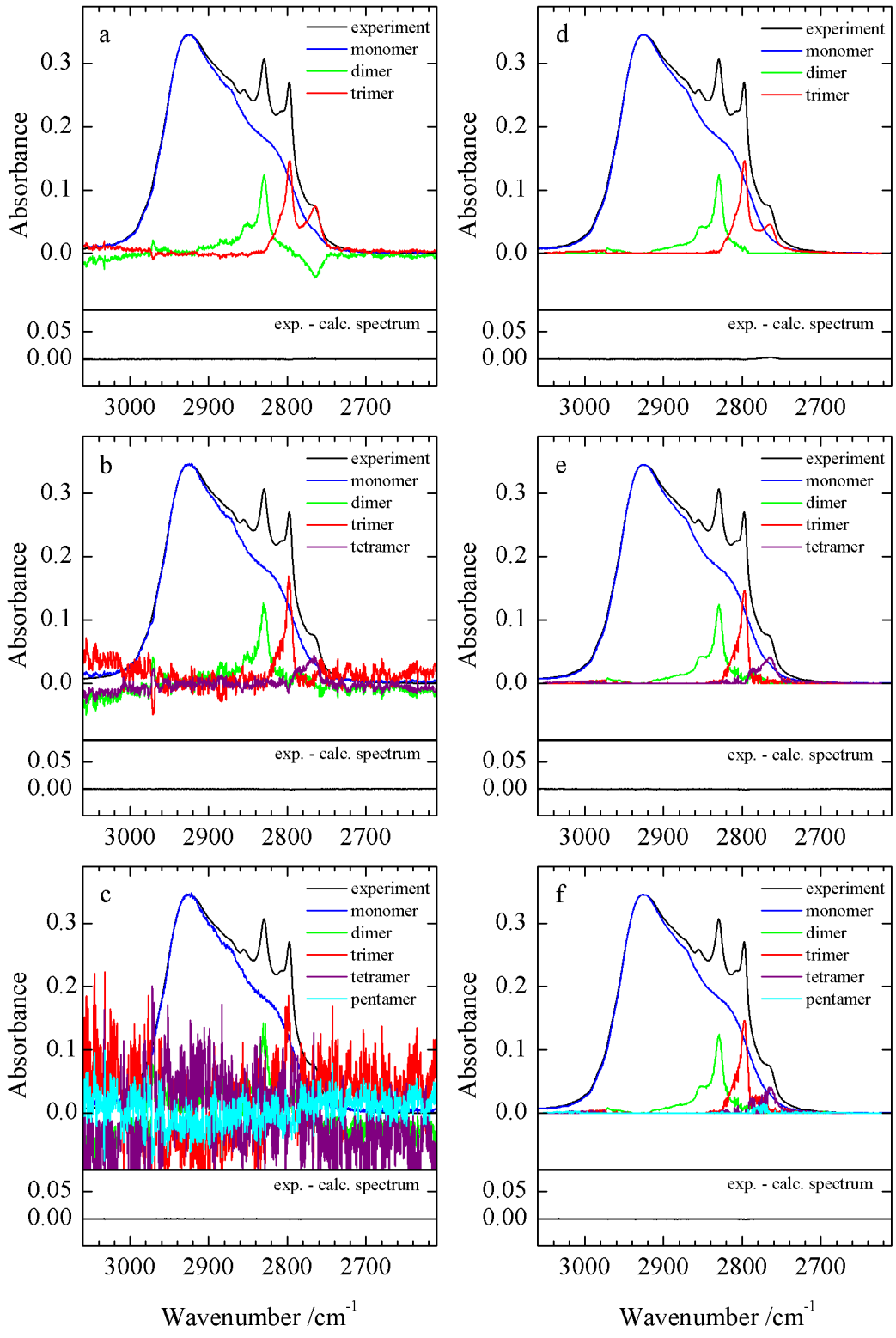

Fig. 3 The experimental data analyzed using only the 3rd degree polynomial (top), the 4th degree polynomial (middle) and the 5th degree polynomial (bottom) for all wavenumbers with regular least-squares (a,b,c) and adding a nonnegative constraint (d,e,f). The bottom panel shows the difference between the experimental spectrum and the model.

3.5 Polynomial selection

Besides using a fixed polynomial degree for all wavenumbers it is also possible to select a model for every wavenumber individually. The panels in Figure 3 show the experimental absorbances for A, B, C, D, E, F, G and H obtained from the database. These plots also show least-squares fitted polynomials with different degrees with the intercept equal to zero and with the polynomial degree p equal to 1, 2, 3, 4 or 5.

$$A_{exp}(\tilde{\nu}_i) = \sum_{p=1}^{n} a_{\mathrm{p}}(\tilde{\nu}_i, \tilde{\nu}_m)\,[A_{mon}(\tilde{\nu}_m)]^p \tag{19}$$

To facilitate analysis, the lower degree polynomial is always plotted in front of the higher degree polynomials so that changes due to the increase of the degree of the polynomial are readily recognized. This principle allows to visually observe when the addition of a higher-degree term has little or no extra contribution to the quality of the fit.

From the data in Figure 4, it can be seen that for case A the experimental data are well reproduced using a linear regression. As expected, the spectral features in this region therefore are fully reproduced using a monomer only approach. For cases B, C and D, a quadratic polynomial model describing monomer and dimer contributions is required, while for cases E and F, a cubic model incorporating monomer, dimer and trimer contributions is needed. For case G, good agreement requires a fourth degree term showing that for this wavenumber a tetramer contribution is needed. Finally, for wavenumber H, located far away from any absorbance, there is no change in the measured absorbance with increasing monomer concentration, therefore a 0^{th} degree polynomial or constant is the best model to mimic this relation. This visual observation of the appropriate degree can be calculated using statistical selection criteria explained in the following paragraphs.

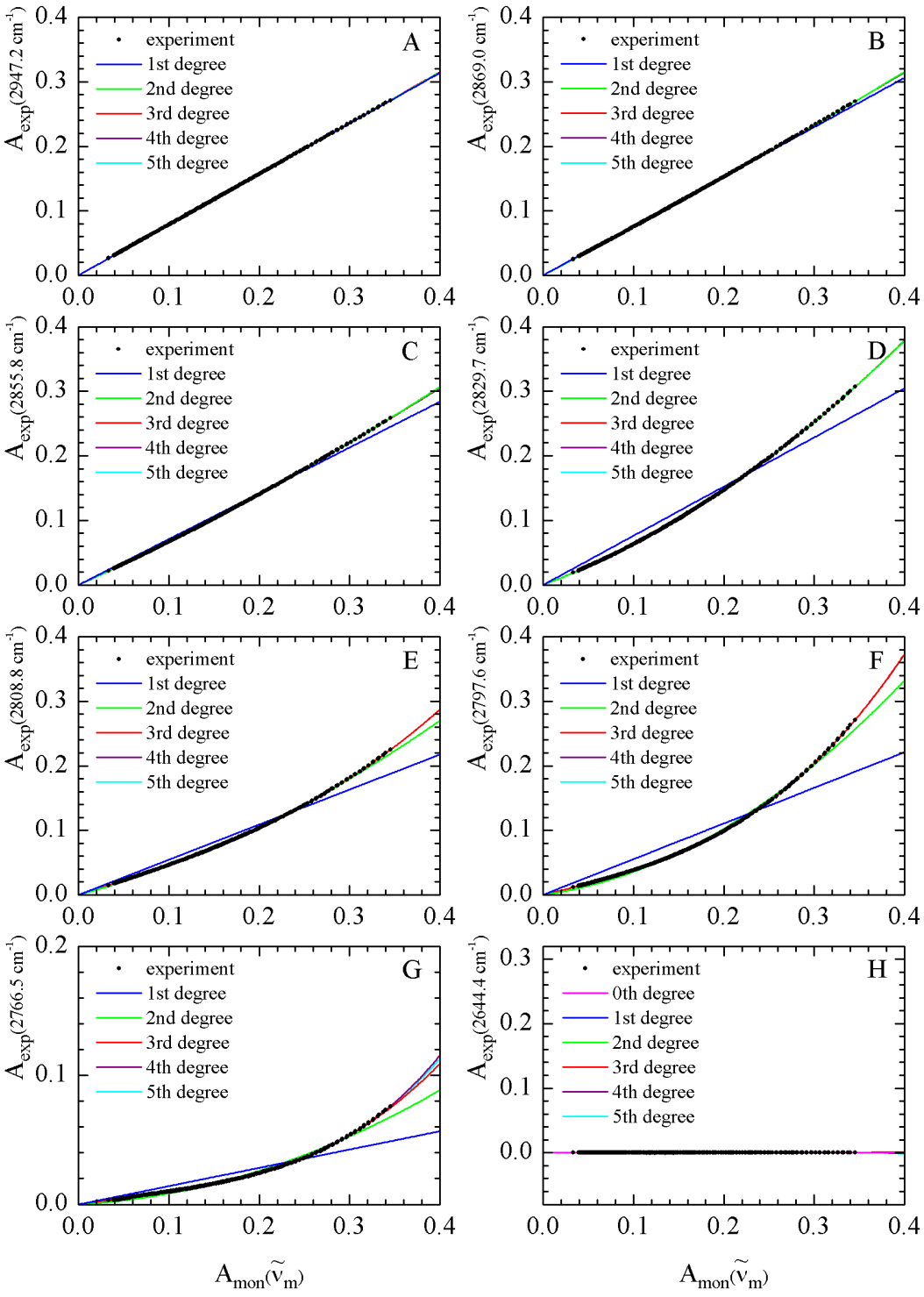

Fig. 4 Experimental absorbances for the different wavenumbers selected in Figure 2 and results from least-squares fitting procedures using different degree polynomials. To facilitate visual comparison, the lower degree polynomials are plotted in front of the higher degree polynomials. For panels A to G where the experimental absorbance versus the monomer absorbance is increasing significantly, the 0^{th} degree polynomial is not shown.

Apart from the visual comparison, the choice of the optimal polynomial can be guided using several statistical criteria. These criteria are a tool to evaluate whether a certain model performs "better" than another model. Here, we define a "better" model to be one that is parsimonious in the polynomial degree *p* and one that has a smaller residual sum of squares, *RSS*

$$RSS_p(\tilde{\nu}_i) = \sum_{k=1}^{n} [A_{exp,k}(\tilde{\nu}_i) - A_{calc,k}\,(\tilde{\nu}_i)]^2 \tag{20}$$

with n the number of spectra in the dataset, i.e. number of measured datapoints for each wavenumber. The top of Table 1 indicates the *RSS* values for the polynomials illustrated in Figure 4 for the wavenumbers A to H defined in Figure 2. Because adding an extra parameter to the polynomial increases the flexibility of the model, the *RSS* values decrease with increasing polynomial degree. If we use the smallest RSS value as a criterion to select the appropriate polynomial degree we would end up with a plot almost identical to Figure 3c with a fixed polynomial degree of five. For completeness this plot is shown in Figure S1 of the supporting information.

In this paper we compare the polynomial models resulting from two commonly used criteria, the A and B information criteria (*AIC*, *BIC*) that have a penalty factor for the number of parameters in the model. The *AIC* was published by Akaike [39] in 1973 and is defined as

$$AIC_p(\tilde{\nu}_i) = \mathrm{n} \cdot \ln\left(\frac{RSS_p(\tilde{\nu}_i)}{n}\right) + r \cdot 2 \tag{21}$$

with *p* the polynomial degree, *n* the number of spectra in the dataset and *r* the number of fitted parameters (r=1 for p=0 and r=p-1 for p>0). The 'better' model is then selected as the one with the smallest *AIC* value when comparing with the values for all suggested polynomials. In 1978, Schwarz [40] published a modified information criteria, the *BIC*, which is defined as

$$BIC_p(\tilde{\nu}_i) = \mathrm{n} \cdot \ln\left(\frac{RSS_p(\tilde{\nu}_i)}{n}\right) + k \cdot \ln(n) \tag{22}$$

As with the *AIC*, the most appropriate model is the one with the minimum *BIC* value. The difference between both criteria is that the *BIC* has a log(*n*) penalty factor, in contrast to the factor 2 in the AIC. This implies that for large datasets, the *BIC* has a heavier penalty for the number of parameters in the model and therefore tends to yield models with fewer parameters, e.g. a lower polynomial degree.

The bottom of Table 1 contains the *AIC* and the *BIC* values for the polynomials illustrated in Figure 4 for the wavenumbers A to H defined in Figure 2. It can be seen that the *RSS* values for the higher degree polynomials are all very small, which is expected as these

polynomialsshow good fit with the data in the panels of Figure 4. To select the appropriate polynomial, the *AIC* and *BIC* values for 8 wavenumbers were calculated. The smallest *AIC* and *BIC* values are indicated in italic and the accompanying polynomial degree is mentioned. Overall we can conclude that both criteria result in rather large polynomial models. Were we would have expected models 0^{th}, 1^{st} or 2^{nd} degree models for wavenumbers A to D and H, the criteria indicate that 4^{th} or 5^{th} models would be more appropriate. It should also be noted that for wavenumbers C to E a smaller polynomial degree was selected when using the *BIC* then instead of the *AIC*. This is an example of the fact that *BIC* indeed tends to yield in smaller models.

Table 1. *RSS*, *AIC* and *BIC* values for the least-squares fit of the 0^{th} to 5^{th} degree polynomial at wavenumbers indicated in Figure 1. The polynomial regressions of the 1^{st} to 6^{th} degree have an intercept value equal to zero. For both the *AIC* and *BIC* the appropriate polynomial degree is selected as the one with the minimum *AIC* or *BIC* value, which are given in italic.

	p	A	B	C	D	E	F	G	H
		2947.2	2869.0	2855.8	2829.7	2808.8	2797.6	2766.5	2644.4
RSS	0	1.45	1.43	1.31	1.81	0.90	1.22	8.40×10^{-2}	1.32×10^{-5}
	1	2.44×10^{-5}	8.78×10^{-4}	5.49×10^{-3}	5.60×10^{-2}	2.77×10^{-2}	1.32×10^{-1}	1.15×10^{-2}	1.73×10^{-5}
	2	1.62×10^{-5}	1.51×10^{-5}	1.64×10^{-5}	2.11×10^{-5}	6.94×10^{-4}	3.94×10^{-3}	1.00×10^{-3}	1.66×10^{-5}
	3	1.54×10^{-5}	1.50×10^{-5}	1.57×10^{-5}	2.05×10^{-5}	1.044×10^{-5}	1.464×10^{-5}	3.71×10^{-5}	1.26×10^{-5}
	4	1.390×10^{-5}	1.445×10^{-5}	1.48×10^{-5}	2.03×10^{-5}	1.035×10^{-5}	1.458×10^{-5}	1.70×10^{-5}	1.07×10^{-5}
	5	1.398×10^{-5}	1.440×10^{-5}	1.47×10^{-5}	2.01×10^{-5}	1.034×10^{-5}	1.455×10^{-5}	1.63×10^{-5}	1.01×10^{-5}
AIC	0	-1891.8	-1897.4	-1928.4	-1816.4	-2056.9	-1952.3	-2877.7	-5908.0
	1	-5695.2	-4455.8	-3821.9	-3018.0	-3262.2	-2722.6	-3566.7	-5814.4
	2	-5834.5	-5859.0	-5830.7	-5744.1	-4535.1	-3934.1	-4409.0	-5826.8
	3	-5850.3	-5859.6	-5845.0	-5751.8	-5985.4	*-5868.5*	-5546.7	-5921.0
	4	*-5884.4*	*-5871.1*	-5861.9	-5754.0	*-5986.4*	-5867.8	-5815.3	-5974.7
	5	-5882.5	-5870.2	*-5863.4*	*-5754.7*	-5984.8	-5866.6	*-5828.2*	*-5994.6*
Selected degree		4	4	5	5	4	3	5	5
BIC	0	-1887.9	-1893.5	-1924.6	-1812.5	-2053.1	-1948.5	-2873.9	-5904.1
	1	-5691.4	-4451.9	-3818.1	-3014.2	-3258.4	-2718.7	-3562.9	-5810.6
	2	-5826.8	-5851.3	-5823.0	-5736.4	-4527.4	-3926.4	-4401.3	-5819.1
	3	-5838.8	-5848.1	-5833.4	*-5740.3*	*-5973.8*	*-5856.9*	-5535.1	-5909.5
	4	*-5869.0*	*-5855.7*	*-5846.6*	-5738.6	-5971.0	-5852.4	-5799.9	-5959.3
	5	-5863.3	-5851.0	-5844.2	-5735.5	-5965.6	-5847.3	*-5808.9*	*-5975.4*
Selected degree		4	4	4	3	3	3	5	5

The panels on the left in Figure 5 demonstrate the results when applying the *AIC* and the *BIC* to every wavenumber of the recorded dataset. Successively, the appropriated polynomial degree is selected according the *AIC* or *BIC* criterion and the monomer and oligomer contributions are plotted in the panels below. Comparison of the spectra obtained with both criteria show that the *BIC* is tending towards lower polynomial degrees, slightly reducing the noise caused by compensation effects. When comparing the *BIC* result with Figure 3c it can be seen that the *BIC* method reduces noise but still a large amount of negative contributions remain. As shown before the noise in the resulting spectrum can be brought down by prohibiting polynomial coefficients to be negative by introducing nonnegative constraints.

Table 2 presents the *RSS*, *AIC* and *BIC* values for the polynomial regression with added nonnegative constraints for the wavenumbers A to H defined in Figure 2. The polynomial degrees resulting from the A and B information criteria are lower than the ones in Table 1 and show a better match with the polynomial degree made by visual inspection of Figure 2. It should be noted that the plots of the polynomial models with nonnegative constraints are almost identical with those in Figure 2 and can be found in the supporting information as Figure S2. Although the degrees selected by both criteria are identical in Table 2, this is not the case for all wavenumbers. Inspection of the data in the upper panels of Figure 5 shows that for the analysis involving nonnegative constraints, *BIC* generally yields lower values. The predicted contributions for monomer and complexes are almost identical, except for the 2860-2880 cm^{-1} region where the monomer, dimer, trimer and tetramer contributions show substantial overlap. The pentamer contribution observed in Figure 3f and in the right (AIC) panel of Figure 5 is strongly reduced. This suggests that at least for the systems studied here the *BIC* is able to better separate the different absorbance bands in regions where several contributions overlap.

It can be noted that when looking at the top of Table 2 the *RSS* value at every wavenumber stays constant from a certain polynomial degree. This means that when including the nonnegative constraint, the addition of an extra term in the polynomial yields little or no improvement of the fit. Because in this case the *RSS* can be used as another criterion for selection of the polynomial model we added the result in the right picture of Figure S1 of the supporting information for completeness.

We like to point out that although we used three different strategies, a fixed polynomial degree or selecting the appropriate polynomial degree with the *AIC* or the *BIC*, at the end the

resulting spectra show high resemblance if the nonnegative constraints are added, Figure 3f and the figures on the right of Figure 5.

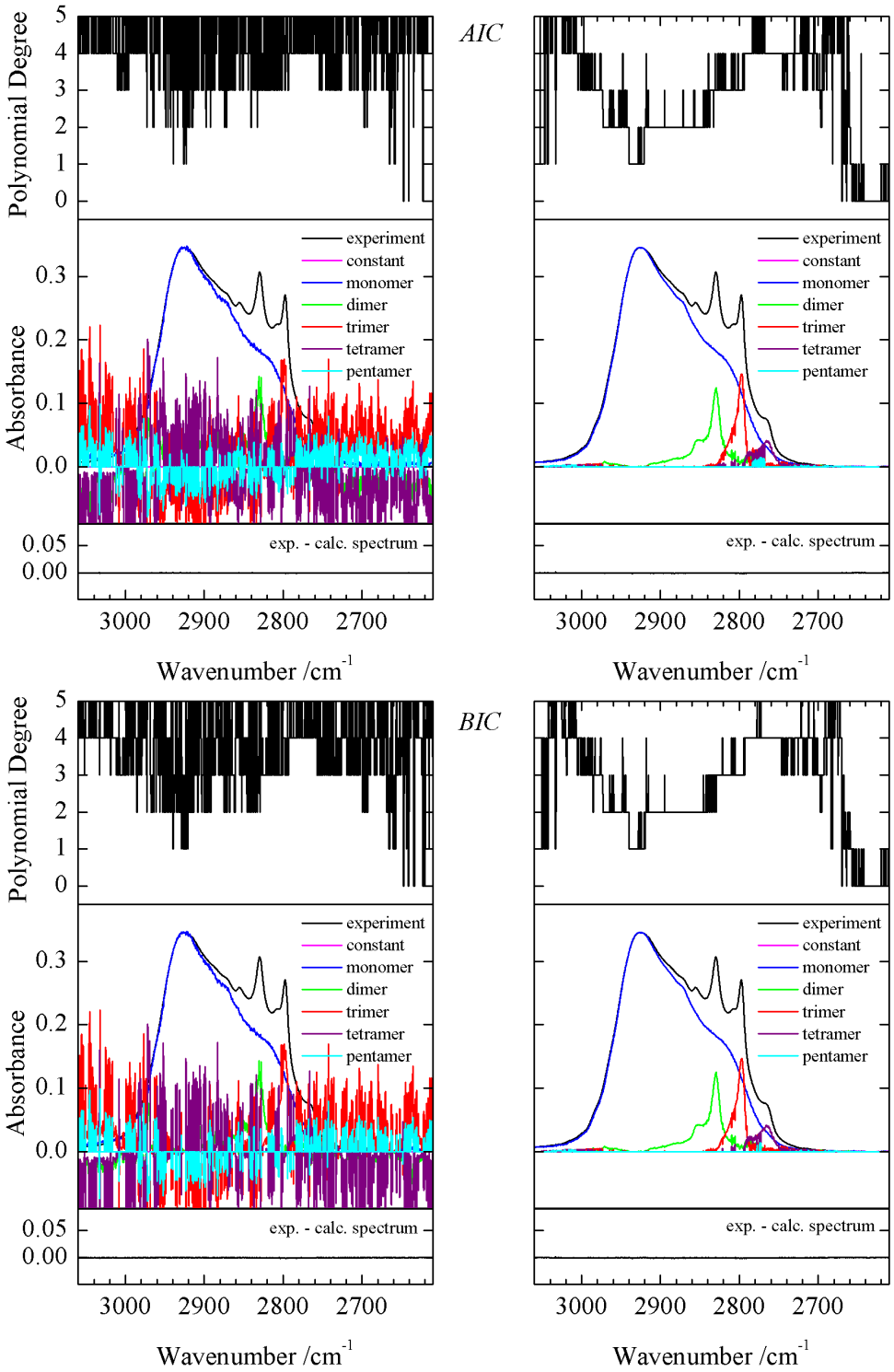

Fig. 5 The experimental data analyzed with polynomial regression using the minimum value of the *AIC* and the *BIC* to select the appropriate polynomial degree varying between 0 and 5 for every wavenumber. The panels on the left show the different monomer and oligomer contributions using regular least-squares procedures and the results in the right panels are obtained by adding nonnegative

constraints for the coefficients. The bottom panel in each figure show the difference between the experimental spectrum and the model.

Table 2. *RSS*, *AIC* and *BIC* values for the least-squares fit with nonnegative constraints of the 0^{th} to 5^{th} degree polynomial at wavenumbers indicated in Figure 1. The polynomial regressions of the 1^{st} to 6^{th} degree have an intercept value equal to zero. For both the *AIC* and *BIC* the appropriate polynomial degree is selected as the one with the minimum *AIC* or *BIC* value, which are given in italic.

	p	A	B	C	D	E	F	G	H
		2947.2	2869.0	2855.8	2829.7	2808.8	2797.6	2766.5	2644.4
RSS	0	1.45	1.43	1.31	1.81	0.90	1.22	8.40×10^{-2}	1.32×10^{-5}
	1	2.44×10^{-5}	8.78×10^{-4}	5.49×10^{-3}	5.60×10^{-2}	2.77×10^{-2}	1.32×10^{-1}	1.15×10^{-2}	1.73×10^{-5}
	2	1.62×10^{-5}	1.51×10^{-5}	1.64×10^{-5}	2.11×10^{-5}	6.94×10^{-4}	3.94×10^{-3}	1.00×10^{-3}	1.73×10^{-5}
	3	1.56×10^{-5}	1.51×10^{-5}	1.64×10^{-5}	2.05×10^{-5}	1.04×10^{-5}	1.46×10^{-5}	1.80×10^{-4}	1.73×10^{-5}
	4	1.56×10^{-5}	1.51×10^{-5}	1.64×10^{-5}	2.05×10^{-5}	1.04×10^{-5}	1.46×10^{-5}	1.70×10^{-5}	1.73×10^{-5}
	5	1.56×10^{-5}	1.51×10^{-5}	1.64×10^{-5}	2.05×10^{-5}	1.04×10^{-5}	1.46×10^{-5}	1.70×10^{-5}	1.73×10^{-5}
AIC	0	-1891.8	-1897.4	-1928.4	-1816.4	-2056.9	-1952.3	-2877.7	*-5908.0*
	1	-5695.2	-4455.8	-3821.9	-3018.0	-3262.2	-2722.6	-3566.7	-5814.4
	2	-5834.5	*-5859.0*	*-5830.7*	-5744.1	-4535.1	-3934.1	-4409.0	-5812.4
	3	*-5847.3*	-5857.0	-5828.7	*-5751.8*	*-5985.4*	*-5868.5*	-4999.6	-5810.4
	4	-5845.3	-5855.0	-5826.7	-5749.8	-5983.4	-5866.5	*-5814.9*	-5808.4
	5	-5843.3	-5853.0	-5824.7	-5747.8	-5981.4	-5864.5	-5812.9	-5806.4
Selected degree		3	2	2	3	3	3	4	0
BIC	0	-1887.9	-1893.5	-1924.6	-1812.5	-2053.1	-1948.5	-2873.9	*-5904.1*
	1	-5691.4	-4451.9	-3818.1	-3014.2	-3258.4	-2718.7	-3562.9	-5810.6
	2	-5826.8	*-5851.3*	*-5823.0*	-5736.4	-4527.4	-3926.4	-4401.3	-5804.7
	3	*-5835.7*	-5845.5	-5817.2	*-5740.3*	*-5973.8*	*-5856.9*	-4988.0	-5798.9
	4	-5829.9	-5839.6	-5811.4	-5734.4	-5968.0	-5851.1	*-5799.5*	-5793.0
	5	-5824.0	-5833.8	-5805.5	-5728.6	-5962.1	-5845.2	-5793.7	-5787.2
Selected degree		3	2	2	3	3	3	4	0

4. CONCLUSIONS

In this paper, we report on the development and validation of numerical methods in which the concentration dependent behavior observed for HCl in liquid argon is analyzed by least-squares fitting approaches. In these methods, each wavenumber a polynomial is used to mimic the relation between monomer concentrations and measured absorbances. It was shown that least-squares fitting of higher degree polynomials tends to overfit and at the same time leads to unphysical compensation effects where a contribution due to one species is compensated for by a negative contribution of another. These issues can be corrected for by carefully analyzing, using AIC or BIC information criteria, the differences between consecutive least-squares fits observed when the degree of the polynomial used is systematically increased, and by introducing additional constraints prohibiting negative absorbances to occur for the monomer or for one of the oligomers. We believe that, the models and approaches developed, and the Matlab based software packages used for their implementation can now be considered robust and thus should allow other systems to be analyzed with a much higher accuracy than before. Results obtained for solutions of ammonia and ammonia-d_3 dissolved in liquid xenon, for which similar but more complicated spectral features due to self-association are known to exist [30, 31], are currently investigated, and will be reported in a consecutive study.

ACKNOWLEDGMENT

Liene I. De Beuckeleer acknowledges the Research Foundation – Flanders, FWO, for the appointment of a PhD Fellowship and for financial help towards the purchase of spectroscopic equipment used in this study. The authors thank the Flemish Community for financial support through the Special Research Fund (BOF). Sam Jacobs is thanked for valuable discussions.

REFERENCES

[1] M. Fárník, D.J. Nesbitt, J. Chem. Phys., 121 (2004) 12386-12395.
[2] M. Fárník, S. Davis, D.J. Nesbitt, Faraday Discuss., 118 (2001) 63-78.
[3] M.D. Schuder, C.M. Lovejoy, R. Lascola, D.J. Nesbitt, J. Chem. Phys., 99 (1993) 4346.
[4] A. Furlan, S. Wulfert, S. Leutwyler, Chem. Phys. Lett., 153 (1988) 291-295.
[5] T. Häber, U. Schmitt, M.A. Suhm, Phys. Chem. Chem. Phys., 1 (1999) 5573-5582.
[6] D. Skvortsov, M.Y. Choi, A.F. Vilesov, J. Phys. Chem. A, 111 (2007) 12711-12716.
[7] A.J. Barnes, K. Szczepaniak, W.J. Orville Thomas, J. Mol. Struct., 59 (1980) 39-53.
[8] A. Engdahl, B. Nelander, J. Chem. Phys., 94 (1990) 8777-8780.
[9] B.J. van der Veken, F.R. De Munck, J. Chem. Phys., 97 (1992) 3060-3071.
[10] W.A. Herrebout, J. Van Gils, B.J. van der Veken, J. Mol. Struct., 563-564 (2001) 249-255.
[11] E.R. Malinowski, Factor Analysis in Chemistry, 3rd ed., John Wiley & Sons, Inc., New York, 2002.
[12] B.J. van der Veken, J. Phys. Chem., 100 (1996) 17436-17438.
[13] N. Nagels, Y. Geboes, B. Pinter, F. De Proft, W.A. Herrebout, Chem. Eur. J., 20 (2014) 8433-8443.
[14] B. Michielsen, C. Verlackt, B.J. van der Veken, W.A. Herrebout, J. Mol. Struct., 1023 (2012) 90-95.
[15] B. Michielsen, J.J. Dom, B.J. van der Veken, S. Hesse, M.A. Suhm, W.A. Herrebout, Phys. Chem. Chem. Phys., 14 (2012) 6469-6478.
[16] J.J. Dom, B.J. van der Veken, B. Michielsen, S. Jacobs, Z.F. Xue, S. Hesse, H.M. Loritz, M.A. Suhm, W.A. Herrebout, Phys. Chem. Chem. Phys., 13 (2011) 14142-14152.
[17] J.J.J. Dom, B. Michielsen, B.U.W. Maes, W.A. Herrebout, B.J. van der Veken, Chem. Phys. Lett., 469 (2009) 85-89.
[18] D. Hauchecorne, N. Nagels, B.J. van der Veken, W.A. Herrebout, Phys. Chem. Chem. Phys., 14 (2012) 681-690.
[19] B. Michielsen, W.A. Herrebout, B.J. van der Veken, ChemPhysChem, 8 (2007) 1188-1198.
[20] B. Michielsen, J.J. Dom, B.J. van der Veken, S. Hesse, Z. Xue, M.A. Suhm, W.A. Herrebout, Phys. Chem. Chem. Phys., 12 (2010) 14034-14044.
[21] B. Michielsen, W.A. Herrebout, B.J. van der Veken, ChemPhysChem, 9 (2008) 1693-1701.
[22] D. Hauchecorne, R. Szostak, W.A. Herrebout, B.J. van der Veken, ChemPhysChem, 10 (2009) 2105-2115.
[23] D. Hauchecorne, B.J. van der Veken, A. Moiana, W.A. Herrebout, Chem. Phys., 374 (2010) 30-36.

[24] D. Hauchecorne, A. Moiana, B.J. van der Veken, W.A. Herrebout, Phys. Chem. Chem. Phys., 13 (2011) 10204-10213.
[25] D. Hauchecorne, B.J. van der Veken, W.A. Herrebout, P.E. Hansen, Chem. Phys., 381 (2011) 5-10.
[26] D. Hauchecorne, W.A. Herrebout, J. Phys. Chem. A., 117 (2013) 11548-11557.
[27] N. Nagels, D. Hauchecorne, W.A. Herrebout, Molecules, 18 (2013) 6829-6851.
[28] N. Nagels, W.A. Herrebout, Spectrochim. Acta, Part A, 136 A (2015) 16-26.
[29] Y. Geboes, N. Nagels, B. Pinter, F. De Proft, W.A. Herrebout, J. Phys. Chem. A, 119 (2015) 2502-2516.
[30] W.A. Herrebout, S.M. Melikova, S.N. Delanoye, K.S. Rutkowski, D.N. Shchepkin, B.J. van der Veken, J. Phys. Chem. A, 109 (2005) 3038-3044.
[31] K.S. Rutkowski, W.A. Herrebout, S.M. Melikova, P. Rodziewicz, B.J. van der Veken, A. Koll, Spectrochim. Acta A, 61 (2005) 1595-1602.
[32] W. Herrebout, Top. Curr. Chem., 358 (2015) 79-154.
[33] MATLAB, in, The MathWorks Inc., Natick, Massachusetts, 2014.
[34] B.J. van der Veken, W.A. Herrebout, J. Phys. Chem. A, 105 (2001) 7198-7204.
[35] A.I. Fishman, W.A. Herrebout, B.J. van der Veken, J. Phys. Chem. A, 106 (2002) 4536-4542.
[36] M.O. Bulanin, G.Y. Zelikina, Phase Equilibria and Spectral Analysis of Cryosystems., in: R.J.H. Clark, R.E. Hester (Eds.) Molecular Cryospectroscopy. Advances in Spectroscopy., Wiley, Chichester, 1995, pp. 22-34.
[37] A.I. Fishman, A.I. Noskov, R.M. Aminova, R.A. Skochilov, Spectrochim. Acta A, 136 (2015) 100-106.
[38] A.B. Remizov, R.A. Skochilov, Proc. SPIE 5507, 2004, pp. 195-202.
[39] H. Akaike, IEEE Transactions on Automatic Control, 19 (1974) 716-723.
[40] G. Schwarz, The Annals of Statistics, 6 (1978) 461-464.