ELSEVIER

Journal of Molecular Structure (Theochem) 464 (1999) 163–170

THEO CHEM

Factorial design – principal component regression calculation of fundamental vibrational frequencies

Ieda S. Scarminio[a,*], Anselmo E. de Oliveira[b], Roy E. Bruns[b]

[a]*Departamento de Química,Universidade Estadual de Londrina, 86010-150 Londrina, PR, Brazil*
[b]*Instituto de Química, Universidade Estadual de Campinas, 13083-970 Campinas, SP, Brazil*

Abstract

A factorial design-principal component regression method is proposed for calculating fundamental vibrational frequencies. The method is illustrated by estimating both observed and anharmonicity-corrected frequencies of CHF_3 using theoretical frequencies of CH_3F, CH_2F_2 and CHF_3 and observed and corrected frequencies of CH_3F and CH_2F_2. Theoretical frequencies are provided by ab initio molecular orbital calculations prescribed by 2^{4-1} fractional factorial designs. The estimated observed and anharmonicity-corrected CHF_3 frequencies are 0.1%–2.8% of the experimental values. These results are much more accurate than those obtained by simple linear regression (0.2%–11.9%) or those of the 2^{4-1} factorial design.

Keywords: Factorial design; Infrared frequencies; Principal component regression

1. Introduction

The calculation of vibrational frequencies for molecules has followed two general tendencies. In earlier times, force constants determined from the vibrational frequencies of small molecules were transferred to larger molecules in order to predict their frequencies. Within the harmonic oscillator approximation this method can be very accurate as least squares procedures [1–3] can be used to determine force constants from anharmonic-corrected frequencies of isotopomers. Errors arise owing to uncertainties in the anharmonic corrections as well as frequency corrections for Fermi resonance interactions between individual bands. Further, force constants are not exactly transferable from one molecule to another even if the two molecules have very similar electronic structures. More recently, with the availability of operationally simple ab initio molecular orbital programs, scientists not trained as specialists in vibrational spectroscopy can perform frequency calculations. Furthermore, frequency estimates are even possible for relatively unstable molecules. Unfortunately comparisons of molecular orbital results with observed band positions are also hampered by anharmonicity and Fermi resonance correction estimates. Besides this, force constant scaling factors must normally be applied as ab initio frequency estimates are often higher than the experimental values [4].

Principal component analysis was used in our laboratory to compare experimentally observed spectral parameters for several molecules, CH_3F[5], CH_2F_2[6] and the difluoro- and dichloroethylenes [7,8], with the results obtained from diverse molecular wave functions in attempts to identify those providing the most accurate results. As each one of these molecules has 3N-6 vibrational degrees of freedom, a half dozen or

* Corresponding author.
E-mail address: ieda@qui.uel.br (I.S. Scarminio)

0166-1280/99/$ - see front matter
PII: S0166-1280(98)00548-X

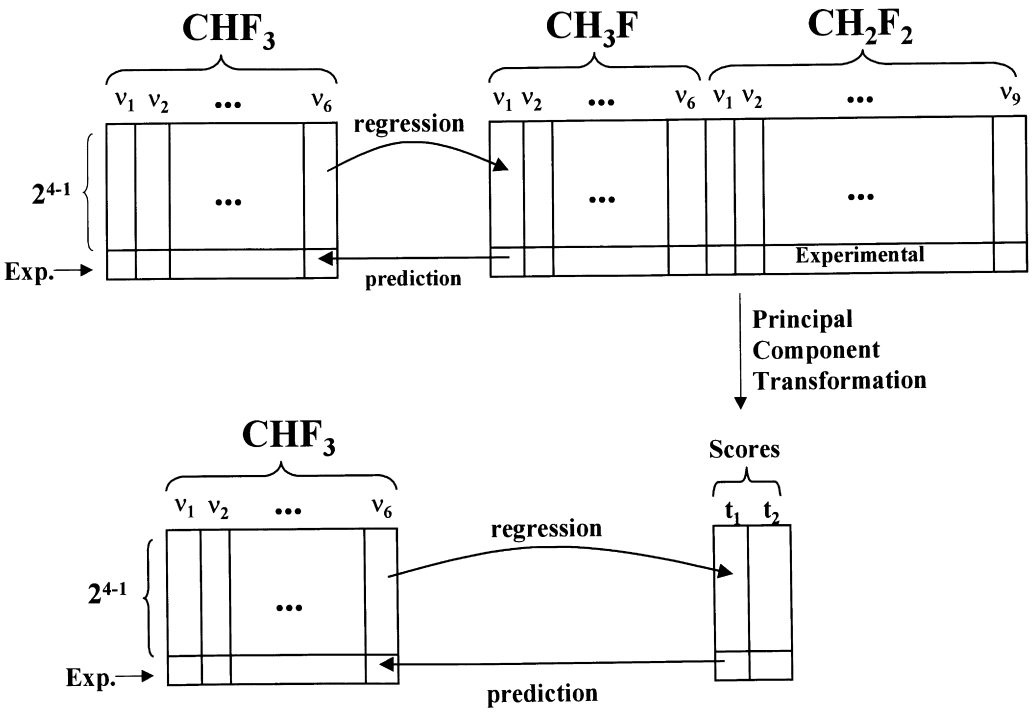

Fig. 1. The observed frequencies of CHF_3 can be predicted by regressing their molecular orbital frequency values on individual (simple linear regression) or all (multiple linear regression) CH_3F and CH_2F_2 frequencies. For principal component regression, scores are determined for theoretical CH_3F and CH_2F_2 frequencies and used as regressors for the CHF_3 theoretical frequencies. In both cases experimental CH_3F and CH_2F_2 values are used to provide predictions of the observed or anharmonic-corrected CHF_3 values.

more vibrational frequencies or intensities can be treated simultaneously with this methodology.

The choice of wave functions to be included in the principal component analysis is conveniently made using a statistical factorial design [9]. These designs are made using statistical criteria and are normally associated with experimental endeavors aimed at minimizing work and expenditures in the laboratory while still providing precise models capable of reproducing and even predicting measurement results as a function of varying experimental conditions.

In this work factorial design and principal component analysis are used to calculate the CHF_3 vibrational frequencies from those of CH_3F and CH_2F_2. Correlations of frequency values calculated from wave functions prescribed by a factorial design for CH_3F and CH_2F_2 with those calculated for the same design for CHF_3 are used along with the experimental frequency values of the former molecules to estimate CHF_3 experimental frequency values. As the calculations are based on statistical correlations rather than physical models, anharmonicity and Fermi resonance corrections, always nagging sources of uncertainty in conventional calculations, are automatically included.

2. Calculations

The proposed procedure is diagrammed in Fig. 1 where an ordinary linear regression situation is shown. Theoretical CHF_3 frequencies are individually regressed on one or more of the theoretical CH_3F and CH_2F_2 frequencies. Then the corresponding experimental CH_3F and CH_2F_2 frequencies are substituted into the regression equations to obtain estimates of the experimental CHF_3 frequencies. For principal component regression, the autoscaled theoretical CH_3F and CH_2F_2 frequency data are first subjected to a principal component transformation. The coordinates of the

Table 1
2^{4-1} Fractional factorial design for the calculation of the vibrational frequencies of CH_3F, CH_2F_2 and CHF_3

Factors	Levels	
	–	+
Basis set	6-31G	6-311G
Polarization functions	Absent	present
Diffuse functions	Absent	present
Electron correlation	Hartree–Fock	Møller–Plesset 2
Wave Function	Factorial designation	
MP2/6-31G	– – – +	
HF/6-311G	+ – – –	
HF/6-31G(d,p)	– + – –	
MP2/6-311G(d,p)	+ + – +	
HF/6-31 + + G	– – + –	
MP2/6-311 + + G	+ – + +	
MP2/6-31 + + G(d,p)	– + + +	
HF/6-311 + + G(d,p)	+ + + –	

original data in the new space are called principal component scores, represented by the *t* matrix in the figure. Theoretical CHF_3 frequencies are individually regressed on these scores to obtain regression equations. Finally, principal component transformed CH_3F and CH_2F_2 experimental frequency values are substituted into these equations to estimate the CHF_3 experimental frequencies.

A 2^{4-1} fractional factorial design, shown in Table 1 requiring eight wave function calculations, is used to predict the CHF_3 frequency values. The four varying factors are (1) the use of a 6-31G or 6-311G valence basis set, (2) the presence or not of polarization functions in the basis set, (3) the presence or not of diffuse functions in this basis and (4) the use or not of second-order Møller–Plesset perturbation corrections to the Hartree–Fock level calculations. The theoretical CH_3F and CH_2F_2 values are taken from Refs. [5 and 6].

The ab initio molecular orbital calculations were carried out using the GAUSSIAN 94 computer program [10] on IBM RISC 6000 at CENAPAD-SP and DIGITAL ALFA 1000 workstations. The principal component transformation and regressions were performed using software prepared in our laboratory.

3. Results

The molecular orbital values of the CH_3F, CH_2F_2 and CHF_3 vibrational frequencies for the fractional factorial design in Table 1 are presented in Tables 2–4. Included in these tables are the observed experimental frequency values as well as values that have been corrected for anharmonicity [1,11–13]. Note that the CH symmetric stretching frequency values for CH_3F and CHF_3 in these tables were also corrected for Fermi resonance.

The simple regressions were carried out by regressing the theoretical values of the 2^{4-1} fractional factorial design for CHF_3 on the corresponding CH_3F values, i.e. the theoretical $\nu_i(CHF_3)$ values were each regressed on the corresponding $\nu_i(CH_3F)$ values for $i = 1,2\cdots6$. For the principal component regression, scores were first calculated for the

Table 2
Calculated and experimental fundamental vibrational frequencies of CH_3F (cm^{-1}) for the 2^{4-1} fractional factorial design

Wave function	ν_1 (A_1) CH_3 str.	ν_2 (A_1) CH_3 bend	ν_3 (A_1) CF str.	ν_4 (E) CH_3 str.	ν_5 (E) CH_3 def.	ν_6 (E) CH_3 def.
– – – +	3096.2	1534.6	996.2	3202.2	1569.6	1172.4
+ – – –	3212.4	1621.4	1080.8	3309.0	1644.7	1264.0
– + – –	3203.5	1637.0	1186.1	3286.3	1633.0	1307.2
+ + – +	3087.6	1536.6	1105.6	3186.3	1519.1	1224.0
– – + –	3251.7	1608.6	1047.2	3359.5	1644.7	1254.3
+ – + +	3050.4	1495.4	924.4	3171.9	1541.9	1144.6
– + + +	3143.5	1525.3	1056.7	3259.6	1548.5	1211.8
+ + + –	3193.5	1610.9	1156.6	3276.9	1614.0	1295.5
Observed[a]	2930[b]	1464	1048.6	3005.8	1467	1182
Harmonic[a]	3055	1500	1067	3165	1510	1212

[a] Ref. [1]
[b] corrected for Fermi resonance

Table 3
Calculated and experimental fundamental vibrational frequencies of CH_2F_2 (cm^{-1}) for the 2^{4-1} fractional factorial design

Wave function	ν_1 (A_1) CH str.	ν_2 (A_1) CH_2 scissor	ν_3 (A_1) CF str.	ν_4 (A_1) CF_2 bend.	ν_5 (A_2) CH_2 twist	ν_6 (B_1) CH str.	ν_7 (B_1) CH_2 rock	ν_8 (B_2) CH_2 wag.	ν_9 (B_2) CF_2 str.
− − − +	3157.3	1594.9	1039.5	467.1	1258.1	3254.9	1156.5	1486.7	1066.6
+ − − −	3300.3	1683.2	1135.5	520.4	1362.9	3391.3	1245.5	1591.3	1157.8
− + − −	3256.4	1695.3	1238.0	571.3	1402.8	3328.7	1299.3	1629.1	1256.4
+ + − +	3126.7	1570.1	1145.7	537.2	1315.7	3209.5	1214.8	1523.4	1148.0
− − + −	3336.7	1668.2	1116.6	514.5	1359.1	3439.6	1223.3	1566.0	1133.6
+ − + +	3117.5	1554.5	982.1	453.8	1214.5	3227.4	1128.6	1439.8	966.1
− + + +	3185.1	1583.1	1115.6	514.3	1293.7	3282.8	1196.5	1487.4	1100.6
+ + + −	3252.5	1665.6	1221.9	580.5	1349.9	3325.5	1282.9	1595.8	1221.0
Observed[a]	2948	1508	1113	529	1262	3014	1178	1435	1090
Harmonic[a]	3071	1539	1124	539	1288	3140	1202	1464	1101

[a] values taken from Ref. [13]

Table 4
Calculated and experimental fundamental vibrational frequencies of CHF_3 (cm^{-1}) for the 2^{4-1} fractional factorial design

Wave function	ν_1 (A_1) CH str.	ν_2 (A_1) CF sym. str.	ν_3 (A_1) CF sym. bend	ν_4 (E) CH bend	ν_5 (E) CF antisym. str.	ν_6 (E) CF antisym bend
− − − +	3260.2	1069.1	634.6	1412.9	1131.7	452.4
+ − − −	3410.3	1155.2	694.5	1523.2	1239.1	504.0
− + − −	3359.9	1262.0	761.9	1574.0	1326.7	551.1
+ + − +	3220.2	1163.7	709.8	1449.6	1203.3	516.2
− − + −	3439.4	1142.4	689.3	1490.5	1226.1	500.4
+ − + +	3208.0	1015.7	614.0	1347.4	1035.3	442.3
− + + +	3268.8	1140.4	687.0	1420.1	1169.0	496.2
+ + + −	3350.1	1246.8	764.3	1539.6	1298.9	558.3
Observed[a]	2991.1[b]	1141.3	700.1	1377.7	1157.5	507.8
Harmonic[c]	3077	1154.7	709.7	1397.5	1187.5	518.9

[a] Ref. [11]
[b] corrected for Fermi resonance
[c] Ref. [12]

Table 5
Predicted CHF_3 fundamental frequency values, in cm^{-1}

Observed frequencies					Frequencies corrected for anharmonicity				
ν_{exp}	ν^{a}SR	relative error[b]	ν^{c}_{PCR}	Relative error[b]	ν_{exp}	$\nu^{a}_{SP}(cm^{-1})$	relative error[b]	ν^{c}_{PCR}	Relative error[b]
		(%)		(%)			(%)		(%)
2991.1	3046.8	1.86	3033.8	1.42	3077 ± 30	3195.8	3.86	3161.6	2.75
1141.3	1020.5	10.58	1142.6	0.11	1154.7 ± 10	1063.6	7.94	1148.3	0.55
700.1	616.9	11.88	698.6	0.21	709.7 ± 7	628.3	11.47	698.4	1.59
1377.7	1265.7	8.13	1373.9	0.28	1397.5 ± 14	1395.2	0.16	1400.1	0.19
1157.5	1054.7	8.88	1129.0	2.46	1187.5 ± 10	1107.1	6.77	1165.5	1.85
507.8	467.2	8.00	504.8	0.59	518.9 ± 5	487.5	6.05	517.8	0.21

[a] simple linear regression of theoretical CF_3H frequencies on corresponding CH_3F frequencies

[b] percentage relative error, $\left(\frac{\nu_{calc} - \nu_{obs}}{\nu_{obs}}\right) \times 100\%$

[c] principal component regression of CF_3H frequencies on CH_3F and CH_2F_2 frequencies

combined molecular orbital CH_3F–CH_2F_2 frequency values of Tables 2 and 3. Then separate regressions of the theoretical values of each CHF_3 fundamental frequency (Table 4) were performed on the scores of the first two principal components of this CH_3F–CH_2F_2 data set. To predict the CHF_3 frequencies, the PC scores of the observed and anharmonicity-corrected frequencies of CH_3F and CH_2F_2 were substituted in the regression equations. Table 5 presents predicted frequencies for CHF_3 obtained by simple linear regression and principal component regression. Observed CHF_3 frequency values and those corrected for anharmonicity are included in this table for comparison.

Simple linear regression results in calculated frequency values deviating from the observed values by relative errors, $[(\nu_{cal} - \nu_{obs})/\nu_{obs}] \times 100\%$, between 1.9% and 11.9%. Its predictions of anharmonicity-corrected values are slightly better deviating by 0.2%–11.5%. The largest relative deviations occur for the ν_3 CF symmetric bending frequency in both cases.

The calculated results for PCR are much more accurate. Relative errors of the calculated frequencies from the observed values range from only 0.1% to 2.5%. The relative errors in the PC regression results aimed at predicting the anharmonic-corrected CHF_3 frequencies are about the same with values from 0.2% to 2.8%. A more rigorous comparison of calculated and experimental values involves considering errors in the anharmonic-corrected frequencies that have been estimated by Kirk and Wilt [12]. These errors include contributions from band position measurements and uncertainties in correcting the observed frequencies for anharmonicity. They are reproduced in the first column of the second half of Table 5. Except for the CH stretching frequency, ν_1, the differences between the PCR calculated and experimental values of the harmonic frequencies are less than two times the error estimate in the observed values. The ν_1 calculated value is 85 cm^{-1} larger than the ν_1 value of Ref. [12]. Although this difference is about three times the experimental uncertainty, the PCR ν_1 estimate is clearly superior to the one obtained by simple linear regression and significantly better than any of the ab initio results in Table 4, which are 130–260 cm^{-1} above this value.

A critical assumption made for the calculation of these results involves the use of only two PC's to describe the CH_3F–CH_2F_2 frequency space. Together they account for 97.3% of the total variance in these frequency values. The third PC describes an additional 2% of the total variance and could provide valuable information for frequency prediction. For this reason PCR calculations of all CHF_3 fundamental frequencies were also carried out using three principal components. Comparison of these results with the calculated frequencies using only two PC's showed a maximum difference of only 3 cm^{-1} for all fundamental frequencies, confirming the validity of the two PC model.

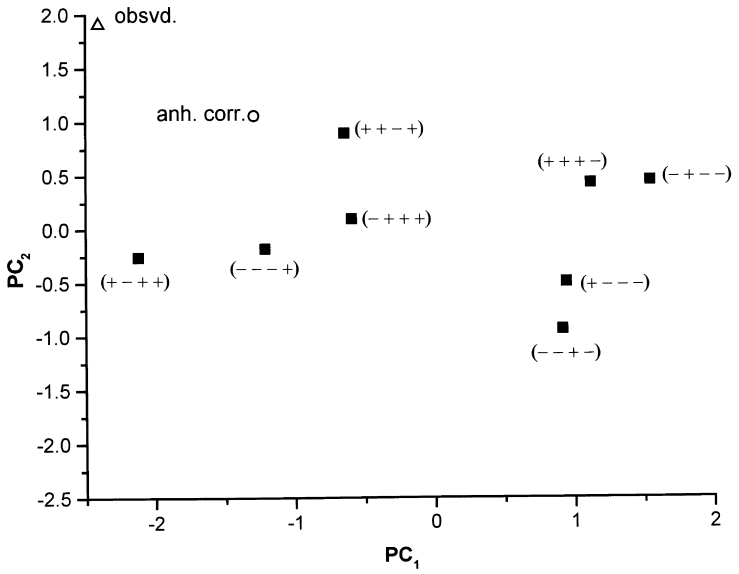

Fig. 2. Principal component score graph of the 2^{4-1} fractional factorial design CH_3F and CH_2F_2 autoscaled frequencies.

4. Discussion

Statistical models are not destined to compete with mechanistic models. When sufficient physical chemical information is available the latter are very efficient and provide insight about the nature of chemical bonding in the molecules being studied. However, if the main objective is to simply predict observed frequency values, multivariate statistical methods can be very useful especially for large molecules. With an adequate calibration set there is no reason to expect empirical multivariate statistical frequency predictions to be less accurate for large molecules than for small ones.

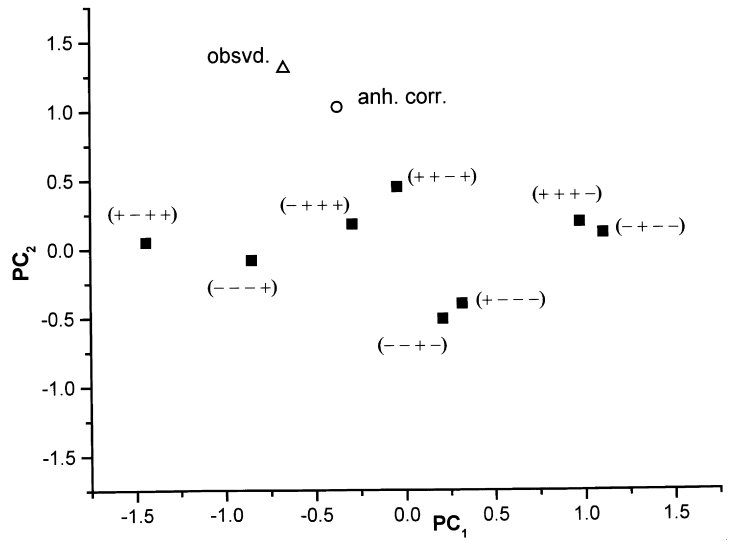

Fig. 3. Principal component score graph of the 2^{4-1} fractional factorial design CHF_3 autoscaled frequencies.

The empirical statistical approach to calculating vibrational frequencies works well because the molecular orbital calculated frequencies of CHF_3 are highly correlated with those of CH_3F and CH_2F_2 for the 2^{4-1} fraction factorial design employed here. However, owing to their high dimensionalities (15 for the combined CH_3F–CH_2F_2 space and six for the CHF_3 space) these correlations are not easily visualized. The principal component transformation, besides allowing a feasible multiple linear regression of the fractional factorial results, permits projections of these spaces in two dimensions. The principal component graphs for the combined CH_3F–CH_2F_2 and the individual CHF_3 frequency spaces are shown in Figs. 2 and 3. The coordinates of the two principal component axes in Fig. 2 for the CH_3F–CH_2F_2 frequency space are the same as the scores symbolized in Fig. 1 and used in the principal component regression. The correlations between the scores of the two spaces are clearly evident upon comparing the spatial arrangements of the points in these figures. They are practically the same for both graphs; the theoretical (+ + + −) and (− + − −) points fall in the first quadrants, the observed, corrected and (− + + +) and (+ + − +) points are positioned in the second ones, the (− − − +) and (+ − + +) results are in or close to the third and the (+ − − −) and (− − + −) results occupy the fourth quadrants. As these spaces explain 97.3% and 98.8% of the total variances of their respective frequency data sets they accurately represent the original 15 and six dimensional spaces.

It is of interest to understand why the correlations between characteristic frequencies of CH_3F, CH_2F_2 and CHF_3 calculated for this set of wave functions are so high. Inspection of the values in Tables 2–4 indicates that the frequency changes provoked by the wave function modifications are systematic. Note that the frequency values for Hartree–Fock level results are almost always larger than the Møller–Plesset 2 level results. If full factorial design results were given in these tables the systematic deviations would be more obvious. See, for example, the full factorial design results already reported for CH_3F[5]and CH_2F_2[6].

The systematic changes in the calculated frequencies for wave function modifications are most conveniently analyzed using factorial models, containing average values quantitatively describing the effects of changing factor levels on the frequency results. On the average the effect of introducing Møller–Plesset perturbation lowers the CH stretching frequencies in these molecules by between 100–150 cm^{-1}. The basis change effects (6-31 to 6-311) for these frequencies fall in the narrow -35 and -42 cm^{-1} range and even their polarization function–electron correlation interaction effect values are almost the same (32 to 40 cm^{-1}).

The CF stretching frequencies also have very similar effect values, between –96 and -138 cm^{-1} for electron correlation, between $+91$ and $+115$ cm^{-1} for addition of polarization functions and between -26 and -49 cm^{-1} for the addition of diffuse functions. All other effect values are much less. The CH bends are also highly correlated with the CF stretches having very similar effect values. In fact the two principal component model seems to be valid for these data as there appears to be two main sources of variation in the theoretical frequency data of these molecules, one for the CH stretching frequencies and another for the CF stretching and CH bending frequencies.

In closing, it should be mentioned that frequency predictions for larger molecules will be accurate if there exist correlations between theoretical values of the calibration set molecules and those of the molecules whose vibrational frequencies are to be predicted. Appropriate planning of the calibration set molecules seems able to insure that adequate correlations are present in the data sets. Also more sophisticated multivariate statistical techniques, such as partial least squares (PLS) regression, might provide even more accurate results than PCR.The authors thank FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) and CNPq (Conselho Nacional de Pesquisa) for partial financial support. AEO acknowledges a doctoral fellowship from CNPq.

References

[1] J. Aldous, I.M. Mills, Spectrochim. Acta. 18 (1962) 1073.
[2] J. Overend, J. Scherer, J. Chem. Phys. 32 (1960) 1296, 1296, 1720.
[3] J. Overend, J. Scherer, J. Chem. Phys. 33 (1960) 1720.
[4] P. Pulay, G. Fogarasi, J.E. Boggs, J. Chem. Phys. 74 (1981)

3999; G. Fogarasi, P. Pulay, J. Mol. Struct. 39 (1977) 275; C.E. Blom, P.J. Sinqevland, C. Altona, Mol. Phys. 11 (1976) 1359.
[5] A.L.M.S. Azevedo, B.B. Neto, I.S. Scarminio, A.E. Oliveira, R.E. Bruns, J. Comp. Chem. 17 (1996) 167.
[6] R.E. Bruns, P.H. Guadagnini, I.S. Scarminio, B.B. Neto, J. Mol. Struct. (Theochem) 394 (1997) 197.
[7] J.B.P. da Silva, M.N. Ramos, R.E. Bruns, Spectrochim. Acta A 53 (1997) 733.
[8] J.B.P. da Silva, M.N. Ramos, R.E. Bruns, Spectrochim. Acta A 53 (1997) 1563.
[9] G.E.P. Box, W.G. Hunter, J.S. Hunter, Statistics for Experimenters, Ch. 10–13, Wiley, New York, 1978.
[10] M.J. Frisch, G.W. Trucks, H.B. Schlegel, P.M.W. Gill, B.G. Johnson, M.A. Robb, J.R. Cheeseman, T. Keith, G.A. Petersson, J.A. Montgomery, K. Raghavachari, M.A. Al-Laham, V.G. Zakrewski, J.V. Ortiz, J.B. Foresman, J. Cioslowski, B.B. Stefanov, A. Nanayakkra, M. Challacombe, C.Y. Peng, P.Y. Ayala, W. Chen, M.W. Wong, J.L. Andres, E.S. Replogle, R. Gomperts, R.L. Martin, D.J. Fox, J.S. Binkley, D.J. Drfrees, J. Baker, J.P. Stewart, M. Head-Gordon, C. Gonzalez J.A. Pople, Gaussian 94, Revision D.2, Gaussian Inc., Pittsburgh, PA, 1995.
[11] N.J. Fuke, P. Lockett, J.K. Thompson, P.M. Wilt, J. Mol. Spectrosc. 58 (1975) 87.
[12] R.W. Kirk, P.M. Wilt, J. Mol. Spectrosc. 58 (1975) 102.
[13] S. Kondo, T. Nakanaga, S. Saeki, J. Chem. Phys. 73 (1980) 5409.