



Error Analysis in Determining Parameters of Overlapping Peaks Using Separation of Complex Spectra into Individual Components. Case Study: Doublets

J. Dubrovkin

Computer Department, Western Galilee College, 2421 Acre, Israel

Abstract: Errors in determining peak parameters of the huge sets of Gaussian doublets using nonlinear least squares curve fitting have been evaluated. Doublets were perturbed by normal noise (additive and proportional) and by constant background. Fitting of original spectrum (the normal mode) and of its derivative (the derivative mode) was performed using Gauss-Newton with Levenberg-Marquardt modifications algorithm. The probability that the relative error in estimating each model parameter is not greater than a priori given limit for a given fitting error has been calculated. It was shown that using the derivative mode for curve fitting has advantage over the normal mode only in the case of non-modelled background. We found that the errors in determining maximum peak positions of doublet are the most sensitive to the noise. Increasing of the additive noise causes increasing of the relative number of the good-fitting models both in the normal and the derivative modes; however, the probability decreases.

Keywords: spectrochemistry, overlapping peaks, curve fitting, Gauss-Newton algorithm, errors of peak parameters

I. Introduction

Mathematical separation of overlapping analytical signals (peaks) in spectroscopy and chromatography becomes routine in physicochemical research labs and in the process analytical technology [1, 2]. Numerous commercial user-friendly software tools (solvers) have been developed for evaluating parameters of overlapping peaks (Appendix). These solvers allow researchers to find the best fitting model for experimental data set and to estimate the confidence intervals for parameters of resolved peaks. These estimations are based on linear Taylor expansion of the functional relationship of the nonlinear model [3].

Analysis of the total error in determining parameters of unresolved peaks was performed theoretically by [4]. The error in peak parameters, obtained by the curve-fitting of single Raman peak to the Gaussian shape model, was estimated using analytical calculations and Monte Carlo simulations [5]. The impact of the instrumental factors (number of points and background noise) on the accuracy of the estimated parameters was studied. The errors have been also evaluated numerically using huge data sets of simulated spectra (Gaussian triplets and quartets) [6, 7]. In the last case, the probability (Ψ_{gf}), that the relative error in estimating peak parameter will be not greater than *a priori* given value for a given fitting errors, was considered as a useful measure of the effectiveness of the curve fitting algorithm [7]. It was found that Ψ_{gf} depends on the data set size. Accurate probabilities values can be only obtained by time-consuming processing of extremely large data sets of triplets and quartets. As for the doublets, the size of data sets, required for the error analysis, may be decreased because the number of peak parameters of doublet is less than the corresponding number of triplet and quartet. However, data compression is also needed.

Recently, we evaluated the errors in determining the errors of the Gaussian peak parameters using curve fitting algorithm based on the least squares criterion [8]. The upper boundaries of the minimum uncertainties (UB) of the maximum intensity (R), width (w) and maximum peak position (x_0) of Gaussian peaks were found to be: $UB(R) = C_R/\sqrt{w}$, $UB(w) = C_w\sqrt{w}/R$ and $UB(x_0) = C_{x_0}\sqrt{w}/R$, respectively, where coefficients C depend on the accepted minimum value of the objective function.

It was demonstrated that the data sets size is significantly reduced if these values are used as increments of peak parameters. On the other hand, the probability Ψ_{gf} , obtained from compressed data sets, was shown to be accurate.

In this paper we present an analysis of the errors in determining peak parameters of overlapping peaks using the least squares decomposition of complex spectra into individual components. Decomposition was carried out via the Gauss-Newton with Levenberg-Marquardt modifications algorithm. The spectra were simulated by the 667,106 combinations of two Gaussian peaks which parameters are varying in a very large range. We showed that this number of the combination is optimal in given ranges of doublet parameters because further decreasing of the peak parameters increments do not change the probability Ψ_{gf} . The doublets were perturbed by additive and proportional normal noise and by constant background.

For randomization each numerical experiment was repeated 20 times with regenerated noise and background. The results, that do not meet the constraints, were rejected. The mean relative errors were calculated. In what follows, for the sake of simplicity, term “peak” is used instead of terms “line and band”. The standard algebraic notations are used throughout the article. All calculations were performed and the plots were built using the MATLAB software. Detailed description of the curve-fitting algorithms can be found in [3] and in MATLAB documentation [9].

II. Computer Modeling

Consider decomposition of Gaussian doublet:

$$Q(x) = R_1 \exp\{-(x + x_{01})/r_1\}^2 + R_2 \exp\{-(x - x_{02})/r_2\}^2, \quad (1)$$

where $x = 2\sqrt{\ln 2}j$, $j = [-4:0.05:6]$, x_{0i} , R_i and r_i are the maximum position, the intensity and the width of the i^{th} peak, respectively.

Parameter values of doublets are listed in Table 1. Total number of doublets $N_{total} = 667,106$. Selection of the optimal decrements of x_{01} , x_{02} and r_2 was performed using expressions given in Introduction (see [7]; we hope to publish mathematical details as soon as possible).

Table 1. Parameter values

Parameter	x_{01}, x_{02}	R_1	R_2	r_1	r_2
Values	[0.2, 1.5]	1	0.2:0.05:5	1	[1/3, 3]

Curve fitting was performed using the Gauss-Newton algorithm with Levenberg-Marquardt modifications (MATLAB function *nlinfit*). The initial values of parameters were uniformly distributed random variables in the wide range $[0.3, 1.7] * P_i^{true}$:

$$P_i^{init} = (0.3 + 1.4\eta)P_i^{true} \quad (2)$$

where P_i^{true} is a true parameter value; η is a random variable uniformly distributed in the range 0-1.

Additive (v_{ad}) and proportional (v_{prop}) noise was added to the doublet (1):

$$v_{ad} = \varphi_{ad}(\sigma), \quad (3)$$

$$v_{prop}(x) = \varphi_{prop}(\sigma)Q(x), \quad (4)$$

where $\varphi(\sigma)$ is the normally distributed random variable with zero mean and standard deviation σ .

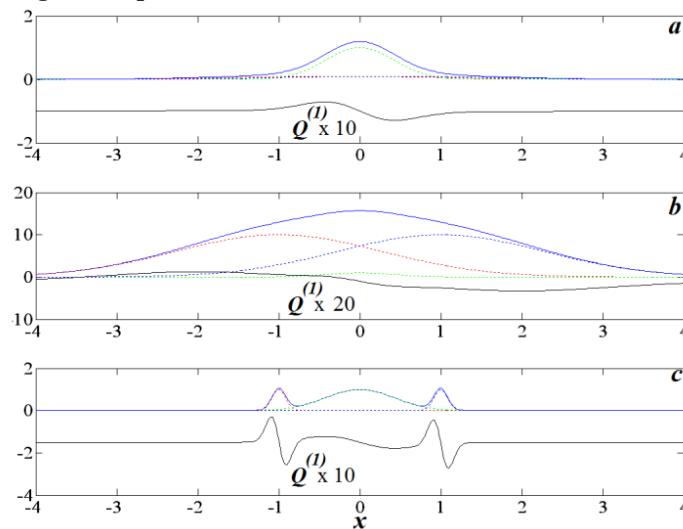
Doublets were also perturbed by constant background ($a_{bkgr} = const$):

$$Bkgr = a_{bkgr}\varphi_{bkgr}(\sigma), \quad (5)$$

where $\sigma(\varphi_{bkgr}) = 1$.

For randomization doublet parameters of the same doublet were estimated 20 times; each time the doublet was perturbed by newly generated noise and background. All of the peak coefficients in Eq. (1) have been constrained to be positive numbers. In the case of the rank deficient program warnings of the Gauss-Newton algorithm the obtained data was rejected.

Figure 1. Spectral doublets and their first-order derivatives



$[x_{01}, R_1, r_1, x_{03}, R_3, r_3] = [-0.3, 0.1, 3, 0.3, 0.1, 3]$ (a), $[-1, 10, 3, 1, 10, 3]$ (b) and $[-1, 1, 0.2, 1, 1, 0.2]$ (c). The fitting error:

$$\xi_{fit} = \sqrt{\sum_{j=1}^{201} (Q(x_j) - \hat{Q}(x_j))^2 / 201}, \quad (6)$$

where $\hat{Q}(x_j)$ is obtained from Eq. 1 by substituting the mean values ($\bar{P}_i^{estim}$) of the estimates of the doublet parameters (P_i^{estim}).

The indicator for good model fit is

$$\xi_{fit} \leq K_{noise} \sigma, \quad (7)$$

where K_{noise} is the noise multiplier. The number of the good-fitting models $N_{gf} < N_{total}$. The relative number (%):

$$\theta_{gf} = 100N_{gf}/N_{total} \quad (8)$$

The relative error (%) in estimating parameter P_i is equal to

$$\xi_{P_i} = 100(\bar{P}_i^{estim} - P_i^{true})/P_i^{true} \quad (9)$$

From N_{gf} models only those doublets for which

$$\xi_{P_i} \leq K_\xi, \quad (10)$$

have been chosen. $K_\xi = [20, 10, 5, 1]$ for all P_i , except P_{x02} for which $K_\xi = [0.05, 0.025, 0.01, 0.002]$.

Number of selected triplets (N_{sel}) depends on the coefficients K : $N_{sel}(K_{noise}, K_\xi)$. The relative value (%):

$$\Psi_{gf} = 100N_{sel}(K_{noise}, K_\xi)/N_{gf} \quad (11)$$

may be considered as the probability that the relative error in estimating parameter P_i is not greater than K_ξ for a given ξ_{fit} over good-fitting models [7].

III. Results and Discussion

Table 2 represents the relative number of the good-fitting models θ_{gf} (8) and the statistical parameters (mean, two-fold standard deviation and skewness) of the error distributions. We conclude that:

1. Increasing of the standard deviation of the additive normal noise causes increasing θ_{gf} in the both modes due to the widening ranges of possible solutions (fittings). However, increasing of the proportional noise does not demonstrate this effect, possibly because this noise perturbs significantly only the central, most intense part of the peak.

2. If spectra are perturbed by the proportional noise then θ_{gf} decreases significantly compared to the additive spectral noise, especially in the derivative mode. Constant background further decreases the θ_{gf} -values. E.g., $\theta_{gf}^{D1}(K_{noise} = 2) = 0.033\%$.

3. The first moments of the error distribution (μ_1) for the additive noise do not differ significantly from zero both in the normal and the derivative modes. The absolute values of μ_1 increase when the standard deviation of the noise is changed from 0.01 to 0.05. Constant background increases $|\mu_1|$ (R_1) and $|\mu_1|(r_1)$ in the normal mode for additive and proportional noise. Differentiation decreases $|\mu_1|$ because the constant background is suppressed. The first moments obtained in case of proportional noise are approximately equals in both modes.

4. The skewness of the error distributions does not reflect any systematic dependence on the noise and doublet parameters. The anomaly high absolute values of the skewness are coloured red in Table 2.

The relative error probability is a very complex function $\Psi_{gf}(P_i, lim, K_{noise}, K_\xi)$. Tables 3 represents some numerical values of this function in the normal and the derivative mode: Ψ_{gf}^{D0} and Ψ_{gf}^{D1} , respectively.

These values are accurate because they were estimated for the optimal compressed set of doublet parameters.

Contrary to the studies [6, 7] dedicated to the estimation of the errors in determining parameters of triplet and quartet peaks, Ψ_{gf} -values obtained in this paper, cannot be refined by further decreasing of the parameter increments both in the normal and the derivative modes.

From Table 3 the following conclusions can be drawn:

1. The last four rows demonstrate high reproducibility of the estimation of the probability Ψ_{gf} .

2. Ψ_{gf} decreases in both modes while the standard deviation of the additive noise increases.

3. Probability Ψ_{gf} increases while increasing K_ξ .

4. If there is no background then usually $\Psi_{gf}^{D0} > \Psi_{gf}^{D1}$.

5. If there is a constant background then for $K_\xi = 5$ $\Psi_{gf}^{D0} < \Psi_{gf}^{D1}$. However, if $K_\xi = 1$ then for $\sigma_{ad} = 0.01$, $\sigma_{prop} = 0$, $K_{bkg} = 0.01$ only $\Psi_{gf}^{D0}(x_0) < \Psi_{gf}^{D1}(x_0)$; for $\sigma_{ad} = 0.002$, $\sigma_{prop} = 0.1$, $K_{bkg} = 0.01$ $\Psi_{gf}^{D0}(R_1) > \Psi_{gf}^{D1}(R_1)$ and $\Psi_{gf}^{D0}(r_1) > \Psi_{gf}^{D1}(r_1)$.

6. The most probabilities which satisfied the condition $\Psi_{gf} < 0.95 * \bar{\Psi}_{gf}$, are $\Psi_{gf}(x_{01})$ and $\Psi_{gf}(x_{02})$ for both types of the noise. For the proportional noise only the probabilities of the first peak parameters satisfy the above

Table 2. Statistical parameters of the error distributions

σ_{ad}	σ_{prop}	K_{bkgr}	Mode	θ_{gf}	Parameters of doublets					
					x_{01}	R_1	r_1	x_{02}	R_2	r_2
0.01	0	0	Normal	30.6	-0.51± 0.006	0	0	-0.50± 0.006	0	-0.10± 0.001
					-1.2	3.7	2.4	-1.1	-7.7	-2.6
			Derivative	11.8	-0.59± 0.01	-0.79± 0.01	0.74± 0.01	-0.57± 0.01	-0.69± 0.01	0.35± 0.008
					-0.97	-1.1	1.1	-0.93	-1.7	2.4
Normal			45.1	-0.72± 0.007	0	0	-0.71± 0.007	0	-0.14± 0.001	
				-1.2	0.49	0.13	-1.1	-6.1	-2.3	
Derivative			21.4	-1.12± 0.017	-0.72± 0.009	0.68± 0.008	-1.09± 0.016	-0.84± 0.011	0.30± 0.007	
				-1.2	-1.1	1.1	-1.2	-1.6	2.1	
Normal			67.8	-1.20± 0.010	0	0	-1.18± 0.010	-0.12± 0.002	-0.20± 0.002	
				-1.2	-0.24	-0.63	-1.24	-4.4	-2.5	
Derivative			39.0	-1.70± 0.020	-0.68± 0.000	0.37± 0.007	-1.75± 0.020	-1.96± 0.009	0.10± 0.006	
				-1.2	-1.3	0.55	-1.2	-1.5	0.69	
0.01	0	0.01	Normal	11.8	-0.13± 0.01	2.2± 0.03	2.9± 0.05	0	0	0.20± 0.01
					-2.0	1.0	1.0	-0.44	-1.6	-0.15
			Derivative	12.0	-0.61± 0.01	-0.79± 0.01	0.74± 0.01	-0.59± 0.01	-0.68± 0.01	0.34± 0.008
					-0.99	-1.1	1.1	-0.96	-1.7	2.4
0.002	0.05	0	Normal	5.9	-0.26± 0.010	0	0	-0.11± 0.005	0	-0.11± 0.004
					-1.1	1.3	-0.27	-0.91	-0.19	-1.5
			Derivative	0.33	-0.28± 0.043	-0.65± 0.074	0.26± 0.050	-0.32± 0.060	-0.27± 0.037	0.14± 0.033
					-1.3	-1.3	0.92	-1.4	-1.9	1.5
			Normal*	19.4	-0.63± 0.008	0.15± 0.004	0	-0.46± 0.006	0	-0.18± 0.003
					-1.3	2.1	0.1	-1.1	-2.1	-1.8
			Derivative*	8.2	-1.11± 0.026	-0.52± 0.020	0.12± 0.018	-0.77± 0.021	-0.69± 0.016	0.23± 0.011
					-1.2	-0.97	-0.1	-1.2	-1.7	1.6
0.002	0.10	0	Normal	1.7	-0.25± 0.012	0	0	0	0	-0.12± 0.010
					-1.2	0.61	-0.31	-0.27	0.42	-1.3
			Derivative	0.033	-0.26± 0.12	-0.53± 0.17	0.21± 0.15	-0.25± 0.16	-0.35± 0.14	0.19± 0.14
					-1.8	-1.3	0.84	-2.0	-2.2	1.4
			Normal*	26.9	-0.81± 0.011	0.31± 0.006	0.12± 0.007	-0.39± 0.006	-0.10± 0.004	-0.34± 0.005
					-1.3	-1.8	0.50	-0.95	-0.85	-1.7
			Derivative*	3.5	-1.35± 0.046	-0.55± 0.030	0.21± 0.032	-0.79± 0.032	-0.93± 0.032	0.15± 0.022
					-1.2	-0.96	0.33	-1.3	-1.4	0.74
0.002	0.10	0.01	Normal*	11.5	-0.39± 0.013	2.21± 0.03	2.90± 0.05	0	-0.19± 0.010	0.13± 0.010
					-1.8	1.1	1.1	-0.30	-1.6	-0.20
			Derivative*	3.4	-1.41± 0.048	-0.54± 0.030	0.22± 0.033	-0.82± 0.034	-0.96± 0.034	0.14± 0.022
					-1.2	-0.95	0.37	-1.3	-1.4	0.66

$\mu_1 \pm 2\mu_2^{1/2}$ and $\gamma_{as} = \mu_3/\mu_2^{3/2}$ (skewness) are given in the upper and lower rows, respectively.
 μ_i is the i^{th} -order central moment. $K_{noise} = 2$ and 10 (*).

Table 3. Relative error probability (Ψ_{gf})

σ_{ad}	σ_{prop}	K_{bkg}	Mode	Parameters of doublets						$\bar{\Psi}_{gf}$
				x_{01}	R_1	r_1	x_{02}	R_2	r_2	
0.01	0	0	Normal	99.2	99.9	100.0	99.4	100.0	100.0	99.8
				75.0	99.3	99.7	76.7	99.6	99.7	91.7
			Derivative	96.9	99.6	100.0	97.3	99.2	99.1	98.7
				42.1	82.1	83.5	43.0	79.0	83.6	68.9
0.02			Normal	96.5	99.5	99.9	99.9	99.9	99.9	99.3
				50.2	97.4	98.2	53.2	98.7	98.1	82.6
			Derivative	90.0	97.3	99.7	92.1	98.1	98.4	95.9
				24.9	68.6	69.6	25.4	73.0	75.9	56.2
0.05			Normal	85.5	97.9	99.4	87.0	99.4	99.9	94.9
				30.2	88.8	86.8	35.1	95.7	95.6	72.0
			Derivative	63.4	89.7	96.7	68.8	97.0	98.1	85.6
				12.2	48.9	50.5	13.2	66.7	66.8	45.1
0.01	0	0.01	Normal	77.9	88.3	70.8	99.0	92.4	98.3	87.8
				62.7	2.2	1.4	89.1	65.2	59.7	46.7
			Derivative	97.1	99.7	100.0	97.6	99.3	99.1	98.8
				41.9	82.3	83.7	42.7	79.3	83.8	69.0
0.002	0.05	0	Normal	99.1	100.0	100.0	99.9	100.0	100.0	99.8
				80.3	94.1	88.8	96.7	97.5	97.5	92.5
			Derivative	99.9	99.6	100.0	98.5	99.7	99.9	99.6
				80.1	58.2	60.5	66.0	89.9	87.1	73.6
			Normal*	95.8	98.5	99.5	99.1	99.8	100.0	98.8
				60.7	85.7	81.3	75.7	94.6	95.0	82.2
			Derivative*	90.1	88.0	87.4	93.0	97.3	98.5	92.4
				28.8	39.5	35.6	35.8	65.5	70.5	46.0
0.002	0.10	0	Normal	99.3	100.0	100.0	99.8	99.9	100.0	99.8
				85.1	89.8	83.7	94.9	95.7	94.6	90.6
			Derivative	99.5	100.0	100.0	99.5	100.0	99.1	99.7
				85.9	70.5	64.1	82.7	87.7	79.1	78.3
			Normal*	92.3	96.5	97.6	98.7	99.7	99.9	97.5
				51.5	70.1	62.0	74.3	79.9	81.2	69.8
			Derivative*	87.1	90.8	82.8	91.9	93.6	96.8	90.5
				25.3	37.4	29.5	37.0	49.5	50.2	38.2
0.002	0.10	0.01	Normal*	79.6	89.7	68.4	99.0	96.4	98.1	88.5
				54.5	5.2	3.1	85.9	62.5	56.1	44.6
			Derivative* Duplicate estimations	86.8	90.3	81.8	92.0	93.0	96.6	90.1
				(86.5)	(90.7)	(82.0)	(91.6)	(93.6)	(96.8)	(90.2)
				24.0	36.6	29.1	36.0	48.3	49.4	37.2
				(24.5)	(36.9)	(29.2)	(36.5)	(47.8)	(49.9)	(37.5)

$K_{\xi} = 5$ and 1 from upper to lower rows, respectively. $K_{noise} = 2$ and 10 (*). Values $\Psi_{gf} < 0.95 * \bar{\Psi}_{gf}$ are given in bold italic.

condition. The last result is qualitatively explained as follows. If intensity of the second doublet peak increases ($R_2 \geq 0.5$) then the proportional noise of more intensive doublet spectra also increases. The region of the first peak will be noisier. Therefore, the errors in determining parameters of this peak increase. Because the number of doublets with relative intensity of the second peak $R_2 \geq 0.5$ is more than $0.5N_{total}$, the relative error probability Ψ_{gf} for the first peak decreases.

7. Non-modelled background significantly decreases the probability of the accurate determination of the overlapping peak parameters. The probabilities $\Psi_{gf}^{D0}(R_1)$ and $\Psi_{gf}^{D0}(r_1)$ are the most sensitive to this background.

Conclusion

Error analysis in determining parameters of overlapping doublet peaks using least squares curve fitting showed that the probability Ψ_{gf} is the useful measure of the effectiveness of the curve fitting algorithm. In practice Ψ_{gf} can be increased by adding normal noise to measured spectra. This is not a new spectroscopy method. E.g., addition of noise to the input training data of an artificial neural network is equivalent to Tikhonov regularization of multivariate calibration [10].

We demonstrated that use of the first-order derivative for suppression of constant background decreases the probability Ψ_{gf} . It seems that data fitting to the peaks with a low-order polynomial background, gives more meaningful background subtraction [11].

References

- [1] K. Danzer, Analytical Chemistry. Theoretical and Metrological Fundamentals, Springer, 2008.
- [2] A. Bogomolov, "Multivariate process trajectories: capture, resolution and analysis", Chemometrics and Intelligent Laboratory Systems, vol. 108, 2011, pp. 49–63.
- [3] G. A. F. Seber and C. J. Wild, Nonlinear Regression, Hoboken, NJ: Wiley-Interscience, 2003.
- [4] V.A. Lorenz-Fonfr, E. Padros, Curve-fitting overlapped bands: quantification and improvement of curve-fitting robustness in the presence of errors in the model and in the data, Analyst, vol. 129, 2004, pp. 1243 – 1250.
- [5] W. M. van Spengen, "The accuracy of parameter estimation by curve fitting in the presence of noise", J. Appl. Phys., vol. 111, 2012, (054908) pp. 1-7.
- [6] J. Dubrovkin, "Error Analysis in Determining Parameters of Overlapping Peaks Using Separation of Complex Spectra into Individual Components. Case Study: Triplets" , International Journal of Emerging Technologies in Computational and Applied Sciences, vol. 1-12, 2015, pp.1-7.
- [7] J. Dubrovkin, "Error Analysis in Determining Parameters of Overlapping Peaks Using Separation of Complex Spectra into Individual Components. Case Study: Quartets" , International Journal of Emerging Technologies in Computational and Applied Sciences, vol. 2-12, 2015, pp. 174-179.
- [8] J. Dubrovkin, "Evaluation of minimum uncertainties in determining parameters of analytical signal consisting of overlapping peaks", The European Conference on Analytical Chemistry (EuroAnalysis-2015), Bordeaux, France.
http://www.mathworks.com
- [9] C. M. Bishop, "Training with noise is equivalent to Tikhonov regularization", Neural Computation vol.7, 1995, pp. 108-116.
- [10] P. Pelikan, M. Ceppan, M. Liska, Applications of Numerical Methods in Molecular Spectroscopy, CRC , 1994.

APPENDIX

Commercial software tools for peak separation

Tool	Year	URL
Casa XPS. Peak Fitting in XPS	2006	www.casaxps.com/help_manual/manual_updates/peak_fitting_in_xps.pdf
Fityk manual. Release 1.2.9	2014	www.apapes.physics.uoc.gr/references/fityk-1.2.9-manual.pdf
WaveMetrics. Igor Pro 6	2014	www.wavemetrics.com
OriginLab. Peak Fitting Module	2015	www.OriginLab.com
Spectral Works. AnalyzerPro	2013	www.spectralworks.com
CAMO. The Uncrambler	2015	www.camo.com
Chroma TOF. Software for Separation Science	2013	www.leco.com
Bruker OPUS. Spectroscopic Software. Reference Manual	2004	www.shaker.umh.es/investigacion/OPUS_script/OPUS_5_BasePackage.pdf
GRAM Suite. Software Tools for Spectroscopists	2004	www.thermo.com.cn/Resources/200802/productPDF_24179.pdf