



Separation of Overlapping Doublet Peaks: Comparison of Curve Fitting Algorithms. Constant Additive, Proportional and Correlated Noise

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 estimated. Decomposition of overlapping peak was carried out by 6 algorithms. Probability that the relative error in estimating peak parameters is not greater than a given value over good-fitting models was estimated. We found that the Gauss-Newton with the Levenberg-Marquard modification and the Broyden-Fletcher-Goldfarb-Shanno methods are the best algorithms for deconvolution of overlapping doublet peaks. It was shown that high probability in determining accurate peak parameters is achieved by rejection significant number of badly fitting doublets.

Keywords: spectrochemistry, overlapping peaks, curve fitting algorithms, errors of peak parameters

I. Introduction

Mathematical separation of overlapping signals (peaks) in analytical chemistry and physics is widely used in research labs and in the process analytical technology [1-3]. Numerous commercial user-friendly software tools (solvers) are available for evaluating parameters of overlapping peaks (see Appendix to the paper [4]).

A variety of nonlinear optimization methods [5], used to solve separation problem, can be divided into four main groups:

1. Gradient methods based on the evaluation of the model derivatives with respect to the model parameters during the minimization process (e.g., the Newton-Raphson method, the Gauss-Newton method (including the Levenberg-Marquard modification - GNLM), the Interior-Reflective Newton method (IRN) [6], the Steepest Descent method (SD), the Broyden-Fletcher-Goldfarb-Shanno (BFGS) and the Davidon-Fletcher-Powell (DFP) methods. The last three algorithms use gradient information and named the quasi-Newton methods since they approximate Newton's method. The effectiveness of this group of algorithms depends largely upon the accuracy of calculation of the derivatives.
2. The Nelder-Mead Simplex method (NMS). Simplex method is stable and effective for problems with a large number of parameters. However, NMS does not work well for spectra decomposition, because it finds only local minima [7].
3. The random search methods [7]. These methods guarantee finding the minimum, even if a large number of parameters is involved. Such algorithms allow substantial modifications; they are flexible and adaptable for a variety of problems.
4. The nature-inspired methods (e.g., genetic algorithms (GN), immune method, artificial neural networks (ANN)).

Effectiveness of algorithms of the separation of overlapping peaks has been carefully discussed. Some examples will be given below.

The pioneering research was a comparison study of 7 iterative gradient methods for searching for best-fit model of simulated and measured spectrum based on the least squares criterion [8]. One of the central problems - convergence of these methods to the global minimum has been studied.

Simultaneous fitting of a measured spectrum and a set of its derivatives (up to 8th order) was performed by [9]. The best fit model was obtained by varying the parameters of the model bands visually. Mathematical details of the fitting procedure were absent.

Fitting of Raman scattering and infrared absorption spectra by Gaussians was performed by GN supplemented with the gradient descent method and by the simplex method (NMS) [10]. Higher derivatives were used to determine extrema of the initial spectrum. The authors stated that NMS is more sensitive to falling into a local minimum in comparison with the GN method, but is faster than GN. In some case the relative differences of peak positions and widths obtained by two methods were more than 5-10%.

Fourier deconvoluted spectra and derivative spectra were fitted to complex peak models [11]. It was found that for a noise-free synthetic spectrum the accuracy of decomposition is similar to that obtained by curve-fitting of the original spectra. When the spectrum contains noise and the model is not correct direct curve-fitting can

converge to a biased solution. Curve-fitting of deconvoluted or derivative spectra improves robustness if good peak shapes are used for modeling.

Peak detection was performed using ANN and global search technique [12]. This technique does not need accurate initial estimates and is less sensitive to local optima than SD. While fitting spectrum consisting of many strongly overlapping peaks, SD can fail if initial estimates of the fit parameters are far from the correct values. SD works well only under highly constrained conditions which were imposed without any acceptable physical background. These constraints are only used to improve the convergence. The formulation of the constraints was a time-consuming task.

Since GN is little sensitive to local optima this algorithm is useful in determining band parameters of a high-dimensional problem within reasonable time.

GN and IRN based on minimizing a smooth nonlinear multivariable function, subject to upper and/or lower bounds on some of the variables were used by [13]. The spectrum was modelled by six Gaussians disturbed by the low-frequency fractal noise. Peak numbers and initial peak positions were estimated by continuous wavelet analysis. It was found that GN-decomposition of the spectrum was stable in the wide range of the noise level and structure.

Fitting several spectra was performed simultaneously by the random search algorithm [7]. This algorithm is based on the generation of the set of randomly distributed multidimensional points. For optimal set the object function is minimal. Generation process is repeated each time with a new distribution.

In our opinion essential drawback of all researches was a small number of measured or simulated spectra used to test curve-fitting algorithms. Therefore, the findings are unlikely to be common.

Recently we performed error analysis in determining parameters of overlapping peaks using more than one million of simulated spectra (Gaussian doublets, triplets and quartets) [4, 14-16]. Although it is not possible to mimic all real-life situations, these models allowed studying most practical cases of overlapping peaks encountered in spectra.

The new feature of our analysis was evaluation of the probability that the relative error in estimating each peak parameter will be less than some predefined value for a given fitting error. Separation of overlapping peaks has been performed using GNLM. Triplets were also decomposed by GN [14]. All spectra were disturbed by a constant additive noise. Proportional noise was also added to doublets. However, the correlated noise has not considered. Consideration of the impact of this noise on the peak separation problem is of a great importance because "successive measurements "remember" each other at least their nearest neighbours" (cited from [17]). In the present paper we set out to investigate the problem of separation doublets disturbed by constant additive, proportional and correlated noise (AR(1)) using GNLM, GN, IRN, BFGS, DFP and SD algorithms [18]. These algorithms are implemented as built-in functions of the MATLAB software.

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.

II. Computer Modeling

Consider decomposition of Gaussian doublet:

$$Q(x) = R_1 \exp\{-(x + x_0)/r_1\}^2 + R_2 \exp\{-(x - x_0)/r_2\}^2, \quad (1)$$

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_0 and r_2 was performed according to [16].

Table 1. Parameter values

Parameter	x_0	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 GNLM, GA, IRN, BFGS, DFP and SD algorithms (MATLAB function *nlinfit*, *ga*, *fminsearch* and *fminunc* (for last four algorithms), respectively).

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}), proportional (v_{prop}) and correlated noise (v_{cor}) which has the AR(1) structure [18] was added to the doublet (1):

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

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

$$v_{cor}(j) = k^2 v_{cor}(j-2) + k v_{cor}(j-1) + \varphi_{ad}(\sigma), \quad (5)$$

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

$j > 2$, $v_{cor}(1) = \varphi_{ad}(\sigma)$, $v_{cor}(2) = kv_{cor}(1) + \varphi_{ad}(\sigma)$, the coefficient $k \ll 1$.

For randomization doublet parameters of the same doublet were estimated 20 times; each time the doublet was perturbed by newly generated noise. 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 GNLM obtained data was rejected. The number of decomposed doublets $N_{dec} < N_{total}$.

The relative number (%):

$$\theta_{dec} = 100N_{dec}/N_{total} \quad (6)$$

The fitting error:

$$\xi_{fit} = \sqrt{\sum_{j=1}^{201} (Q(x_j) - \hat{Q}(x_j))^2 / 201}, \quad (7)$$

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

$$\xi_{fit} \leq K_{noise}\sigma, \quad (8)$$

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

$$\theta_{gf} = 100N_{gf}/N_{dec} \quad (9)$$

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

$$\xi_{P_i} = 100(\bar{P}_i^{estim} - P_i^{true})/P_i^{true} \quad (10)$$

From N_{gf} models only those doublets for which

$$\xi_{P_i} \leq K_\xi, \quad (11)$$

have been chosen. $K_\xi = [20, 10, 5, 1]$.

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

$$\psi_{gf} = 100N_{sel}(K_{noise}, K_\xi)/N_{gf} \quad (12)$$

was 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.

III. Results and Discussion

Statistical parameters of the error distributions are given in Table 2. The maximum signal-to-noise ratio of doublets was about 50-100 depending on the noise type. It is seen that in all cases the absolute value of the mean of the relative error distribution was less than 1%. In the case of the constant and correlated noise distributions of r_1 and R_2 for GNLM and of r_1, r_2 and R_2 for IRN, BFGN and DFP were significantly asymmetrical. However, generally speaking, no clear pattern of the skewness has been identified.

It was found (Table 2) that the relative number of decomposed doublets (θ_{dec} , Eq. (6)) is very close to 100% for GNLM, IRN and BFGS; for GN and DFP $\theta_{dec} \approx 55 - 65\%$. Since for SD θ_{dec} is very small (less than 3%) condition (11) is not satisfied for all doublets which were decomposed into separated peaks by this algorithm. Therefore SD is not applicable to the decomposition problem solving (corresponding data is absent in Tables 2 and 3).

Computer modeling showed that the relative number of the good fitting doublet models is low than $\theta_{gf} < 9\%$ for all algorithms in the case of proportional noise; for GN the relative number $\theta_{gf} < 2\%$. For DFP $\theta_{gf} < 6\%$ for all types of the noise.

For all models (except GN) the mean probability $\bar{\psi}_{gf} > 90\%$ for $K_\xi = 1$ (Table 3). For GN $\bar{\psi}_{gf} > 59\%$. However, as was pointed above, this high probability in determining accurate peak parameters is achieved by rejection significant number of badly fitting doublets (e.g., for proportional noise more than 90%). Therefore the performance of the curve fitting process may be improved at the cost of growth of permissible uncertainty of obtained peak parameters.

IV. Conclusion

Peak separation of the huge dataset of simulated doublets showed that the best algorithms for deconvolution of overlapping peaks are GNLM and BFGS. The drawback of DFP is a small relative number of the good fitting models. We found that the results obtained by SD have no informational value. Therefore, SD is not applicable to the decomposition problem solving. It turned out that the GN is the most time-consuming algorithm.

We found that the high probability in determining accurate peak parameters (the relative errors less than 1%) has been achieved by rejection significant number of badly fitting doublets. Therefore effectiveness of the curve fitting process may be improved at the cost of growth of permissible uncertainty of obtained peak parameters.

Error analysis in determining peak parameters of overlapping peaks must be an integral part of the separation procedures.

Table 2. Statistical parameters of the error distributions ($K_{noise} = 2$)

Algorithm	Noise	Parameters of doublets						$\theta_{dec}, \%$	$\theta_{gf}, \%$
		x_{01}	R_1	r_1	x_{02}	R_2	r_2		
GNLM	Constant	-0.51± 0.006	0	0	-0.50± 0.006	0	-0.10± 0.001	>99.9	30.6
		-1.2	3.7	2.4	-1.1	-7.7	-2.6		
	Proportional	-0.26± 0.010	0	0	-0.11± 0.005	0	-0.11± 0.004	>99.9	5.9
		-1.1	1.3	-0.27	-0.91	-0.19	-1.5		
	Correlated	-0.50± 0.006	0	0	-0.50± 0.006	0	-0.10± 0.001	>99.9	30.4
		-1.2	3.0	2.2	-1.1	-6.3	-2.2		
GN	Constant	-0.59± 0.010	-0.25± 0.008	0.86± 0.014	0.15± 0.006	-0.32± 0.007	0.17± 0.007	55.1	87.0
		-1.7	-0.71	1.3	0.71	-1.3	0.60		
	Proportional	-0.12± 0.021	0	0.23± 0.031	0	-0.17± 0.030	0.42± 0.044	55.9	1.8
		-0.74	-1.0	0.76	-0.72	-1.1	1.8		
	Correlated	-0.59± 0.01	-0.25± 0.009	0.84± 0.01	0.15± 0.006	-0.32± 0.007	0.17± 0.007	55.0	86.8
		-1.7	-0.69	1.3	0.74	-1.3	0.62		
IRN	Constant	-0.21± 0.002	0	-0.14± 0.002	-0.24± 0.002	0	0	>99.9	60.3
		-1.5	0	-5.9	-2.2	-6.1	-9.0		
	Proportional	-0.14± 0.050	0	0	0	0	0	>99.9	8.6
		-1.8	0.41	-0.40	-0.52	0.09	-1.3		
	Correlated	-0.21± 0.002	0	-0.14± 0.002	-0.24± 0.002	0	0	>99.9	70.5
		-1.5	0.32	-5.8	-2.3	-6.1	-7.5		
BFGS	Constant	-0.21± 0.002	0	-0.14± 0.002	-0.24± 0.002	0	0	>99.9	67.4
		-1.5	0	-5.9	-2.2	-6.9	-7.3		
	Proportional	-0.13± 0.005	0	0	0	0	0	>99.9	8.6
		-1.4	0.38	-0.53	-0.85	0.24	-1.2		
	Correlated	-0.21± 0.002	0	-0.13± 0.002	-0.24± 0.002	0	0	>99.9	69.2
		-1.6	0.42	-5.9	-2.2	-6.3	-7.9		
DFP	Constant	-0.12± 0.010	0	-0.11± 0.008	-0.12± 0.010	0	-0.12± 0.007	64.2	5.7
		-0.13	0.55	-7.6	0	-7.3	-3.4		
	Proportional	0	0	0	0	0	0	65.0	1.7
		-1.4	0.45	-0.66	-1.2	0.25	-1.2		
	Correlated	-0.11± 0.010	0	-0.10± 0.007	-0.12± 0.010	0	-0.14± 0.007	64.2	5.7
		0	1.7	-7.4	0	-6.4	-3.2		

Skewness is given in the las rows

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

Algorithm	Noise	Parameters of doublets						$\bar{\Psi}_{gf}$
		x_{01}	R_1	r_1	x_{02}	R_2	r_2	
GNLM	Constant	75.0	99.3	99.7	76.7	99.6	99.7	91.7
	Proportional	85.1	89.8	83.7	94.9	95.7	94.6	90.6
	Correlated	74.1	99.0	99.5	76.2	99.5	99.6	91.3
GN	Constant	62.5	52.2	34.2	75.0	70.6	65.4	60.0
	Proportional	94.8	87.8	80.0	96.5	83.6	77.2	86.7
	Correlated	62.2	50.0	32.9	74.9	70.2	64.9	59.2
IRN	Constant	94.0	97.8	97.0	94.8	99.5	99.4	97.1
	Proportional	95.1	90.9	90.5	97.7	98.6	98.3	95.2
	Correlated	93.5	97.5	96.9	94.4	99.5	99.4	96.9
BFGS	Constant	93.8	97.8	97.0	94.6	99.5	99.4	97.0
	Proportional	95.3	91.2	90.5	97.7	98.6	98.3	95.3
	Correlated	93.7	97.6	97.0	94.7	99.5	99.4	97.0
DFP	Constant	83.8	99.1	98.3	84.7	99.4	94.9	93.4
	Proportional	99.9	92.7	96.5	99.9	97.3	98.7	97.5
	Correlated	83.0	99.2	98.4	84.0	99.4	95.2	93.2

$K_{noise} = 2, K_{\xi} = 1$. Values $\Psi_{gf} < 0.95 * \bar{\Psi}_{gf}$ are given in bold italic.

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] P. Pelikan, M. Ceppan, M. Liska, Applications of Numerical Methods in Molecular Spectroscopy, CRC , 1994.
- [4] J. Dubrovkin, "Error Analysis in Determining Parameters of Overlapping Peaks Using Separation of Complex Spectra into Individual Components. Case Study: Doublets" , International Journal of Emerging Technologies in Computational and Applied Sciences, vol. 1-13 , 2015, pp. 36-41.
- [5] E.M.T. Hendrix and B. G.-Tóth, Introduction to Global and Nonlinear Optimization, Springer, New York, 2010.
- [6] T.F. Coleman, Y. Li, "On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds", Mathematical Programming, vol. 67 (1994) pp. 189-224.
- [7] T. R. Griffith, D. A. Nerukh and S. A. Eremenko, "The application of theoretical models of complex shape to the fitting of experimental spectra having closely overlapping bands", Phys. Chem. Chem. Phys., vol. 1 (1999) pp. 3199-3208.
- [8] J. Pitha , R. N. Jones, "A comparison of optimization methods for fitting curves to infrared band envelopes", Can. J. Chem., vol. 44 (1966) pp. 3031-3050.
- [9] I. K. Mikhailyuk, H. Lokstein, T. A. P. Razjivin, "A method of spectral subband decomposition by simultaneous fitting the initial spectrum and a set of its derivatives", J. Biochem. Biophys. Methods, vol. 63 (2005) pp. 10–23.
- [10] S. A. Burikov, T. A. Dolenko, I. S. Kurchatov, S. V. Patsaeva, and Yu. V. Starokurov, "Computer analysis of vibrational spectra of aqueous ethanol solutions", Russian Physics Journal, vol. 55 (2012) pp.383-388.
- [11] V. A. Lórenz-Fonfría, E. Padrós, "Curve-fitting of Fourier manipulated spectra comprising apodization, smoothing, derivation and deconvolution", Spectrochim. Acta Part A, vol. 60 (2004) pp. 2703–2710.
- [12] P. De Weijer, C. B. Lucasius, L. Buydens, and G. Kateman, "Curve Fitting Using Natural Computation", Anal. Chem., vol. 66 (1994), pp. 23-31.
- [13] E. P. Shaimuhametova, D. Z. Galimullin, M. E. Sibgatullin, D. I. Kamalova and M. H. Salakhov, "Impact of the noise type on decomposition of complex spectral profiles", Bulletin of the Russian Academy of Sciences: Physics, vol. 76 (2012) pp. 347-349 [Russian].
- [14] 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.
- [15] 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.
- [16] 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.
- [17] R. Nigmatullin, R. Rakhmatullin, "Detection of quasi-periodic processes in repeated measurements: New approach for the fitting and clusterization of different data", Commun. Nonlinear. Sci. Numer. Simulat., vol. 19 (2014) pp.4080–4093.
- [18] G. A. F. Seber, C. J. Wild, Nonlinear Regression, Wiley, 2005.