



Bayesian Estimation of Minimum Uncertainties in Determining Parameters of Analytical Signal Consisting of Overlapping Symmetrical Peaks

J. Dubrovkin
Computer Department,
Western Galilee College, 2421 Acre, Israel

Abstract: We present the results of the error analysis of Bayesian estimates of peak parameters. The analysis was based on the sampling of the posterior distribution. Peak shape was modelled by Gaussian and Lorentzian functions. Obtained results confirmed our previous Least Squares estimates.

Keywords: spectrochemistry, overlapping peaks, curve fitting, Bayesian estimates of the errors of peak parameters

I. Introduction

Mathematical separation of overlapping peaks is one of the most common pre-processing techniques for instrumental analytical chemistry. It necessarily requires error analysis of parameter estimates. Uncertainty in determining peak maximum position using the second order derivatives and deconvolution was estimated by computer modeling [1-3]. For this goal more than one million doublets, triplets and quartets were deconvoluted by gradient and genetic algorithms. The probability that the relative error in estimating each peak parameter will be less than some predefined value for a given fitting error was evaluated.

The upper bounds (UB) of the total error in determining all peak parameters [4] and of the errors in determining maximum peak position (x_0), peak width (w) and peak maximum intensity (A_m) of Gaussians were obtained theoretically using Least Squares Estimation (LSE) combined with Taylor expansion [5]. We found that the errors in determining x_0 and w are proportional to $\sigma\sqrt{w}/A_m$ (σ is the standard noise deviation). The corresponding error of A_m : $UB(A_m) \propto \sigma/\sqrt{w}$.

The disadvantage of the LSE approach is that it requires cumbersome algebraic calculations of derivatives of object function with respect to the peak parameters.

An alternative method to LSE is the Bayesian estimation (BE) based on the estimation of the posterior distribution of unknown parameters of the model of the analytical signal for a given signal [6]. For integration, a sampling of the posterior distribution is usually carried out by Monte Carlo Markov chain (MCMC) which significantly simplified calculations [7]. For MCMC-applications (e.g., using the Metropolis-Hastings algorithm) the distribution may be known up to a constant of proportionality.

Bayesian method was firstly used in instrumental analytical chemistry for blind separation of the analytical signal of a linear mixture model [8]. Mathematical subtleties of this method were taken from signal processing (see references in [8]).

In this report we present the results of the error analysis of BE of peak parameters. The analysis was based on the sampling of the posterior distribution. The advantage of the BE approach to our analysis is that it can be carried out by simple numerical integration of the posterior distribution.

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. Theory

1. LSE-method

Consider non-linear curve fitting performed by minimizing the object function:

$$S(\theta) = \sum_{l=1}^L [F_{ms}(l) - F_{fit}(l, \theta)]^2 \leq \varepsilon^2 L, \quad (1)$$

where $F_{ms}(l)$ is the analytical signal, $l = 1, 2, \dots, L$; the model

$$F_{fit}(l, \theta) = \sum_{i=1}^N f(l, \theta_i) \quad (2)$$

consists of N peaks of the same shape, $\theta (\theta_{i1}, \theta_{i2}, \dots, \theta_{it})$ is the vector of the best-fitting model parameters; t is the number of the shape parameters; $\varepsilon = k_{noise} \sigma$ is a perturbation; k_{noise} is the noise coefficient, and σ stands for standard deviation of the noise. For precise fit, $\theta = \theta_0$ contains the correct parameters.

Assume that the j^{th} parameter of the i^{th} peak is incremented by $\Delta\theta_{ij} \ll 1$, then according to the Taylor approximation:

$$\Delta F_{fit}(l, \theta_0, \Delta \theta_{ij}) \cong \frac{\partial F_{fit}(l, \theta_0)}{\partial \theta_{ij}} \Delta \theta_{ij} \quad (3)$$

Increments $\Delta \theta_{ij}$ are obtained by substituting $F_{fit}(l, \theta) = F_{fit}(l, \theta_0) + \Delta F_{fit}(l, \theta_0, \Delta \theta_{ij})$ into Eq. (1) with the equal sign. Corresponding dependences $\Delta \theta(x_0, R, r)$ up to a constant of proportionality will be calculated by following method.

Let us substitute Eq. (3) into Eq. (1). Then it can be shown that

$$\sum_{l=1}^L v_l^2 + v_l \Delta F_{fit}(l) \leq k_1^2 L \sigma^2, \quad (4)$$

where $v_l = F_{ms}(l) - F_{fit}(l, \theta_0)$ is a random noise, k_1 is a constant.

Taking into account Eq. (4) with the equal sign and by replacing sums of Eq. (1) with integrals with infinite limits for Gaussian and Lorentzian peaks:

$$f_G(x) = R \exp\{-arg^2\}, \quad (5)$$

$$f_L(x) = R/(1 + arg^2), \quad (6)$$

where x is a dimensionless variable; $arg = \beta(x - x_0)/r$, $\beta_G = 2\sqrt{\ln 2}$, $\beta_L = 2$, we have

$$\Delta \theta \propto \frac{\sigma}{\sqrt{\sum_{l=1}^L \left(\frac{\partial F_{fit}}{\partial \theta}\right)^2}} \propto \frac{\sigma/\sqrt{r}}{\sqrt{\int_{arg=-\infty}^{+\infty} \left(\frac{\partial F_{fit}}{\partial \theta}\right)^2 darg}} \quad (7)$$

According to Eqs. (5) and (6) Eq. (2) can be transform to

$$F_{fit}(l, \theta) = \sum_{i=1}^N R_i \varphi(arg_i^2) \quad (8)$$

In what follows, for the sake of simplicity, we omit index i of arg .

Consider following derivatives:

$$\frac{\partial F_{fit}}{\partial R_i} = \varphi(arg^2), \quad \frac{\partial F_{fit}}{\partial r_i} = R_i \frac{\partial \varphi(arg^2)}{\partial arg} \frac{\partial arg}{\partial r}, \quad \frac{\partial F_{fit}}{\partial x_{0i}} = R_i \frac{\partial \varphi(arg^2)}{\partial arg} \frac{\partial arg}{\partial x_0}, \quad (9)$$

where $\partial arg / \partial r = -arg * r$, $\partial arg / \partial x_0 = -\beta/r$.

Substituting Eqs. (9) into Eq.(5) we have the following proportions:

$$\Delta R \propto \frac{\sigma/\sqrt{r}}{\sqrt{\int_{arg=-\infty}^{+\infty} (\varphi(arg^2))^2 darg}} \propto \sigma/\sqrt{r} \quad (10)$$

$$\Delta r \propto \frac{\sigma\sqrt{r}/R}{\sqrt{\int_{arg=-\infty}^{+\infty} (\varphi(arg^2))^2 arg^2 darg}} \propto \sigma\sqrt{r}/R \quad (11)$$

$$\Delta x_0 \propto \frac{\sigma\sqrt{r}/R}{\sqrt{\int_{arg=-\infty}^{+\infty} (\varphi(arg^2))^2 darg}} \propto \sigma\sqrt{r}/R \quad (12)$$

since integrals in denominators are constants.

LSE-method requires calculation of derivatives and summation (according to Eq. (7)). Coefficients of proportionality were calculated numerically by computer modeling of Gaussian doublets which parameters were varied in the wide range. Obtained errors were used as optimal increments of the peak parameters for creating huge dataset consisting of 667106 Gaussian doublets and their parameters (more than 1GB). (Final results will be published as soon as possible).

2. BE-method

Suppose that $p(\theta, \sigma)$ is the prior distribution function of the curve fit model parameters (Eq., (2)). Then according to Bayes' theorem [6] the posterior density function is

$$p(\theta, \sigma | F_{ms}) \propto p(F_{ms} | \theta, \sigma) p(\theta, \sigma), \quad (13)$$

where $p(F_{ms} | \theta, \sigma) = (2\pi\sigma^2)^{-L/2} \exp(-S(\theta)/2\sigma^2)$, $p(\theta, \sigma) \propto 1/\sigma$.

It was shown [6] that

$$p(\theta, | F_{ms}) = S(\theta)^{-L/2} \int S(\theta)^{-L/2} d\theta \quad (14)$$

As was pointed out [6] analytic integration in Eq. (14) is usually impossible. Therefore, the posterior probability of precise fit parameters for a given measured analytical signal is calculated using following approximation of Eq. (14) [6]:

$$Pr(\theta = \theta_0 | F_{ms}) = S(\theta_0)^{-L/2} / \sum_{j=1}^M S(\theta_j)^{-L/2}, \quad (15)$$

where θ_j is defined over t -dimensional rectangular grid of M points equally spaced by $\Delta \theta_{ij}$, $S(\theta_0) = \sum_{l=1}^L v_l^2$. Also for comparison with the LSE-method we took the constrains $S(\theta_j) \leq \varepsilon^2 L$ into account. Increment $\Delta \theta_{ij}$ is obtained numerically from Eq. (15) while $\theta_{lm \neq ij}$ are constant parameters (that is, another increments are zeroes: $\Delta \theta_{lm \neq ij} = 0$).

III. Results and Discussion

For modeling Gaussian and Lorentzian peaks (Eqs. (5) and (6)) were taken, where $x = [-5:0.05:5]$, $x_0 = [0.1 \ 0.2 \ 0.3 \ 0.5 \ 0.7 \ 1.0]$, R and $r = [0.25 \ 1/3 \ 0.5 \ 1 \ 2 \ 3 \ 4]$. Peaks were perturbed by additive normal noise with zero mean and standard deviation σ .

For a given set of parameters $\{x_0, R, r\}$, predefined probability Pr and the noise standard deviation increments $\Delta\theta_{ij}$ were calculated from Eq. (15) 25 times; for randomization each time with newly generated noise. The mean values were taken.

According to the remarks in the LSE-method section $\Delta\theta_{ij}$, obtained theoretically does not depend on x_0 since sums were replaced by integrals with infinite limits. However, in the case of the numerical definite summation infinite tails of the peaks are cut off. Therefore possible dependence $\Delta\theta_{ij}(x_0)$ was taken into account: the matrix $\Delta\theta$ was averaged with respect to x_0 (Fig.1). According to our previous findings (see section "LSE-method") [5] we expected the following dependences of the minimum parameters increments on the peak parameters:

$$\Delta p = a\alpha + b, \quad (16)$$

where $\alpha = 1/\sqrt{r}$, $\sqrt{r}/R$ for ΔR , Δr and Δx_0 , respectively, a is proportional to the standard deviation of the noise. Numerical solution of Eq. (16), obtained in the wide range of the doublet parameters fully confirmed Eq. (16) (Fig.2 and Table). The coefficient of linear correlation (χ) is close to 1. It is seen from the tabular data the slope the dependences $\Delta R(\alpha)$ for Gaussian and Lorentzian peak shapes is approximately equal ($a_G \approx a_L$), for dependences $\Delta r(\alpha)$ $a_G/a_L \approx \beta_G/\beta_L$. The slope of the dependences $\Delta x_0(\alpha)$ for Lorentzian peak is only slightly larger than the slope for Gaussian.

Since for given R and r $\Delta x_0 < \Delta r$, curve fitting procedure is less sensitive to the shift of the peak maximum than to the uncertainty of the peak width.

Dependence $\Delta\theta(Pr)$ is approximately linear in the range $Pr = 0.50 - 0.90$ (Fig.3). Further growth of probability strongly increases increments $\Delta\theta$.

It is seen from Eq. (16) that all relative uncertainties have the same dependence on the peak parameters up to a constant of proportionality:

$$\Delta R/R, \Delta r/r, \Delta x_0/r = a/R\sqrt{r} \quad (17)$$

It is follow from Eq. (17) and the tabular data that for high probability $Pr \geq 0.95$ and small R and r the relative uncertainties in determining peak parameters may be more than 15-30%.

Obtained results clearly show that increasing of the increments of peak parameters increase the probability of distinctiveness of analytical signals $F_{fit}(\theta)$ and $F_{fit}(\theta + \Delta\theta)$. Therefore Bayesian computer modeling allows estimating the feasible intervals of "the best" peak parameters obtained by a curve-fitting procedure. The intervals depend on the probability of distinctiveness.

Figure1. Averaging

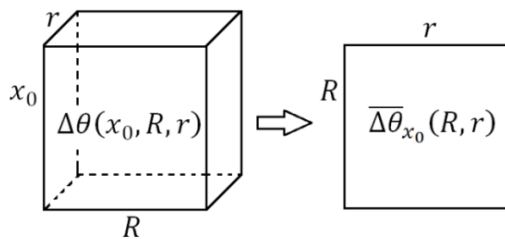


Figure 2. Example of the linear dependence

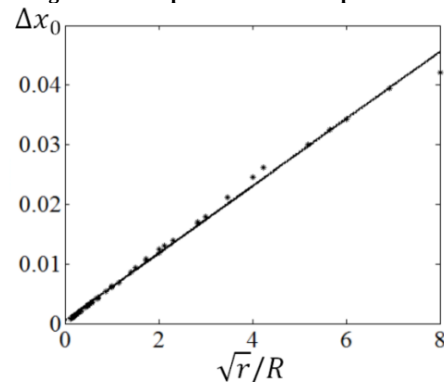


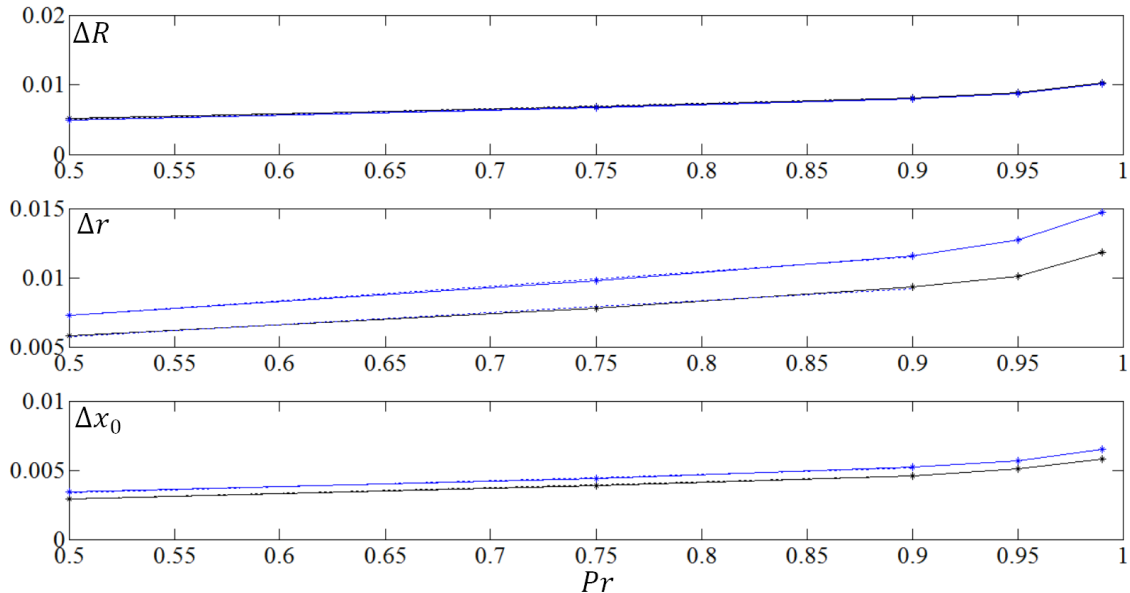
Table. Dependences of the minimum parameter increments on the peak parameters

$\Delta\theta$	Pr	Gaussians			Lorentzians		
		$\Delta\theta(\alpha, a, b)$	$\Delta\theta(\alpha, a)$	χ	$\Delta\theta(\alpha, a, b)$	$\Delta\theta(\alpha, a)$	χ
ΔR	0.50	51α	51α	0.9998	$49\alpha + 1$	49α	0.9995
	0.75	$67\alpha + 1$	68α	0.9999	$66\alpha + 1$	67α	1.0000
	0.90	$80\alpha + 1$	81α	0.9999	80α	80α	0.9999
	0.95	89α	89α	1.0000	87α	87α	0.9999
	0.99	103α	103α	1.0000	$99\alpha + 2$	101α	0.9999

Δr	0.50	$57\alpha + 1$	58α	0.9987	$74\alpha - 4$	73α	0.9964
	0.75	78α	78α	0.9997	$99\alpha - 3$	98α	0.9982
	0.90	$92\alpha + 1$	93α	0.9996	$118\alpha - 4$	116α	0.9983
	0.95	$101\alpha + 1$	101α	0.9997	$129\alpha - 5$	127α	0.9984
	0.99	$117\alpha + 3$	118α	0.9998	$148\alpha - 3$	147α	0.9984
Δx_0	0.50	$29\alpha + 2$	29α	0.9986	$34\alpha + 1$	34α	0.9992
	0.75	$39\alpha + 2$	39α	0.9990	$43\alpha + 4$	44α	0.9987
	0.90	$46\alpha + 3$	46α	0.9989	$51\alpha + 5$	52α	0.9986
	0.95	$50\alpha + 3$	51α	0.9989	$55\alpha + 6$	57α	0.9976
	0.99	$56\alpha + 5$	58α	0.9977	$62\alpha + 9$	65α	0.9941

Pr is the probability. Coefficients a and b are multiplied by 10^4 . χ is the linear correlation coefficient. $\sigma = 0.01$.

Figure 3. Dependences of the minimum errors on the probability



Dependences for Gaussian and Lorentzian peaks are black and blue coloured, respectively.

IV. Conclusion

Error analysis of Bayesian estimates of peak parameters allowed us to calculate minimum uncertainties in determining parameters of analytical signal consisting of overlapping peaks. The theoretical probabilistic approach to the error analysis is the main advantage of this method. Unlike the least squares estimation Bayesian approach does not require cumbersome algebraic calculations. It is applicable to any peak shape, and it is not a very time-consuming process.

However, simple dependences (similar to Eqs. (10)- (12)) of the increments on the peak parameters cannot be obtained analytically for asymmetrical peak shapes since corresponding integrals have no closed-form solution. E.g., for Polynomial-modified Gaussian model [9]: $Rexp\{-[arg/(1 - \tau arg)]^2\}$, where τ is the coefficient of asymmetry, the derivatives with respect to the peak parameters (Eq. (9)) depend on τ . Therefore, integrands are also parameter-dependent functions. In this case numerical approximation is needed.

References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.

- [5] J. Dubrovkin, "Evaluation of minimum uncertainties in determining parametrs of analytical signal consisting of overlapping peaks", The European Conference on Analytical Chemistry (EuroAnalysis-2015), Bordeaux, France.
- [6] G. A. F. Seber, C. J. Wild, Nonlinear Regression, Wiley, 2005.
- [7] Handbook of Markov Chain Monte Carlo, Ed., S.Brooks, A. Gelman, G.L. Jones and X-L. Meng, Chapman & Hall/CRC, 2011.
- [8] S. Moussaoui , , C. Carteret , D. Brie and A. Mohammad-Djafari, "Bayesian Analysis of Spectral Mixture Data using Markov Chain Monte Carlo Methods", Chemometrics and Intelligent Laboratory Systems, vol.81(2006) pp. 137-148.
- [9] J. J. Baeza-Baeza, C. Ortiz-Bolsico, M. C. García-Álvarez-Coque, "New approaches based on modified Gaussian models for the prediction of chromatographic peaks" , Anal. Chim. Acta, vol. 758, 2013, pp. 36–44.