Hindawi Publishing Corporation
Applied Computational Intelligence and Soft Computing
Volume 2014, Article ID 276741, 10 pages
http://dx.doi.org/10.1155/2014/276741

Research Article

A Decomposition Model for HPLC-DAD Data Set and Its Solution by Particle Swarm Optimization

Lizhi Cui,[1,2] Zhihao Ling,[1] Josiah Poon,[2] Simon K. Poon,[2] Junbin Gao,[3] and Paul Kwan[4]

[1] *Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China*
[2] *School of Information Technologies, The University of Sydney, Sydney, NSW 2006, Australia*
[3] *School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia*
[4] *School of Science and Technology, University of New England, Armidale, NSW 2350, Australia*

Correspondence should be addressed to Zhihao Ling; zhhling1957@gmail.com

Received 17 July 2014; Revised 1 November 2014; Accepted 1 November 2014; Published 25 November 2014

Academic Editor: Samuel Huang

This paper proposes a separation method, based on the model of Generalized Reference Curve Measurement and the algorithm of Particle Swarm Optimization (GRCM-PSO), for the High Performance Liquid Chromatography with Diode Array Detection (HPLC-DAD) data set. Firstly, initial parameters are generated to construct reference curves for the chromatogram peaks of the compounds based on its physical principle. Then, a General Reference Curve Measurement (GRCM) model is designed to transform these parameters to scalar values, which indicate the fitness for all parameters. Thirdly, rough solutions are found by searching individual target for every parameter, and reinitialization only around these rough solutions is executed. Then, the Particle Swarm Optimization (PSO) algorithm is adopted to obtain the optimal parameters by minimizing the fitness of these new parameters given by the GRCM model. Finally, spectra for the compounds are estimated based on the optimal parameters and the HPLC-DAD data set. Through simulations and experiments, following conclusions are drawn: (1) the GRCM-PSO method can separate the chromatogram peaks and spectra from the HPLC-DAD data set without knowing the number of the compounds in advance even when severe overlap and white noise exist; (2) the GRCM-PSO method is able to handle the real HPLC-DAD data set.

1. Introduction

After more than 100 years' development, the technology of chromatography has become the collective term for a set of laboratory technique for quality control of various mixtures such as herbal medicine, grape wine, agriculture, and petroleum. With the development of the chromatographic instrument, the High Performance Liquid Chromatography with Diode Array Detector (HPLC-DAD) technology is used in many researches to generate a data set containing the chromatogram peaks and spectra for all compounds. Figure 1 shows the principle of the HPLC-DAD data set. The sample is injected at the sample injection. The high pressure pump drives the solvent to carry the sample to go through the column with absorbent. Different compounds will receive different resistance when they go through the column. Given an ultraviolet detector at the bottom of the column, a chromatogram peak represented by $\mathbf{s}_i$ will be observed when one compound comes out from the column. The position and area of the peak can tell the name and the amount of the compound. If the detector is a DAD, which has more than one thousand channels to detect multiwavelength simultaneously, the spectrum for the same compound represented by $\mathbf{a}_i$ will also be recorded as well. $\mathbf{D}_i = \mathbf{a}_i \times \mathbf{s}_i^T$ represents ith compound and $\mathbf{X}$ represents the mixture. The relationship of the variables in Figure 1 can be shown as

$$\mathbf{X}_{w\times t} = \sum_{i=1}^{n} \mathbf{D}_i = \sum_{i=1}^{n} \mathbf{a}_i \times \mathbf{s}_i^T = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n] \times \begin{bmatrix} \mathbf{s}_1^T \\ \mathbf{s}_2^T \\ \vdots \\ \mathbf{s}_n^T \end{bmatrix} = \mathbf{A} \times \mathbf{S}, \tag{1}$$

where n indicates the number of the compounds.

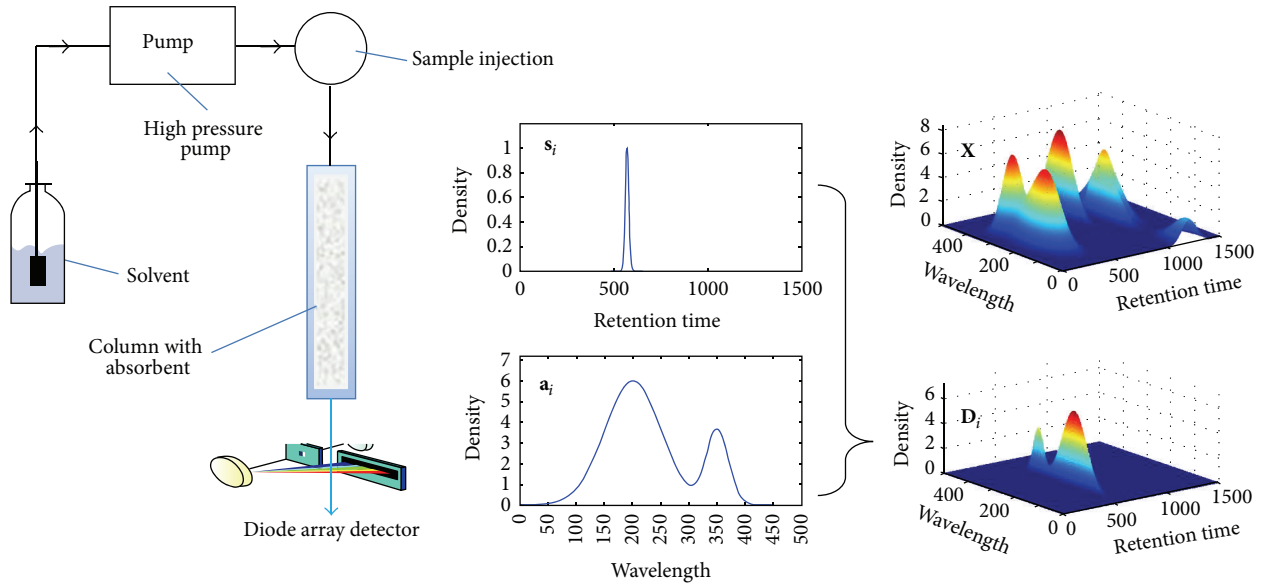


Figure 1: The principle of the HPLC-DAD data set.

For the data set $\mathbf{X}$ in (1), there are already several methods to separate it, but with insufficiencies. The algorithm of evolving factor analysis (EFA) [1, 2] and its improvements such as evolutionary factor analysis (EVOLU) [3], fixed-size moving window evolving factor analysis (FSMWEFA) [4], heuristic evolving latent projections (HELP) [5], and orthogonal projection resolution (OPR) [6] are used for peak purity, but without full quantitative information. The method of Multivariate Curve Resolution with Alternating Least Square (MCR-ALS) [7, 8] can recover the pure species spectra and elution profiles. However, the MCR-ALS method will be unavailable when the compounds become complex (see the simulations). And the performance of the MCR-ALS method depends on two important parameters: (1) a threshold for deciding the number of the compounds; (2) the noise level of the data set for estimating initial spectra. Usually, it is not easy to decide these two parameters when noise exists (see Appendix A for explanation). The immune algorithm (IA) [9, 10] can extract the compounds from noise. But, the standard chromatogram peaks for compounds are needed from experiments in advance. The method of independent component analysis (ICA) [11] can separate the HPLC-DAD data set without knowing the number of the compounds in advance. But the cluster methods are still needed to select compounds from the obtained independent components. Our previous works proposed a model named independent components analysis constrained by reference curve (ICARC) and its solution by multiarea genetic algorithm (mGA) [12] and by multitarget Particle Swarm Optimization (mPSO) [13], respectively, which can extract the chromatogram peaks from the HPLC-DAD data set directly. However, through further analysis, we find that it is not necessary for the chromatogram peaks (source signals) to be independent from each other. So a method based on the model of Generalized Reference Curve Measurement (GRCM) and the algorithm of Particle Swarm Optimization (PSO) is proposed in this paper.

The remainder of this paper is arranged as follows: Section 2 introduces the principle of the GRCM-PSO method; Section 3 gives the simulations and experiments; finally, Section 4 draws the conclusions and future works.

2. Mathematical Methods

It is difficult to extract the $\mathbf{a}_i$ and $\mathbf{s}_i$ in (1) only based on the data set $\mathbf{X}$ without any other knowledge. Fortunately, the fact that the shape of a chromatogram peak looks like a Gaussian curve [14] can help. Based on this "a priori" knowledge, the GRCM-PSO method is proposed as shown in Figure 2. Firstly, a reference curve $\mathbf{r}(\boldsymbol{\theta})$ with parameter $\boldsymbol{\theta}$ is constructed based on the general shape of the chromatogram peak, according to which the initial population $\boldsymbol{\theta}_i$, $i = 1, 2, \ldots, n$ are generated. Then, the GRCM model calculates the errors ε_i, $i = 1, 2, \ldots, n$ for the parameters. Following, a search category is used to obtain the rough solutions $\boldsymbol{\theta}_i^r$, $i = 1, 2, \ldots, m$ ($m \ll n$). In the dashed rectangular box in Figure 2, a step called reinitialization generates t parameters randomly around one rough solution, for example, $\boldsymbol{\theta}_{1i}^r$, $i = 1, 2, \ldots, t$ in Figure 2 for the first rough solution. The GRCM model calculates the errors ε_{1i}^r for these $\boldsymbol{\theta}_{1i}^r$. Based on these errors ε_{1i}^r, the PSO algorithm is adopted to obtain the optimal parameter $\boldsymbol{\theta}_1^*$ around $\boldsymbol{\theta}_1^r$. Similarly, other optimal parameters $\boldsymbol{\theta}_i^*$, $i = 2, 3, \ldots, m$ can be found. Finally, the approximated chromatogram peaks can be constructed by the reference curve, and the spectra can be obtained by an estimator.

The structure of the parameter used in this paper is the same as that in literature [13], which is shown by (2) and (3). Equation (2) is the Gaussian curve which will be used in

the simulations to demonstrate the performance of the GRCM-PSO method; (3) is a 5-parameter curve which will be used in the experiments to show the practicability of our method:

$$\mathbf{r}_1(\boldsymbol{\theta}) = \mathbf{r}_1(\mu,\sigma) = \exp\left[-\frac{(x-\mu)^2}{2*\sigma^2}\right], \quad \mu \in [1, \mathrm{col}(\mathbf{X})], \quad \sigma \in \left[1, \left(\frac{(\mu_2-\mu_1)}{6}\right)\right], \tag{2}$$

$$\mathbf{r}_2(\boldsymbol{\theta}) = \mathbf{r}_2(\mu,\sigma_L,\sigma_R,h_L,h_R) = \begin{cases} \mathrm{unit}\left\{\exp\left[-\frac{(x-\mu)^2}{(2*\sigma_L^2)}\right] + \frac{h_L}{(1-h_L)}\right\}, & x \le \mu, \\ \mathrm{unit}\left\{\exp\left[-\frac{(x-\mu)^2}{(2*\sigma_R^2)}\right] + \frac{h_R}{(1-h_R)}\right\}, & x > \mu, \end{cases} \quad \mu \in [1, \mathrm{col}(\mathbf{X})], \quad \sigma_L, \sigma_R \in \left[1, \left(\frac{(\mu_2-\mu_1)}{3}\right)\right], \tag{3}$$

where col($\mathbf{X}$) is the column number of $\mathbf{X}$. $\mathbf{r}_2(\boldsymbol{\theta})$ is the combination of two Gaussian curves at the peak position with σ_L and σ_R for each side's width and h_L and h_R for each side's deviation from zero. The ranges of the σ, σ_L, σ_R are limited in order to guarantee that every peak has an integral shape. $h_L \in [0, 0.0001]$, $h_R \in [0, 0.03]$ in the experiments due to the profile of the data set. unit$\{\cdot\}$ is the function to limit the amplitude at 1.

In order to obtain initial parameters with small errors, four times of initialization with the same population of $n = 2000$ are implemented to generate 8000 parameters totally and only the top 2000 parameters according to their errors are chosen as the initialized parameters.

2.1. The Model of GRCM. The function of the GRCM model is to assess the parameters by calculating their errors, which indicate the distance between the reference curves constructed by these parameters and the chromatogram peaks existing in $\mathbf{X}$. As shown in Figure 3, the GRCM model is composed of five elements: reference curve $\mathbf{r}(\boldsymbol{\theta})$, data set $\mathbf{X}$, Reference Curve Measurement (RCM) model, predicted curve (PC) $\mathbf{y}$, and measurement operator (MO) $\|\cdot\|_{\mathrm{MO}}$.

The RCM model is designed by introducing a vector $\mathbf{m}$ so that

$$\mathbf{y} = \mathbf{m}^T \times \mathbf{X} = \mathbf{m}^T \times [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t] \approx \mathbf{s} \longrightarrow \mathbf{r}(\boldsymbol{\theta}). \tag{4}$$

Equation (4) means that let $\mathbf{y}$ approximate $\mathbf{s}$ and look like $\mathbf{r}(\boldsymbol{\theta})$. Then, we have the objective function as

$$\min\left\{\left\|\mathbf{y}^T - \mathbf{r}^T(\boldsymbol{\theta})\right\|_2^2\right\}. \tag{5}$$

Solving (5), we obtain the RCM model as

$$\mathbf{y}^T = \mathbf{r}^T(\boldsymbol{\theta}) \times \left(\frac{1}{t} \times \widetilde{\mathbf{X}}^{*T} \times \widetilde{\mathbf{X}}^*\right), \tag{6}$$

where $\widetilde{\mathbf{X}}^*$ is a matrix generated from $\mathbf{X}$, which will be introduced in Appendix B as well as the deducing process from (5) to (6).

The MO $\|\cdot\|_{\mathrm{MO}}$ is designed as

$$\varepsilon = \frac{\left\|\mathbf{y}^T - \mathbf{r}^T(\boldsymbol{\theta})\right\|_2^2}{\left\|\mathbf{r}^T(\boldsymbol{\theta})\right\|_2^2}. \tag{7}$$

2.2. Search Category and Reinitialization. After the initialization, every parameter $\boldsymbol{\theta}_i$ will search within a small hypersphere to find one parameter with the smallest error as its target $\mathbf{T}_i$. It is possible for $\boldsymbol{\theta}_i$ to find $\boldsymbol{\theta}_j$ as its target, that is, $\mathbf{T}_i = \boldsymbol{\theta}_j$, and for $\boldsymbol{\theta}_j$ to find $\boldsymbol{\theta}_k$ as its target, that is, $\mathbf{T}_j = \boldsymbol{\theta}_k$. In order to accelerate the searching speed, we directly set $\mathbf{T}_i = \mathbf{T}_j = \boldsymbol{\theta}_k$. Finally, only limited parameters have been chosen as targets for others, which are the rough solutions $\boldsymbol{\theta}_i^r$, $i = 1, 2, \ldots, m$ ($m \ll n$). It is assumed that all the real solutions are around the rough solutions because the intensively and randomly distribution of the initializing parameters. So a step named reinitialization only around rough solutions will reduce the search area significantly. The areas for reinitialization are hyperspheres, whose radii are half of the smallest distance between the centre rough solution and other solutions in order to cover all the possible spaces. An example of such a hypersphere is illustrated in Figure 4. There are five rough solutions $\boldsymbol{\theta}_i^r$, ($i = 1, 2, \ldots, 5$) in the two-dimensional space. The distance between $\boldsymbol{\theta}_3^r$ and other rough solutions is $d_{13} < d_{35} < d_{34} < d_{23}$. So the hypersphere for $\boldsymbol{\theta}_3^r$ is shown by the circle in Figure 4, where $R_3 = 0.5 \times d_{13}$. The population in every hypersphere is set to 10.

2.3. Algorithm of PSO. PSO is swarm intelligence that emulates social interaction and individual cognition of bird flocks foraging [15, 16]. Equation (9) gives the algorithm of PSO:

$$\mathbf{v}_{i+1} = \omega \times \Delta\mathbf{v}_i + c_1 \times r_1 \times (\mathbf{p}_i - \boldsymbol{\theta}_i) + c_2 \times r_2 \times (\mathbf{p}_g - \boldsymbol{\theta}_i), \quad \boldsymbol{\theta}_i^+ = \boldsymbol{\theta}_i + \mathbf{v}_{i+1}, \tag{8}$$

where $\boldsymbol{\theta}_i$ and $\mathbf{v}_i$ represent the position and velocity of the ith particle, respectively; ω is the inertia weight; c_1 and c_2 are acceleration constants; r_1 and r_2 are two random numbers in $[0, 1]$; $\mathbf{p}_i$ is the personal best position for $\boldsymbol{\theta}_i$; and $\mathbf{p}_g$ is the global best position. Please see relative literatures for the values of the parameters in (8).

In this paper, all the parameters are divided in several different groups within certain hyperspheres. And every group updates these particles according to (8), respectively, until the value of the best particle in every group does not change for 500 steps, or the maximum step is reached.

2.4. Other Processes. During the process from $\boldsymbol{\theta}_i$, $i = 1, 2, \ldots, n$ to $\boldsymbol{\theta}_i^*$, $i = 2, 3, \ldots, m$ in Figure 2, the random initialization of $\boldsymbol{\theta}_i$ may cause inaccuracy in the results. So this process is executed multiple times to eliminate the influence of the random initialization. Through observation, ten times was chosen. Ten executions will generate 10 candidate solutions.

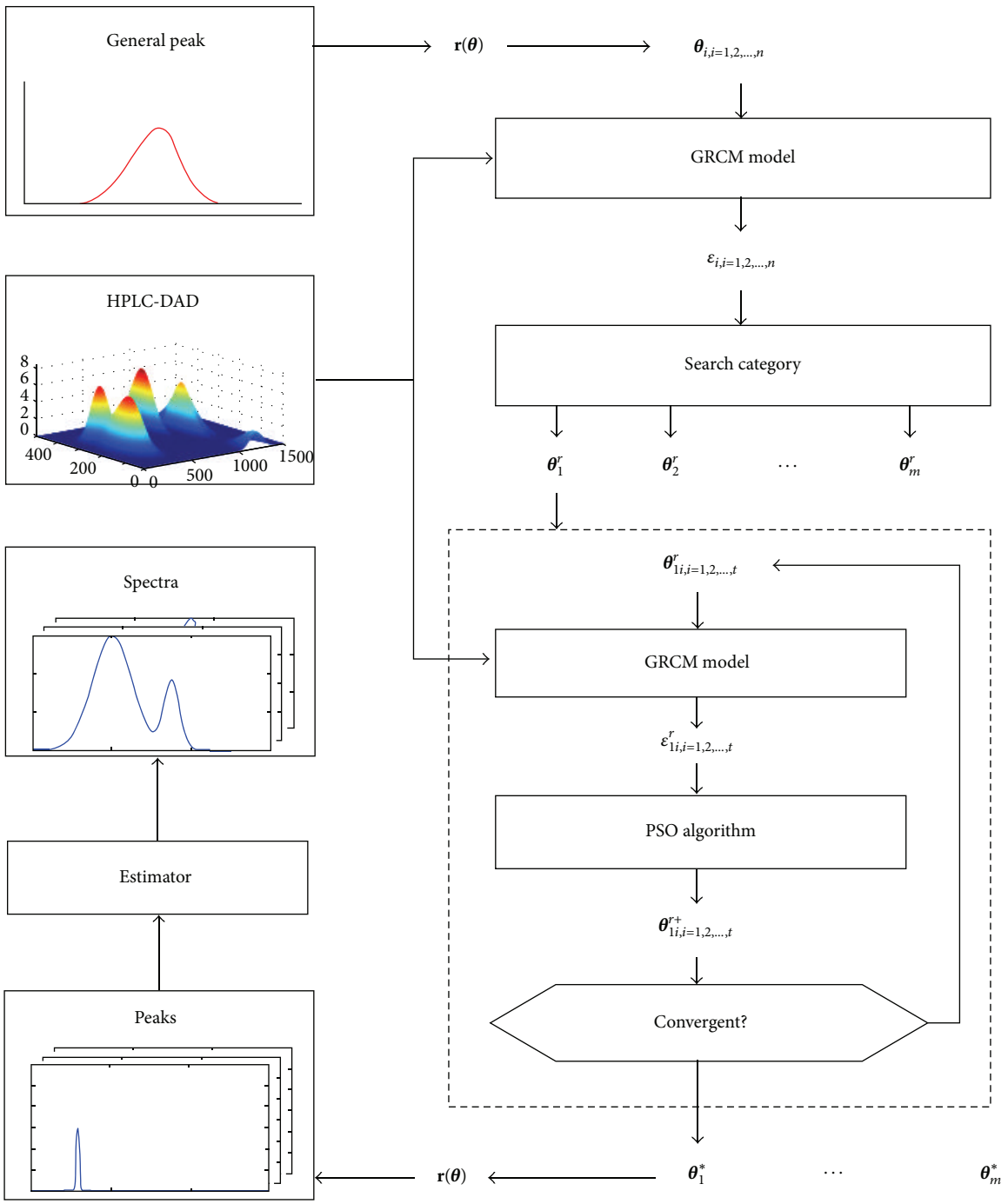


Figure 2: Principle of the GRCM-PSO method.

There can be difference among the value even the number of the optimal parameters among these candidates. Firstly, select one candidate with the maximum number of parameters as a reference. Then, select one parameter from every candidate to be grouped with one parameter in the reference according to the Euclidean distance and count the number of parameters in every group. Only the groups with the number of parameters lager than 6 are selected as valid groups. Finally, choose one parameter with lowest error from every valid group to form the final result.

Finally, the estimator is designed as the following equation to calculate the spectra for all the compounds:

$$\widehat{\mathbf{A}} = \mathbf{X} \times \text{pinv}\left(\widehat{\mathbf{S}}\right), \tag{9}$$

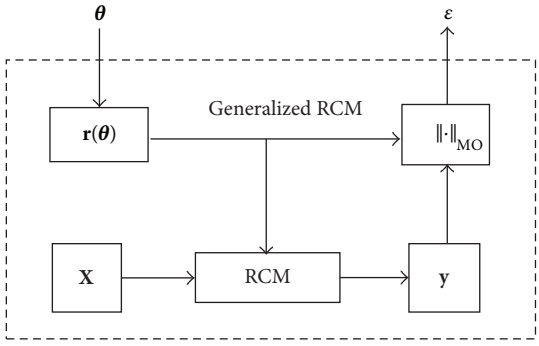


Figure 3: The GRCM model.

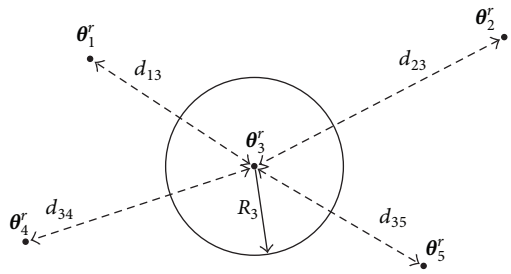


Figure 4: The hypersphere for the reinitialization.

where $\widehat{\mathbf{S}} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m]^T \approx [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m]^T$ is the approximation of chromatogram peaks; pinv(·) is the pseudoinverse function. Equation (9) is derived from (1) directly.

3. Simulation, Experiment, and Discussion

In this section, a group of simulations are given to demonstrate the performance of the GRCM-PSO method. Then experiments on a HPLC-DAD data set are implemented to show the practicability of the GRCM-PSO method. Two criteria are used to evaluate the results: (1) whether all the chromatogram peaks can be found; (2) whether the errors between the true/simulated spectra and estimated spectra are small enough.

3.1. Simulation and Discussion. The simulation data set is shown in graphs (a) and (b) of Figure 5, which contains seven compounds with severe overlap. The seven chromatogram peaks are constructed by (2) with the parameters of [45, 6], [53, 15], [59, 9], [100, 30], [141, 5], [149, 15], and [157, 8]. And the seven spectra are constructed randomly as long as they are uncorrelated with each other. The simulation data set is added with different level of whiten noise. The results are listed in Table 1. From the results, we can see the following.

(1) The GRCM-PSO method can separate the simulation data set without knowing the compounds' number in advance even when severe overlap and white noise exist. This is a big advantage over previous method which needs to know the compounds' number in advance. The values of the ε and the error between the calculated spectra and the simulated spectra are small. The time cost by this method is much less than that by ICARCmPSO [12], which was 13.9 seconds. However, the MCR-ALS method cannot give correct results. The results given by MCR-ALS method are illustrated in graphs (c) and (d) of Figure 5, in which no noise is added to the simulated data set.

(2) The average time cost by the ten implementations is almost the same. This means that the degree of the white noise has no significant influence on the time cost.

(3) The values of the ε and the error between the calculated spectra and the simulated spectra become larger with the increase of the noise level. What should be noted is that when noise becomes severe, the "Errors" for small peaks are influenced more significantly than that for big peaks.

3.2. Experiment and Discussion. The HPLC-DAD data set of "adataset.mat" is downloaded from http://www.mcrals.info/ for free. This data set is a three-compound mixture with two known pesticides and one unknown interferent [8]. The 5-parameter function $\mathbf{r}_2(\boldsymbol{\theta})$ shown by (3) is used in the experiments as the RC. The graphical results are illustrated in Figure 6 as well as the results given by the ICARCmPSO method and the MCR-ALS method. The values of the results are listed in Table 2. From the experiments, we can see the following.

(1) Comparison between the GRCM-PSO method and the ICARCmPSO [13]: the average time and average steps for the GRCM-PSO method are much less. The ε and the "Errors" are similar for both of these two methods.

For the ICARCmPSO method, the parameter of scope for particles to search their local targets should be determined according to the specific application. For the GRCM-PSO method, this parameter is fixed to a small value. So, from the view of operability and speed, the GRCM-PSO method has advantage over the ICARCmPSO method.

(2) Comparison between the GRCM-PSO method and the MCR-ALS method: both of them obtain the same number, 3, of the compounds. The speed of the MCR-ALS method is better than that of the GRCM-PSO method. The accuracy of the MCR-ALS method can be better than that of the GRCM-PSO method. But the parameters in the GRCM-PSO method are easier to be controlled.

For the MCR-ALS method, the important two parameters are the threshold to select the valid singular values and the noise level for initial estimation of the spectra. If noise exists, it will be difficult to decide the first parameter as explained in Appendix A. The performance of the MCR-ALS method is also very sensitive to the second parameter as shown in graph (i) of Figure 6. A small change of the β, which is explained in Appendix A, will cause big error in the calculated spectra, while all the parameters for the GRCM-PSO method are fixed for all applications.

So, from the view of operability and stability, the GRCM-PSO method has advantage over the MCR-ALS method.

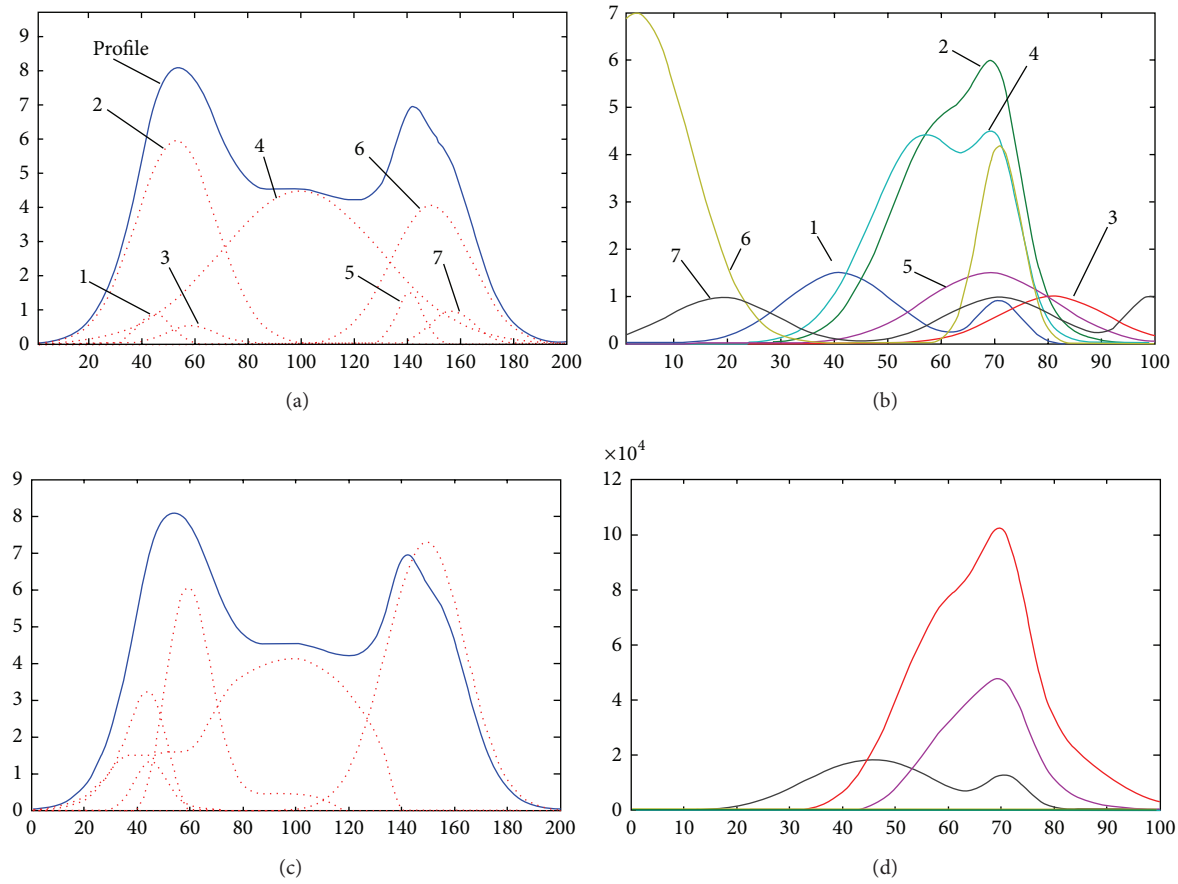


FIGURE 5: (a) The profile and simulated chromatogram peaks; (b) simulated spectra; (c) the calculated chromatogram peaks by MCR-ALS method; (d) the calculated spectra by MCR-ALS method.

4. Conclusions and Future Works

A method named GRCM-PSO was proposed in this paper to separate the chromatogram peaks and spectra for compounds from the HPLC-DAD data set. The GRCM model transformed the separation problem to a multiparameter optimization issue. The PSO algorithm was introduced to calculate the optimal parameters. Groups of simulations with different noise level were implemented. A simulated data set was constructed with severe overlap among seven compounds. The GRCM-PSO method separated the chromatogram peaks and spectra from this simulated data set without knowing the number of the compounds in advance. And the speed was fast. Groups of experiments on a real HPLC-DAD data set were implemented. And comparisons among the results by the GRCM-PSO method, the ICARCmPSO method, and the MCR-ALS method were given. The results showed that the GRCM-PSO method was an effective, efficient, and practical method to separate HPLC-DAD data set even when severe overlap and white noise existed. The speed and practicability of the GRCM-PSO method are better than that of the ICARCmPSO method. The stability and operability of the GRCM-PSO method are better than that of the MCR-ALS method.

Currently, the performance of the GRCM-PSO method depends on the selection of the reference curve. So it is only suitable for the separation task with "a priori" knowledge to be known, such as the separation of HPLC-DAD data set. The accuracy of the result by the GRCM-PSO method can be improved by further research on more accurate reference curves.

Appendices

A. Parameters of MAC-ALS Method

The flowchart of the MCR-ALS method is illustrated in graph (a) of Figure 7. The number of the compounds is calculated by the SVD method. The SVD method needs a threshold to keep the valid singular values, which is noted as α. The initial estimation of $\mathbf{S}_0$ is calculated by the pure variable detection method, which needs to know the noise level of the data set which is noted as β. The variables of $\mathbf{a}$ and $\mathbf{b}$ are two thresholds given according to the data set. Please see relative literature for detailed information of the MCR-ALS method [8].

The singular values from largest to smallest for the HPLC-DAD data set used in the experiment of this paper are listed in

Table 1: The calculated results for the simulations with different level of white noise. The column of "Rate" indicates the rate for this parameter emerging among the total ten candidates. The column of "Error" represents the error between simulated spectra and calculated spectra.

Number	SNR	Parameters		Steps/times	Rate	ε	Error
		μ	σ				
1	No noise	45	6	1048.1/3.17 s	1	$3.34e-27$	$2.63e-25$
		53	15		1	$7.22e-27$	$2.74e-26$
		59	9		1	$5.26e-26$	$1.06e-24$
		100	30		1	$4.12e-28$	$1.63e-27$
		141	5		1	$2.55e-25$	$2.48e-24$
		149	15		1	$5.85e-28$	$1.40e-26$
		157	8		1	$1.95e-27$	$3.14e-24$
2	100	45	6	1059.3/3.23 s	1	$5.29e-11$	$2.57e-09$
		53	15		1	$9.97e-11$	$2.64e-10$
		59	9		1	$6.86e-10$	$5.59e-09$
		99.99	30		1	$2.00e-11$	$1.83e-11$
		141	4.99		1	$2.96e-9$	$2.53e-09$
		149	15		1	$9.02e-13$	$1.42e-09$
		157	8		1	$4.97e-11$	$5.73e-08$
3	50	44.99	5.99	1063.1/3.20 s	1	$5.32e-6$	0.01
		52.96	15.06		1	$9.52e-6$	$2.4e-5$
		58.98	9.01		1	$5.75e-5$	0.09
		99.99	29.99		1	$1.99e-6$	$1.50e-6$
		140.98	5.05		1	$2.59e-4$	0.0003
		149	14.99		1	$1.22e-7$	0.0001
		157	7.99		1	$5.31e-6$	0.005
4	40	45	5.99	1055.7/3.26 s	1	$5.38e-5$	0.11
		52.79	15.17		1	$1.07e-4$	$5.3e-4$
		58.95	8.98		0.9	$6.65e-4$	1.23
		100.04	30.46		1	$1.99e-5$	$2.30e-5$
		140.99	5.03		1	$3.1e-3$	0.008
		149.01	14.99		1	$8.55e-7$	0.008
		156.98	8.01		1	$5.90e-5$	0.09

Table 2: The calculated results for the experiments. The column of "Rate" indicates the rate for this parameter emerging among the total ten candidates. The column of "Error" represents the error between the calculated spectra and the true spectra.

Method	Parameters					Steps/times	Rate	ε	Error
	μ	$\sigma(\sigma_L)$	σ_R	h_L	h_R				
GRCM-PSO	23.93	7.11	12.70	0	0.03	1361/3.61 s	0.8	$1.9e-5$	0.27
	30.29	8.72	14.07	$1.0e-4$	0.01		1	$1.3e-5$	1.59
	59.85	7.91	14.43	$3e-21$	0.03		1	$3.3e-5$	—
ICARC mPSO	23.07	7.37	12.28	$0.9e-4$	0.08	1481.8/16.55 s	1	$1.5e-5$	0.18
	29.91	8.39	14.12	$1.2e-4$	0.01		1	$1.4e-5$	1.73
	59.17	8.26	13.04	$0.8e-4$	0.10		1	$2.1e-5$	—
MCR-ALS	—	—	—	—	—	1000/0.39 s	—	—	0.033
	—	—	—	—	—		—	—	0.153

graph (b) of Figure 7. Although three compounds are known to be contained in the data set, there is no obvious boundary between the third singular value and the fourth one. That is to say, it is not easy to set the value for the parameter α.

As shown in graph (i) of Figure 6, a small change of the parameter β will lead a big error for the calculated spectrum.

B. Deducing Process of RCM Model

In order to make the calculation for (5) simpler, a preprocessing [17] is used to transform $\mathbf{X}$ as

$$\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_{c1}, \widetilde{\mathbf{x}}_{c2}, \ldots, \widetilde{\mathbf{x}}_{ct}] = \mathbf{W} \times \left(\mathbf{X} - \overline{\mathbf{X}}\right), \tag{B.1}$$

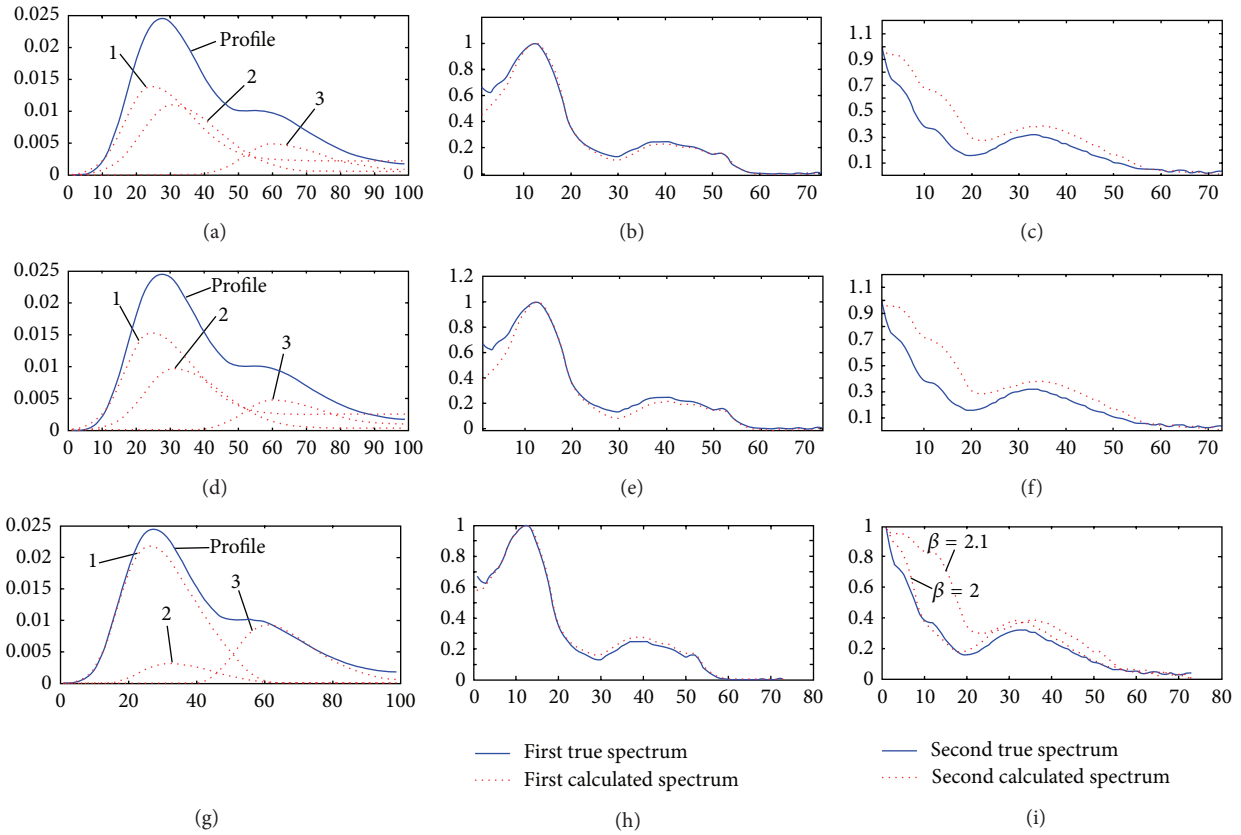
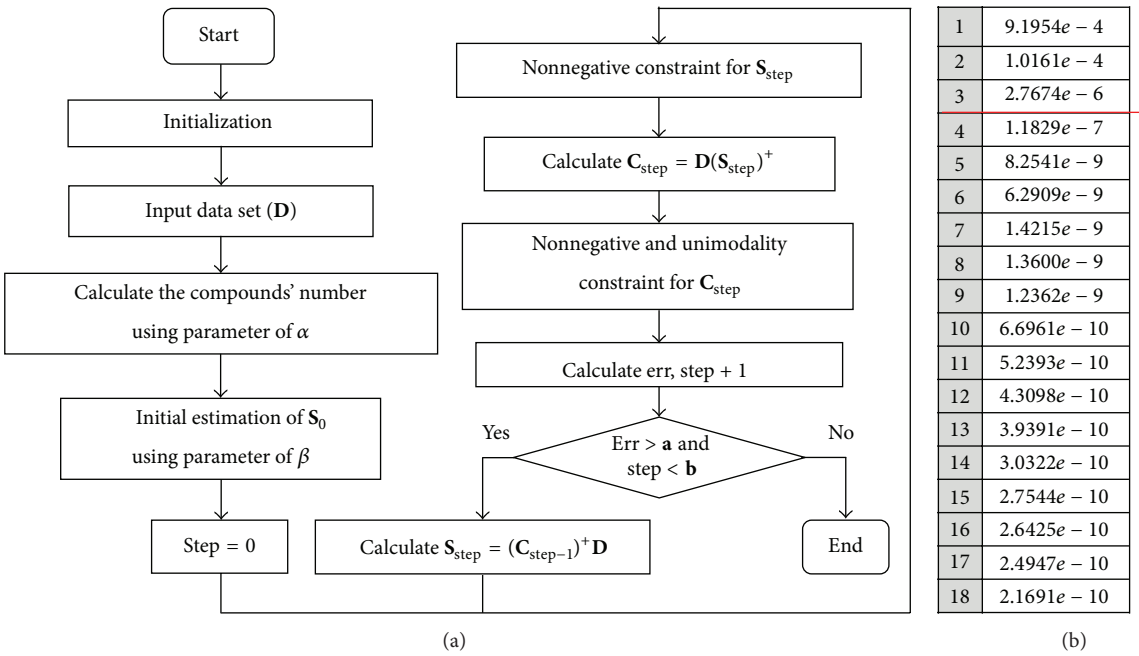


FIGURE 6: The calculated results. (a), (b), and (c) are for the GRCM-PSO method. (d), (e), and (f) are for the ICARCmPSO method. (g), (h), and (i) are for the MCR-ALS method. (a), (d), and (g) are the calculated chromatogram peaks. (b), (e), and (h) are the compares between first spectrum and corresponding calculated spectrum. (c), (f), and (i) are the compares between the second ones.



1	$9.1954e-4$
2	$1.0161e-4$
3	$2.7674e-6$
4	$1.1829e-7$
5	$8.2541e-9$
6	$6.2909e-9$
7	$1.4215e-9$
8	$1.3600e-9$
9	$1.2362e-9$
10	$6.6961e-10$
11	$5.2393e-10$
12	$4.3098e-10$
13	$3.9391e-10$
14	$3.0322e-10$
15	$2.7544e-10$
16	$2.6425e-10$
17	$2.4947e-10$
18	$2.1691e-10$

(b)

FIGURE 7: (a) Flowchart of the MCR-ALS method. (b) The singular values for the HPLC-DAD data set ordered from largest to smallest.

where $\mathbf{W}$ is a matrix generated in the preprocessing; $\overline{\mathbf{X}}$ is a matrix in which every row is filled with the average of every row of $\mathbf{X}$ (see [17] for details). Every column vector $\widetilde{\mathbf{x}}_{ci}(i = 1, 2, \ldots, t)$ in (B.1) satisfies

$$E\left\{\widetilde{\mathbf{x}}_{ci} \times \widetilde{\mathbf{x}}_{ci}^T\right\} = \mathbf{I}_{w\times w}, \tag{B.2}$$

where w is the row number of the matrix $\widetilde{\mathbf{X}}$. Then, (6) can be transformed as

$$\begin{aligned} &\min\left\{\left\|\mathbf{y}_i^T - \mathbf{r}^T(\boldsymbol{\theta}_i)\right\|_2^2\right. \\ &\quad\left. = \left\|\mathbf{m}_i^T \times \overline{\mathbf{X}} + \mathbf{m}_i^T \times \mathbf{W}^{-1} \times \widetilde{\mathbf{X}} - \mathbf{r}^T(\boldsymbol{\theta}_i)\right\|_2^2\right\}, \\ &\widetilde{\mathbf{X}} \times \widetilde{\mathbf{X}}^T = \sum_i^t \widetilde{\mathbf{x}}_{ci} \times \widetilde{\mathbf{x}}_{ci}^T = t \times \mathbf{I}_{w\times w}, \end{aligned} \tag{B.3}$$

where t is the number of the columns of $\widetilde{\mathbf{X}}$; $\overline{\mathbf{X}}$ and $\mathbf{W}$ are the matrices generated in the preprocessing. If we set

$$\begin{aligned} \mathbf{d}_i &= \mathbf{m}_i^T \times \overline{\mathbf{X}} = [d_i, d_i, \ldots, d_i], \\ \mathbf{b}_i^T &= \mathbf{m}_i^T \times \mathbf{W}^{-1}, \end{aligned} \tag{B.4}$$

where $\mathbf{d}_i$ is a vector with the same value for every element referring to a specific $\mathbf{m}_i^T$ and (B.3) is transformed as

$$\begin{aligned} &\min_{\overline{\mathbf{b}}_i^T}\left\{\left\|\overline{\mathbf{b}}_i^T \times \widetilde{\mathbf{X}}^* - \mathbf{r}^T(\boldsymbol{\theta}_i)\right\|_2^2\right\}, \\ &\overline{\mathbf{b}}_i^T = \left[\mathbf{b}_i^T, d_i\right], \\ &\widetilde{\mathbf{X}}^* = \begin{bmatrix}\widetilde{\mathbf{X}} \\ \mathbf{1}\end{bmatrix} = \begin{bmatrix}\widetilde{\mathbf{x}}_{c1} & \widetilde{\mathbf{x}}_{c2} & \cdots & \widetilde{\mathbf{x}}_{ct} \\ 1 & 1 & \cdots & 1\end{bmatrix} = \left[\widetilde{\mathbf{x}}_{c1}^*, \widetilde{\mathbf{x}}_{c2}^*, \ldots, \widetilde{\mathbf{x}}_{ct}^*\right], \\ &\widetilde{\mathbf{X}}^* \times \widetilde{\mathbf{X}}^{*T} = t \times \mathbf{I}_{(w+1)\times(w+1)}. \end{aligned} \tag{B.5}$$

The proof of $\widetilde{\mathbf{X}}^* \times \widetilde{\mathbf{X}}^{*T} = t \times \mathbf{I}_{(w+1)\times(w+1)}$ will be given in Appendix C. Equation (B.5) is an optimization problem. According to the Karush-Kuhn-Tucker (KKT) conditions [18], the solution should satisfy

$$F\left(\overline{\mathbf{b}}^T\right) = \sum_{j=1}^t 2 \times \widetilde{\mathbf{x}}_{cj}^{*T} \times \left(\overline{\mathbf{b}}^T \times \widetilde{\mathbf{x}}_{cj}^* - \mathbf{r}^T(j; \boldsymbol{\theta}_i)\right) = 0, \tag{B.6}$$

where $\mathbf{r}^T(j; \boldsymbol{\theta})$ is the value of jth element under parameter $\boldsymbol{\theta}$. The Jacobean matrix of (B.6) is

$$JF\left(\overline{\mathbf{b}}^T\right) = \sum_{j=1}^t 2 \times \widetilde{\mathbf{x}}_{cj}^* \times \widetilde{\mathbf{x}}_{cj}^{*T}. \tag{B.7}$$

Therefore, the following formula is obtained based on Newton iteration [19]:

$$\begin{aligned} \overline{\mathbf{b}}^+ &= \overline{\mathbf{b}} - \frac{\sum_{j=1}^t 2\widetilde{\mathbf{x}}_j^*\left(\overline{\mathbf{b}}^T\widetilde{\mathbf{x}}_{cj}^* - \mathbf{r}^T(j; \boldsymbol{\theta}_i)\right)}{\sum_{j=1}^t 2\widetilde{\mathbf{x}}_{cj}^* \times \widetilde{\mathbf{x}}_{cj}^{*T}} \\ &= \frac{\sum_{j=1}^t \widetilde{\mathbf{x}}_{cj}^* \times \mathbf{r}(j; \boldsymbol{\theta}_i)}{t \times \mathbf{I}} = \frac{1}{t} \times \widetilde{\mathbf{X}}^* \times \mathbf{r}(\boldsymbol{\theta}_i). \end{aligned} \tag{B.8}$$

With (B.8), we can calculate $\mathbf{y}_i^T$ as

$$\mathbf{y}_i^T = \overline{\mathbf{b}}_i^T \times \widetilde{\mathbf{X}}^* = \mathbf{r}^T(\boldsymbol{\theta}_i) \times \left(\frac{1}{t} \times \widetilde{\mathbf{X}}^{*T} \times \widetilde{\mathbf{X}}^*\right). \tag{B.9}$$

C. Proof for (B.5)

From the definition in (B.5), we have

$$\widetilde{\mathbf{X}}^* = \begin{bmatrix}\widetilde{\mathbf{X}} \\ \mathbf{1}\end{bmatrix} = \begin{bmatrix}\widetilde{\mathbf{x}}_{c1} & \widetilde{\mathbf{x}}_{c2} & \cdots & \widetilde{\mathbf{x}}_{ct} \\ 1 & 1 & \cdots & 1\end{bmatrix}, \tag{C.1}$$

where t is the column number of the matrix $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{x}}_{ci}$ $(i = 1, 2, \ldots, t)$ are the column vectors in $\widetilde{\mathbf{X}}$. Then, we have

$$\begin{aligned} \widetilde{\mathbf{X}}^* \times \widetilde{\mathbf{X}}^{*T} &= \begin{bmatrix}\widetilde{\mathbf{x}}_{c1} & \widetilde{\mathbf{x}}_{c2} & \cdots & \widetilde{\mathbf{x}}_{ct} \\ 1 & 1 & \cdots & 1\end{bmatrix} \times \begin{bmatrix}\widetilde{\mathbf{x}}_{c1}^T & 1 \\ \widetilde{\mathbf{x}}_{c2}^T & 1 \\ \vdots & \vdots \\ \widetilde{\mathbf{x}}_{ct}^T & 1\end{bmatrix} \\ &= \sum_{i=1}^t \begin{bmatrix}\widetilde{\mathbf{x}}_{ci} \\ 1\end{bmatrix} \times \begin{bmatrix}\widetilde{\mathbf{x}}_{ci}^T & 1\end{bmatrix} = \sum_{i=1}^t \begin{bmatrix}\widetilde{\mathbf{x}}_{ci} \times \widetilde{\mathbf{x}}_{ci}^T & \widetilde{\mathbf{x}}_{ci} \\ \widetilde{\mathbf{x}}_{ci}^T & 1\end{bmatrix} \\ &= \begin{bmatrix}\sum_{i=1}^t \widetilde{\mathbf{x}}_{ci} \times \widetilde{\mathbf{x}}_{ci}^T & \sum_{i=1}^t \widetilde{\mathbf{x}}_{ci} \\ \sum_{i=1}^t \widetilde{\mathbf{x}}_{ci}^T & \sum_{i=1}^t 1\end{bmatrix}. \end{aligned} \tag{C.2}$$

Because (B.2), we have

$$\sum_{i=1}^t \widetilde{\mathbf{x}}_{ci} \times \widetilde{\mathbf{x}}_{ci}^T = t \times \mathbf{I}_{w\times w}. \tag{C.3}$$

Because the transformation from $\widehat{\mathbf{x}}_{ci}$ to $\widetilde{\mathbf{x}}_{ci}$ does not change the original amplitude, so we have

$$\begin{aligned} \sum_{i=1}^t \widetilde{\mathbf{x}}_i &= \sum_{i=1}^t \widehat{\mathbf{x}}_i = \mathbf{0}_{w\times 1}, \\ \sum_{i=1}^t \widetilde{\mathbf{x}}_i^T &= \sum_{i=1}^t \widehat{\mathbf{x}}_i^T = \mathbf{0}_{1\times w}, \end{aligned} \tag{C.4}$$

where $\widehat{\mathbf{x}}_i$ is the column vector with zero mean [16]. Substitute (C.3) and (C.4) in (C.2); we have

$$\widetilde{\mathbf{X}}^* \times \widetilde{\mathbf{X}}^{*T} = t \times \mathbf{I}_{(w+1)\times(w+1)}. \tag{C.5}$$

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

Lizhi Cui thanks School of Information Technologies, the University of Sydney, for providing him with a Ph.D. fellowship; thanks are due to Chinese Scholarship Council for providing Lizhi Cui with financial support; the student's no. is 201206740061.

References

[1] M. Maeder, "Evolving factor analysis for the resolution of overlapping chromatographic peaks," *Analytical Chemistry*, vol. 59, no. 3, pp. 527–530, 1987.

[2] M. Maeder and A. Zilian, "Evolving factor analysis, a new multivariate technique in chromatography," *Chemometrics and Intelligent Laboratory Systems*, vol. 3, no. 3, pp. 205–213, 1988.

[3] K. J. Schostack and E. R. Malinowski, "Theory of evolutionary factor analysis for resolution of ternary mixtures," *Chemometrics and Intelligent Laboratory Systems*, vol. 8, no. 2, pp. 121–141, 1990.

[4] H. R. Keller and D. L. Massart, "Peak purity control in liquid chromatography with photodiode-array detection by a fixed size moving window evolving factor analysis," *Analytica Chimica Acta*, vol. 246, no. 2, pp. 379–390, 1991.

[5] O. M. Kvalheim and Y.-Z. Liang, "Heuristic evolving latent projections: Resolving two-way multicomponent data. 1. Selectivity, latent-projective graph, datascope, local rank, and unique resolution," *Analytical Chemistry*, vol. 64, no. 8, pp. 936–946, 1992.

[6] Y.-Z. Liang and O. M. Kvalheim, "Diagnosis and resolution of multiwavelength chromatograms by rank map, orthogonal projections and sequential rank analysis," *Analytica Chimica Acta*, vol. 292, no. 1-2, pp. 5–15, 1994.

[7] R. Tauler, "Multivariate curve resolution applied to second order data," *Chemometrics and Intelligent Laboratory Systems*, vol. 30, no. 1, pp. 133–146, 1995.

[8] R. Tauler, S. Lacorte, and D. Barceló, "Application of multivariate self-modeling curve resolution to the quantitation of trace levels of organophosphorus pesticides in natural waters from interlaboratory studies," *Journal of Chromatography A*, vol. 730, no. 1-2, pp. 177–183, 1996.

[9] X. Shao, Z. Yu, and L. Sun, "Immune algorithms in analytical chemistry," *Trends in Analytical Chemistry*, vol. 22, no. 2, pp. 59–69, 2003.

[10] X. Shao, Z. Liu, and W. Cai, "Resolving multi-component overlapping GC-MS signals by immune algorithms," *Trends in Analytical Chemistry*, vol. 28, no. 11, pp. 1312–1321, 2009.

[11] B. Debrus, P. Lebrun, A. Ceccato et al., "A new statistical method for the automated detection of peaks in UV-DAD chromatograms of a sample mixture," *Talanta*, vol. 79, no. 1, pp. 77–85, 2009.

[12] L. Cui, J. Poon, S. K. Poon et al., "Parallel model of independent component analysis constrained by reference curves for HPLC-DAD and its solution by multi-areas genetic algorithm," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '13)*, pp. 27–28, Shanghai, China, December 2013.

[13] L. Cui, Z. Ling, J. Poon et al., "A parallel model of independent component analysis constrained by a 5-parameter reference curve and its solution by multi-target particle swarm optimization," *Analytical Methods*, vol. 6, no. 8, pp. 2679–2686, 2014.

[14] Z.-M. Zhang, S. Chen, and Y.-Z. Liang, "Peak alignment using wavelet pattern matching and differential evolution," *Talanta*, vol. 83, no. 4, pp. 1108–1117, 2011.

[15] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, pp. 1942–1948, December 1995.

[16] B. Jiang, N. Wang, and L. Wang, "Particle swarm optimization with age-group topology for multimodal functions and data clustering," *Communications in Nonlinear Science and Numerical Simulation*, vol. 18, no. 11, pp. 3134–3145, 2013.

[17] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[18] A. D. Belegundu and T. R. Chandrupatla, *Optimization Concepts and Applications in Engineering*, Cambridge University Press, 2nd edition, 2011.

[19] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons, 1st edition, 1969.