analytical chemistry

Article

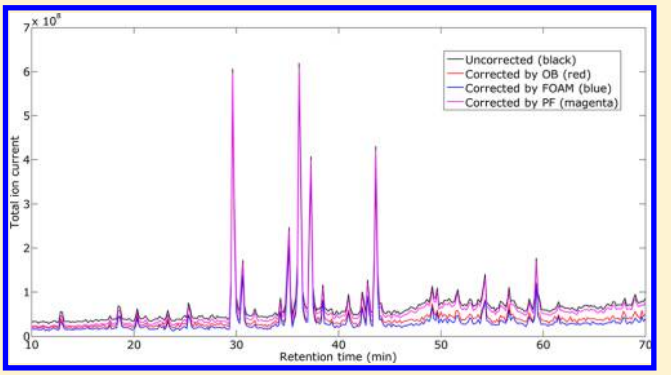
pubs.acs.org/ac

Comparison of Three Algorithms for the Baseline Correction of Hyphenated Data Objects

Zhengfang Wang, Mengliang Zhang, and Peter de B. Harrington*

Center for Intelligent Chemical Instrumentation, Clippinger Laboratories, Department of Chemistry and Biochemistry, Ohio University, Athens, Ohio 45701-2979, United States

ABSTRACT: Three novel two-way baseline correction algorithms, that is, orthogonal basis (OB), fuzzy optimal associative memory (FOAM), and polynomial fitting (PF), were evaluated with high performance liquid chromatography−mass spectrometry (HPLC−MS) and gas chromatography/mass spectrometry (GC/MS) data objects. Among these algorithms, both OB and FOAM are two-way baseline correction algorithms, which reconstruct the entire two-way backgrounds from blank data objects, while the PF algorithm is a pseudo-two-way method, which models each ion chromatogram baseline with a third-order polynomial. The performance of baseline correction methods was first evaluated with respect to the signal-to-noise ratios (SNRs) of 4 major peaks of the HPLC−MS total ion current (TIC) chromatograms of celery seed extracts. Then, the effect of baseline correction on pattern recognition was evaluated by using 42 two-way headspace (HS) solid phase microextraction (SPME) GC/MS data objects of 7 polychlorinated biphenyl (PCB) mixture standard solutions. Two types of classifiers, that is, a fuzzy rule-building expert system (FuRES) and partial least-squares-discriminant analysis (PLS-DA) were evaluated in parallel. Bootstrapped Latin partitions (BLPs) were used to give an unbiased and generalized evaluation of the classification accuracy. Results indicate that SNRs of major peaks of the TIC chromatogram representative of two-way HPLC−MS data objects are increased by baseline correction. In addition, higher prediction accuracies can be obtained by performing baseline correction on the entire GC/MS data set prior to pattern recognition. It is also found that proper data transformation is able to improve the performance of baseline correction. This report is the first of two-way baseline correction methods for hyphenated chromatography/mass spectrometry data objects. Both the orthogonal basis and FOAM baseline correction methods are novel in-house algorithms and proved to be generally effective for two-way baseline correction in the present study. Polynomial fitting is a conventional baseline correction method for one-way data objects and is applied to two-way data objects for the first time. It is applicable when blank data objects are unavailable.



Hyphenated analytical techniques, such as high performance liquid chromatography−mass spectrometry (HPLC−MS) and gas chromatography/mass spectrometry (GC/MS), combine chromatographic and multichannel spectral detectors to exploit the advantages of both methods. However, two-way data objects may contain significant variations that are not pertinent to the measurement especially when temperature or solvent programs are used during the separation stage.[1] For example, a measured signal may consist of the analytical signal, the background signal, and noise.[2] The analytical signal is the systematic response to the analytes, the background is the systematic response not related to the analytes, and noises are associated with random variations.[2] One important reason for noisy two-way signals is that the high sensitivity of mass spectrometers allows trace compounds to be detected at very low levels. Besides, data quality can be adversely affected by small changes of chromatographic conditions, such as unstable flow rates of the mobile phase and column bleed.

Hyphenated measurements by HPLC−MS or GC/MS furnish two-way data objects (i.e., chromatogram × mass spectrum). A variety of methods have been proposed for one-way chromatographic or spectral baseline correction.[3−8] Most of them rely on determining a suitable offset that is subtracted from the original data.[9] Although the total ion current (TIC) chromatogram, which is the summation of the ion current from all mass channels, is still commonly used to process two-way data objects, others have used the total mass spectrum (TMS) or the covariation mass spectrum with respect to retention time. Most commercial baseline correction methods usually are applied to the one-way chromatograms or spectra.[1,10] However, the summation across mass channels or retention times results in a needless loss of information.[11−13]

In the present work, three baseline correction algorithms, that is, orthogonal basis (OB), fuzzy optimal associative

Received: May 5, 2014
Accepted: August 25, 2014

dx.doi.org/10.1021/ac501658k | *Anal. Chem.* XXXX, XXX, XXX−XXX

memory (FOAM), and polynomial fitting (PF), for preprocessing two-way chromatographic/mass spectrometric data objects are compared. Both OB and FOAM are two-way baseline correction methods which reconstruct the best fitting two-way background from blank measurements, while PF is a one-way method which models each mass channel with a third-order polynomial. Related principles are provided in the Theory section.

The performance of three baseline correction methods was first evaluated by using 9 two-way HPLC−MS data objects of celery seed extracts, for which the SNRs of 4 major peaks of the TIC chromatograms before and after baseline correction were compared. Next, the effect of baseline correction on pattern recognition was evaluated by using 42 two-way headspace (HS) solid-phase microextraction (SPME) GC/MS data objects of 7 polychlorinated biphenyl (PCB) mixture standard solutions.[14] Bootstrapped Latin partitions (BLPs) were used to give an unbiased and generalized measure of classification accuracy for the two classifiers, that is, a fuzzy rule-building expert system (FuRES) and partial least-squares-discriminant analysis (PLS-DA). These classifiers were evaluated in parallel so that each classifier had identical training and prediction sets during the bootstrap analysis. Classification accuracies before and after baseline correction were compared.

It is the first time that two-way baseline correction methods for chromatography/mass spectrometry data objects have been reported. Both the orthogonal basis and FOAM baseline correction methods are novel in-house algorithms and proved to be generally effective for two-way baseline correction in the present study. Polynomial fitting is a conventional baseline correction method for one-way data objects and was applied to two-way data objects for the first time. It is useful when blank data objects are unavailable.

Among these algorithms, OB requires a collection of blank data objects and optimization of the rank of the basis; FOAM requires a collection of blank data objects and selection of grid number and fuzzy function; PF does not require blank data objects but it requires the optimization of the polynomial order and fitting threshold.

THEORY

Orthogonal Basis (OB) Baseline Correction. This two-way baseline correction method is based on the singular value decomposition (SVD) algorithm. A collection of two-way data objects of blanks are collected under the same instrumental conditions as the samples. Each blank two-way data object is unfolded into a column by concatenating the ion chromatograms of each mass channel to form a long column vector. The transpose of these vectors are assembled into a matrix for which each row is a blank measurement and the columns correspond to the retention times blocked by each mass channel. Then, this matrix is decomposed by using SVD, as described by eq 1

$$\mathbf{B} = \mathbf{USV}^{\mathbf{T}} + \mathbf{E} \tag{1}$$

for which $\mathbf{B}$ is the $m \times n$ matrix of m blanks and n variables, $\mathbf{U}$ is a $m \times p$ orthogonal matrix of p components, $\mathbf{S}$ is $p \times p$ diagonal matrix with nonnegative real numbers along the diagonal, and $\mathbf{V}$ is a $n \times p$ orthogonal matrix. The diagonal entries of $\mathbf{S}$ are the singular values of $\mathbf{B}$. The p columns of $\mathbf{V}$ comprise the orthogonal basis that will be used for reconstructing the backgrounds.

The orthogonal basis $\mathbf{V}$ is constructed from the blank objects. The number of components or the rank p of the basis $\mathbf{V}$ affects the performance of the baseline correction, especially when the blank objects contain artifact peaks, for example, chromatographic peaks generated from the SPME fiber, septum of the headspace vial cap, or matrix components of the blank sample.[14] Using all the components of $\mathbf{V}$ will model all the information on the blanks, but may result in overfitting during background reconstruction from sample objects. Using values of p that are less than the rank of $\mathbf{B}$ (e.g., p of 5) may reduce overfitting that manifests in negative peaks or attenuated signal peaks in the baseline corrected samples.

Likewise, a collection of the two-way sample data objects constitute a three-way data set, which can be rearranged into a two-way matrix using the same procedure as described previously with each sample data object as a row. The reconstructed background is as follows

$$\hat{\mathbf{X}} = (\mathbf{X_0V})\mathbf{V}^{\mathbf{T}} \tag{2}$$

for which $\hat{\mathbf{X}}$ is the reconstructed background matrix, $\boldsymbol{X_0}$ is the uncorrected data matrix, and $\boldsymbol{V}$ is the orthogonal basis set obtained from the blanks. The reconstructed blank matrix is subtracted from the data matrix, as shown below

$$\mathbf{X}_c = \mathbf{X}_0 - \hat{\mathbf{X}} \tag{3}$$

for which $\mathbf{X}_c$ is the baseline-corrected sample matrix. Lastly, the two-way sample matrix is reshaped again to retrieve the three-way set of the two-way sample data objects.

Fuzzy Optimal Associative Memory (FOAM) Baseline Correction. FOAM is an enhanced optimal associative memory (OAM). It was devised in the 1990s and has been applied to the background estimation of single scan near-infrared (NIR) spectra.[15] A FOAM uses a different paradigm than typical chemometric approaches, that is, the model derives from the processing that occurs when a spectrum or chromatogram is observed by an analyst and the optical cells of the analyst's retina are activated.[15] Very recently FOAMs have been successfully applied to model two-way GC/MS data objects.[13,14,16] The advantage of a FOAM is that an incomplete mass spectrum or a distorted mass spectrum can be reconstructed. This property of FOAMs makes them ideally suited for background correction.

The key features of the FOAM are the fuzzy grid encoding and decoding steps. In the encoding process, a grid of equal sized squares is superimposed on the matrix of blank data objects.[17] Typically each data point defines the abscissa but larger grids could be constructed by binning the variables or measurement channels together. The ordinate grid size or increment is obtained by taking the range of intensity values and dividing the range into bins. Analytical chemists may be familiar with binning measurement channels together, but the concept of binning the intensities is new to chemometrics and chemistry.

For each measurement channel, the grid element which contains the intensity is encoded with a value of unity otherwise the value of the grid is zero. This process maps any vector onto a two-dimensional grid, just as one would fill in the squares of a sheet of graph paper that is overlaid on a spectrum or a chromatogram. A one-way data object is then transformed into a binary image. These binary images can be unfolded into a vector by taking each column of the grid and concatenating them all together to form a single long column.

After the binary encoding, a fuzzy function is applied to the grid, then the construction of basis sets and reconstruction of

dx.doi.org/10.1021/ac501658k | Anal. Chem. XXXX, XXX, XXX−XXX

background can be obtained by using the same procedure as described for the orthogonal basis baseline correction approach. Note that these grid encoded objects are very sparse (i.e., contain mostly values of zero). When the unfolded grid-encoded objects are assembled into a matrix, columns that contain all zero values can be removed, and an index is maintained of the positions of the columns that contain at least a single value of unity for all the objects (i.e., rows) of the data set. This procedure will save a significant amount of memory. Some programming environments provide sparse matrix structures, but matrix operations on these sparse structures are much slower than the simple procedure given here.

The gridding procedure renders peaks or points of the same measurement or variable with different intensities orthogonal to each other. The relationships of convolving points of the binary encoded matrix with a fuzzy function can control the correlation of nearby points (i.e., data points with similar intensities). If the fuzzy function is applied with respect to the ordinate (i.e., intensity) direction, points or peaks with similar intensities (i.e., close but different grid values) will be correlated. Likewise if the fuzzy function is applied with respect to the abscissa (i.e., variable) points with similar retention times or mass to charge ratios will be correlated so that drift or registration problems can be corrected, but that is the subject of another project.[17,18]

In the present study, the number of elements was defined to be 100, which generally is suitable for most problems. Likewise, the triangular fuzzy function is also generally applicable to most problems and will be used throughout this study. The triangular fuzzy function used in this paper is defined as

$$y = 1 - |1 - 0.1 \times x|, \{x = 1, 2, ...19\} \quad (4)$$

To reconstruct the original data set the grid must be decoded. This inverse grid mapping process is quite simple. The grid is reassembled using the indices that were stored during the compression step. Then for each column of the grid, the grid element with the largest value is found. The columns of the grid correspond to the variables of the reconstructed object and the intensity of the reconstructed object is defined by the midpoint of the grid with the maximum value. It is possible to also have overfitting when the reconstructed blank data objects contain larger peaks than the sample data object, and when the reconstructed background is subtracted some negative peaks may occur. In some cases, such as OB, the negative peaks of the background corrected object may be assigned to a zero value. In this study, all the negative peaks that resulted from the correction were retained.

Polynomial Fitting (PF) Baseline Correction. Both OB and FOAM require a collection of blank data objects. When blank objects are unavailable, polynomial fitting methods could be used. The polynomial fitting baseline correction method works on a two-way data object by fitting a third-order polynomial to each ion chromatogram until the entire data object is baseline-corrected. In this sense, polynomial fitting is actually a one-way baseline correction method.

For each ion chromatogram, the algorithm iteratively estimates the polynomial baseline and subtracts it from the ion chromatogram. Outlying data points whose magnitudes are beyond a threshold are excluded from the subsequent iteration. The polynomial is then fit to the uncorrected ion chromatogram with the outliers removed. This procedure is iterated until the algorithm converges. Note, if the threshold is too small, all the points may be eliminated and the algorithm may never converge. The disadvantage of this approach is that it assumes the background of the two-way data is relatively smooth, so it will work well for column bleed, but is ineffective for the removal of artifact peaks.

A third order polynomial was selected for this study, because the third-order polynomials give an approximate estimate of the GC baseline.[6] Higher-order polynomials may cause overfitting and result in a loss of signal.

Signal-to-Noise Ratio (SNR). A chromatographic peak has been defined as a Gaussian shaped peak that contains a maximum signal that is at least three times higher than the noise level.[19] The noise level is defined as the standard deviation obtained from regions where no chromatographic peaks are present. Peak height is measured from the chromatographic peak maximum to the point that is a linear interpolation of the intensity at the start and end point of the chromatographic peak.

For each peak, SNR is calculated by dividing the peak height with the standard deviation of the noise. The noise regions are measured well away from the peak tails. The noise was either 60 s before the peak start point or a combination of 30 s before and after the peak start and end points. The standard deviation was calculated from all the points in these regions to generate the noise estimate. The SNR can measure a decrease in the noise points during the background correction while maintaining the signal intensity.

Classification. Two classification methods, that is, FuRES and PLS-DA, are individually used to construct classification models with a training set, by using the parameters optimized in a previous study.[14] The established classifiers were then used to classify a validation set.

FuRES is a fuzzy classifier while PLS-DA is a crisp one. PLS-DA is a commonly used pattern recognition method, but the inherent fuzziness of FuRES makes it a powerful method for distinguishing overlapped or outlier-containing classes. Both pattern recognition methods are evaluated in parallel in the present work. Related principles can be found elsewhere.[13,14,20]

Data preprocessing is very important because it simplifies classification by reducing noise and decreasing model complexity. In the present study, baseline correction is applied to the entire data set prior to pattern recognition. Prediction accuracies before and after baseline correction are compared.

EXPERIMENTAL SECTION

HPLC−MS Data Objects of Celery Seed Extract. Two-way HPLC−MS data objects were obtained by the chromatographic separation and mass spectrometric detection of celery seed extract. This set of data objects was used to evaluate the quality of the TIC chromatograms after baseline correction with respect to the SNRs of 4 major peaks. Solvent blanks were collected with no parametric change to the instrumental parameters.

Materials and Reagents. HPLC grade acetonitrile (ACN) and methanol (MeOH) were purchased from VWR Scientific (Seattle, WA). Mass spectrometry grade formic acid (FA) was purchased from Sigma-Aldrich (Saint Louis, MI). HPLC grade water was purchased from Fisher Scientific (Pittsburgh, PA).

Dried celery seeds of the same breed were obtained from the Food Composition and Method Development Laboratory, United States Department of Agriculture (USDA). The dried sample was finely powdered and passed through a 20 mesh sieve. The dried celery seed powder (50 mg) was extracted with 5 mL of a MeOH/H_2O (60:40, v:v) solution in a FS30

dx.doi.org/10.1021/ac501658k | Anal. Chem. XXXX, XXX, XXX−XXX

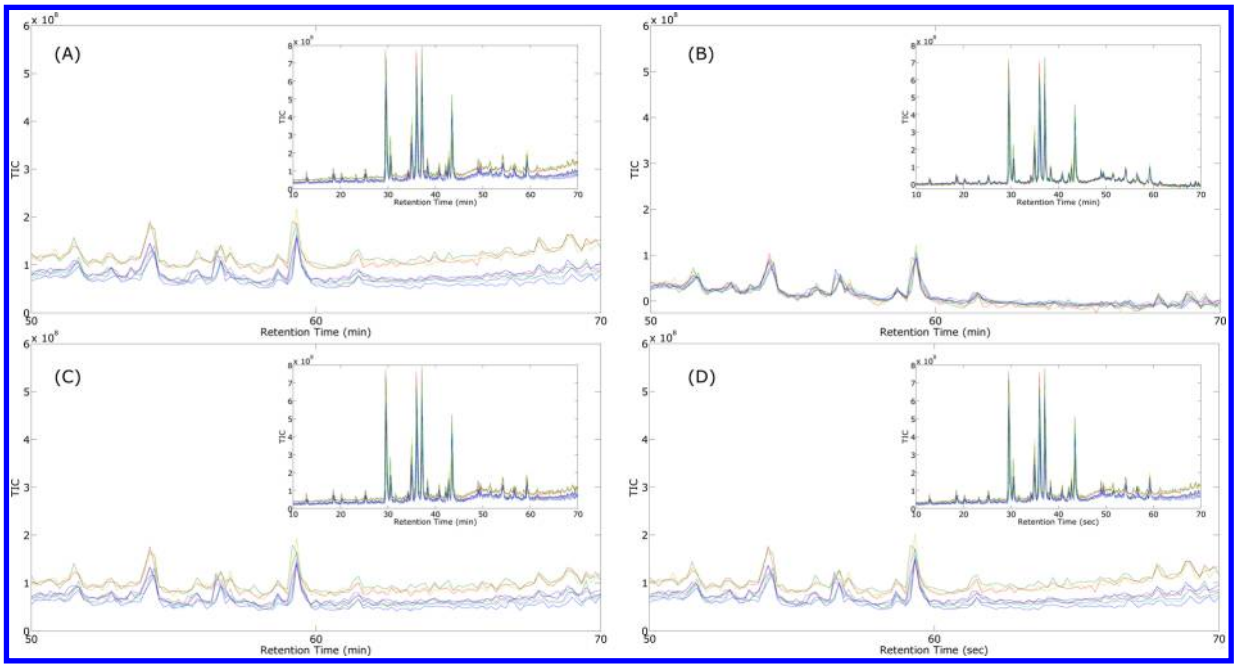


Figure 1. HPLC–MS total ion current (TIC) chromatograms of 9 celery seed extracts from 50 to 70 min (A) without baseline correction, (B) after two-way orthogonal basis (OB) baseline correction, (C) after two-way fuzzy optimal associative memory (FOAM) baseline correction, and (D) after polynomial fitting (PF) baseline correction. The entire TIC chromatograms are provided as inserts.

Ultrasonic sonicator (Fisher Scientific, Pittsburgh, PA) at 40 kHz and 100 W for 60 min at room temperature. The slurry mixture was centrifuged at 5000*g* for 15 min, and the supernatant was filtered through a 0.45 μm polyvinylidene difluoride (PVDF) syringe filter (VWR Scientific, Seattle, WA). A 10 μL aliquot of the extract was then subject to reverse phase HPLC–MS analysis in negative ion mode.

HPLC–MS Data Collection. An Agilent 1200 HPLC system (Agilent Technologies, Santa Clara, CA) was coupled with a Thermo LCQ DECA XP Plus ion trap mass spectrometer (Thermo Scientific, San Jose, CA). The HPLC–MS system was controlled by the Xcalibur software, version 1.4 (Thermo Scientific, San Jose, CA). The HPLC separation was accomplished on a 150 mm × 4.6 mm × 3.5 μm Eclipse XDB-C18 column (Agilent Technologies, Santa Clara, CA) at a flow rate of 1 mL/min. The mobile phase A was H_2O containing 0.1% of FA and the mobile phase B was ACN containing 0.1% of FA. The gradient was linear from 10 to 26% B in 40 min, then to 65% B at 70 min, and finally to 100% B at 71 min. The mobile phase composition was held at 100% B to 75 min. Nine replicates were collected on 9 days. Before each sample run, one blank data object was collected under the same experimental conditions as the sample data objects.

The MATLAB R2013a (MathWorks Inc., Natick, MA) was used to process data, develop and implement the data processing scripts. All the calculations were performed on an Intel Core i7 2.93 GHz personal computer (PC) with 12 GB RAM running a Microsoft Windows XP Professional x64 operation system (Microsoft Corp., Redmond, WA). The HPLC–MS data object was binned by retention time from 10 to 70 min with a 10-s increment, and by mass-to-charge ratio from 100 to 2000 Th with a 1-Th increment. Thereafter, each data object comprised 361 retention time rows and 1901 mass-to-charge ratio columns.

HS-SPME-GC/MS Data Objects of PCB Standard Solutions. The FuRES and PLS-DA classifiers were constructed by using 42 two-way GC/MS data objects of the 7 PCB mixture standard solutions. Headspace solid-phase microextraction was employed as the sample preparation method. Blank data objects were obtained from the SPME of empty headspace vials. They were collected with no parametric change to the experimental parameters. The effects of baseline correction on the classifiers were then evaluated.

Materials and Reagents. Commercial mixtures of PCBs, that is, Aroclor 1016, 1221, 1232, 1242, 1248, 1254, and 1260, at a concentration of 100 μg/mL in MeOH were purchased from AccuStandard Inc. (New Haven, CT). The last two digits of the Aroclor numbers represent the percentage of Chlorine by mass in the mixture. One exception to this nomenclature is Aroclor 1016 that contains 41% Chlorine by mass and not 16%. The potassium dichromate ($K_2Cr_2O_7$), sulfuric acid (H_2SO_4), SPME fibers coated with polydimethylsiloxane (PDMS, 100 μm-thick film), 20 mL headspace glass vials, and crimp seals with PTFE/silicone septa were purchased from Sigma-Aldrich Co. LLC. (St. Louis, MO). Blank soil samples were purchased from RT Corp. (Laramie, WY).

GC/MS Data Collection. All the data were collected on a Thermo Finnigan PolarisQ quadrupole ion trap mass spectrometer/Trace GC system with a Triplus AS2000 autosampler (San Francisco, CA). The GC/MS system was controlled by the XCalibur software, version 2.0.7 (Thermo Scientific Inc., San Francisco, CA). The GC separation was accomplished on a 30 m × 0.25 mm × 0.1 μm 5% diphenyl/95% dimethyl polysiloxane cross-linked capillary column (SHRXI-5MS, Shimadzu Scientific Instruments Inc., Columbia, MD).

The Aroclor stock solution were diluted with MeOH to obtain standard solutions at concentrations of 0.3, 1, and 3 μg/mL in duplicate. A 50 μL aliquot of each standard solution was added to a 20 mL headspace glass vial with seal. After incubation at 100 °C for 5 min, a PDMS fiber was exposed to the headspace for 25 min at 100 °C. The fiber was then

dx.doi.org/10.1021/ac501658k | *Anal. Chem.* XXXX, XXX, XXX–XXX

thermally desorbed in the GC injector at 280 °C for 5 min. The GC oven temperature program was as follows: 50 °C, hold for 1 min; ramp at 20 °C/min to 280 °C, hold for 10 min. The transfer line and the ion source temperature were maintained at 280 °C. The carrier gas helium (99.99% purity) was maintained at a flow rate of 1 mL/min throughout the experiment. Six replicates of an empty vial and each of the 42 PCB standard solutions (i.e., 7 Aroclor stock solutions ×3 concentrations ×2 replicates) were individually collected. A random block design was applied to the data collection process so that any experimental variation would be characterized. Besides, 10 blank data objects were obtained under the same experimental conditions as sample data objects on each day. The arrangement of blank and sample data objects was randomized. A total of 20 two-way blank data objects were collected last.

Each two-way data object was binned by retention time from 4.1 to 22.0 min with a 0.01 min increment and by mass-to-charge ratio from 140 Th to 550 Th with a 1-Th increment. Thereafter, each data object comprised 1801 retention time rows and 411 mass-to-charge ratio columns. Then, each data object was normalized to unit vector length to remove systematic variations caused by slightly varying amounts of samples in different injections.

RESULTS AND DISCUSSION

Effect of Baseline Correction on the TIC Chromatogram. The TIC chromatogram is still the most common and convenient visualization of an HPLC−MS data object. The SNR is usually used to evaluate the quality of a TIC chromatogram. In the first experiment, each of the three baseline correction methods is performed in parallel on the entire three-way data set, which is composed of 9 two-way HPLC−MS data objects, and then the effect of baseline correction on the TIC chromatograms is investigated.

The enlarged HPLC−MS TIC chromatograms from 50 to 70 min of 9 celery seed extracts are provided in Figure 1. The entire TIC chromatograms are inserts. Because these data objects are collected on different days, instrument deviation is obvious before baseline correction (Figure 1A). After orthogonal basis baseline correction, however, baselines are attenuated to a large extent (Figure 1B). For a demonstration, all the negative values will be retained, so any possible negative peaks will be observed. Both FOAM and polynomial fitting are also able to attenuate baselines to some extent, as illustrated in Figure 1C and D. Because background signals are removed from each mass spectrum of a two-way data object, a broad peak of a TIC chromatogram may become sharper after baseline correction.

Note that none of the baselines are zero in Figure 1, because the two-way data objects are corrected and then the intensities across the mass channels are summed to generate the TIC. Because none of the baseline corrections generated negative values, the summation will result in numbers greater than zero. This effect is a key disadvantage of evaluating baseline correction by visualization of the TIC chromatograms.

In addition, the performance of baseline correction is demonstrated more clearly by using one HPLC−MS data object. As illustrated in Figure 2, the TIC chromatograms without, with orthogonal basis, with FOAM, and with polynomial fitting baseline corrections are drawn as black, red, blue, and magenta, respectively. The baseline can be attenuated by each of the baseline correction algorithms while

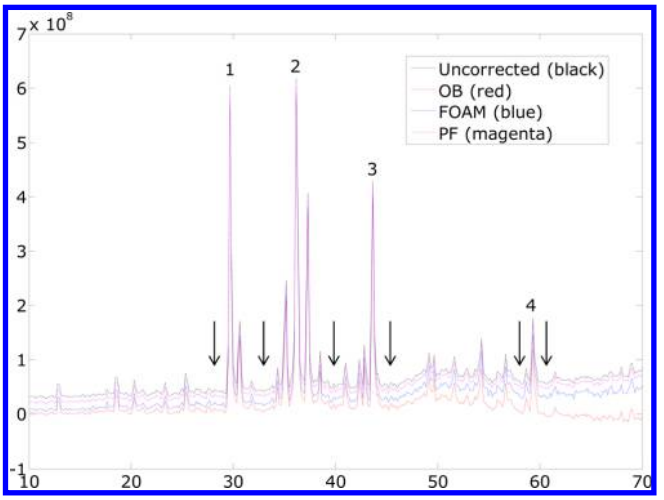


Figure 2. HPLC−MS total ion current (TIC) chromatograms of one celery seed extract without baseline correction (black), after two-way orthogonal basis baseline correction (red), after two-way fuzzy optimal associative memory (FOAM) baseline correction (blue), and after polynomial fitting (PF) baseline correction (magenta). The retention time of 4 major peaks (1) 29.7, (2) 36.2, (3) 43.7, and (4) 59.3 min. Noise regions are indicated by arrows.

the orthogonal basis is the most effective method, which has attenuated the entire baseline to the largest extent.

Four major peaks at retention time of 29.7, 36.2, 43.7, and 59.3 min were chosen for evaluation because these 4 peaks represent different situations. As illustrated in Figure 2, peak 1 at 29.7 min has an adjacent peak at the back, which is not fully separated from peak 1. Peak 2 at 36.2 min is the middle peak of an isotope-like peak cluster. Peak 3 at 43.7 min has an adjacent peak in the front. Peak 4 at 59.3 min is in the region where column bleed was the strongest. Noise regions are also indicated in Figure 2. For Peak 1, which does not have any peak in front of it, the noise sampling region was 60 s before the peak. For peaks 2, 3, and 4, which have small peaks in the front, the noise region is a combination of 30 s before the small peaks and 30 s after the peak of interest.

The average SNRs with 95% confidence intervals of these 4 major peaks of 9 HPLC−MS TIC chromatograms are reported in Table 1. The average SNRs of peaks 1−4 before baseline correction were 3.8 ± 0.7, 2.9 ± 0.5, 1.8 ± 0.2, and 0.6 ± 0.1, respectively. After the orthogonal basis baseline correction, the average SNRs were 6 ± 2, 14 ± 2, 5 ± 3, and 0.8 ± 0.2 for peaks 1−4, respectively. After the FOAM baseline correction, the average SNRs were 7 ± 3, 5 ± 2, 2.6 ± 0.5, and 0.7 ± 0.2,

Table 1. Average Signal-to-Noise Ratios with 95% Confidence Intervals of 4 Major Peaks of the 9 HPLC−-MS TIC Chromatograms of Celery Seed Extracts

RT* (min)	UC	OB	FOAM	PF
29.7	3.8 ± 0.7	6 ± 2	7 ± 3	4.7 ± 0.9
36.2	2.9 ± 0.5	14 ± 2	5 ± 2	3.4 ± 0.7
43.7	1.8 ± 0.2	5 ± 3	2.6 ± 0.5	2.0 ± 0.2
59.3	0.6 ± 0.1	0.8 ± 0.2	0.7 ± 0.2	0.6 ± 0.1

*Denotations: RT, retention time; UC, uncorrected data objects; OB, data objects corrected by using orthogonal basis baseline correction, p of 5; FOAM, data objects corrected by using fuzzy optimal associative memory baseline correction along the intensity direction; PF, data objects corrected by using polynomial fitting baseline correction.

dx.doi.org/10.1021/ac501658k | Anal. Chem. XXXX, XXX, XXX−XXX

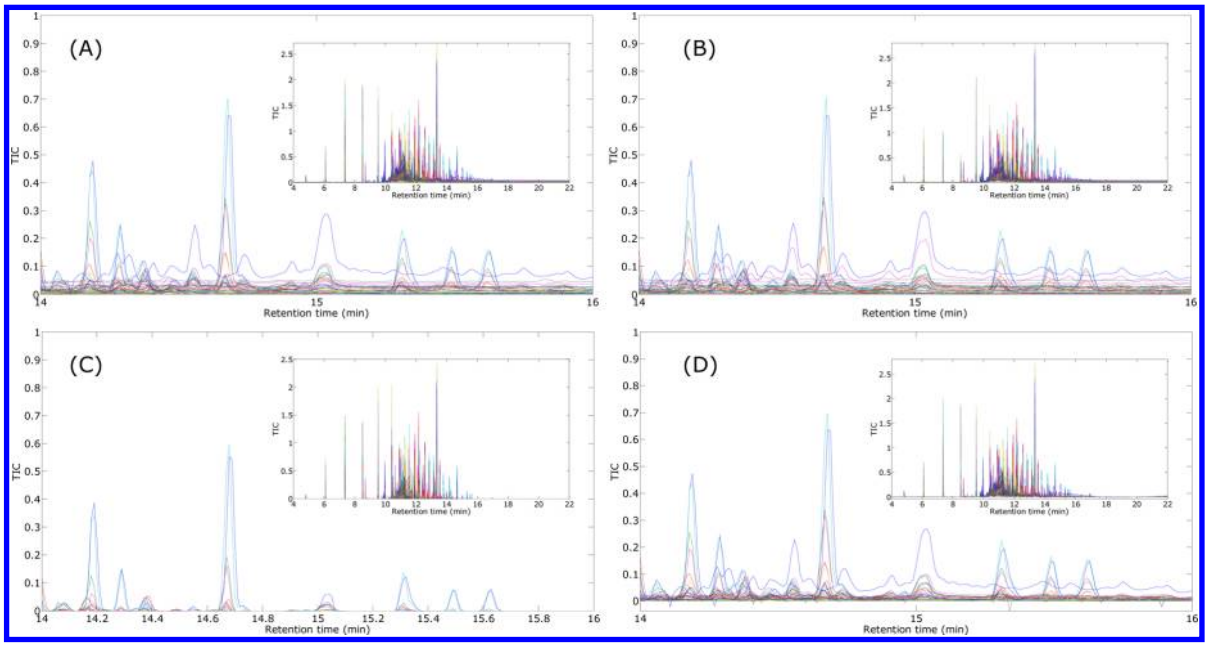


Figure 3. Normalized headspace (HS) solid phase microextraction (SPME) GC/MS total ion current (TIC) chromatograms of 42 polychlorinated biphenyl (PCB) standard solutions from 14 to 16 min (A) without baseline correction, (B) after orthogonal basis baseline correction, (C) after fuzzy optimal associative memory baseline correction, and (D) after polynomial fitting baseline correction. The entire TIC chromatograms are provided as inserts.

respectively. After the polynomial fitting baseline correction, the average SNRs were 4.7 ± 0.9, 3.4 ± 0.7, 2.0 ± 0.2, and 0.6 ± 0.1, respectively. The results indicate that baseline correction has effectively improved the quality of the HPLC−MS TIC chromatograms.

Because 9 data objects were collected on different days, instrument deviation occurred but retention time alignment has yet been performed. Note that the OB baseline correction generates a higher standard deviation, compared with FOAM and PF. This result occurs because OB is least constrained and most susceptible to overfitting.

Effect of Baseline Correction on Classification. Baseline variations will produce common features in data collected from different samples, which may have a deleterious effect on classification or modeling. Therefore, the two-way baseline correction methods are evaluated to see if they may improve classification performance.

For the FOAM, the fuzzy function may be used to blur the grid encoded image with respect to intensity or retention time. The fuzzy function as described by eq 2 is applied along the intensity direction. When the transpose of the fuzzy function is applied, the grid-encoded image will be blurred with respect to the retention time. These two modes of blurring or fuzzification were compared.

In the second experiment, the FuRES and PLS-DA classification models were constructed, from 2 Latin partitions and 100 bootstraps, by using 42 two-way HS-SPME-GC/MS data objects acquired from 7 PCB mixture standard solutions. The enlarged GC/MS TIC chromatograms from 14 to 16 min before baseline correction, after OB baseline correction, after FOAM baseline correction, and after PF baseline correction are provided in Figures 3A−D, respectively. The entire TIC chromatograms are provided as inserts.

The effect of baseline correction on classification can be previewed by a principal component analysis (PCA) score plot for the 7 PCB mixture standard solutions. As illustrated in Figure 4, all the samples cannot be separated without baseline correction. After baseline correction, however, Aroclor 1260, 1254, 1016, and 1221 were individually separated very well from other Aroclor solutions. The performance of baseline correction is further demonstrated by the improvement of classification accuracy.

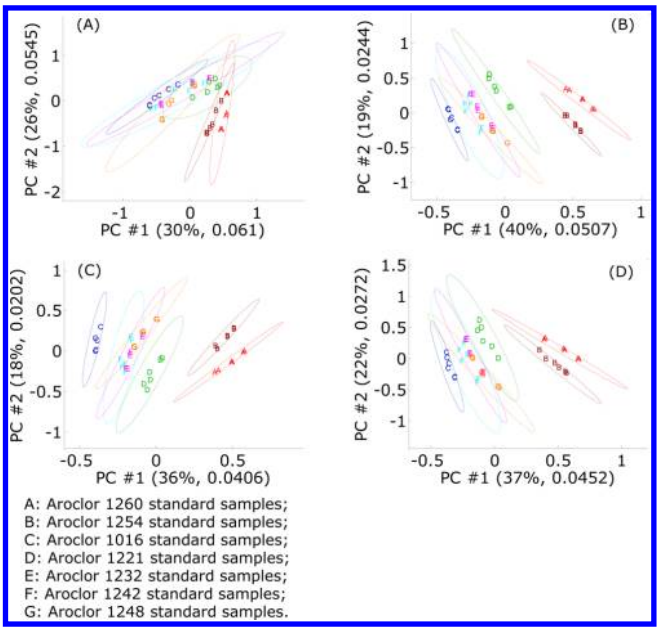


Figure 4. Principal component analysis score plot for normalized two-way HS-SPME-GC/MS data objects of 7 Aroclor standard samples (A) without baseline correction, (B) after the OB baseline correction, (C) after the FOAM baseline correction, and (D) after the PF baseline correction. The 95% confidence intervals are represented by the ellipses.

dx.doi.org/10.1021/ac501658k | *Anal. Chem.* XXXX, XXX, XXX−XXX

The average classification accuracies with 95% confidence intervals of FuRES and PLS-DA classifiers are listed in Table 2.

Table 2. Average Classification Accuracy with 95% Confidence Intervals of FuRES and PLS-DA Classifiers Established by Using 42 HS-SPME-GC/MS Data Objects of PCB Standard Solutions

	FuRES (%)	PLS-DA (%)
UC[a]	85 ± 1	77 ± 1
OB^1	89 ± 1	83 ± 1
OB^2	90 ± 1	87 ± 1
$FOAM^1$	85 ± 1	76 ± 1
$FOAM^2$	77 ± 1	30 ± 1
PF	83 ± 1	76 ± 1
UC_S	93.8 ± 0.9	96.7 ± 0.8
OB^1_S	96.0 ± 0.7	98.7 ± 0.5
OB^2_S	96.2 ± 0.7	98.2 ± 0.6
$FOAM^1$_S	95.9 ± 0.7	97.5 ± 0.6
$FOAM^2$_S	95.5 ± 0.8	98.2 ± 0.6
PF_S	92 ± 1	93 ± 1

[a]Denotations: UC, uncorrected data objects; OB^1, data objects corrected by using orthogonal basis baseline correction, p of 5; OB^2, data objects corrected by using orthogonal basis baseline correction, p of 20; $FOAM^1$, data objects corrected by using fuzzy optimal associative memory baseline correction along the intensity direction; $FOAM^2$, data objects corrected by using fuzzy optimal associative memory baseline correction along the retention time direction; PF, data objects corrected by using polynomial fitting baseline correction; S, by using the square root of the entire data set.

The average classification accuracies of the two-way data objects before baseline correction are 85 ± 1% and 77 ± 1% for the FuRES and PLS-DA classifiers, respectively. After orthogonal basis baseline correction using the reduced basis set (p of 5), the average classification accuracies of the two-way data objects improved to 89 ± 1% and 83 ± 1% for the FuRES and PLS-DA classifiers, respectively. After the orthogonal basis baseline correction using all of the basis set components (p of 20), the average classification accuracies of the two-way data objects improved further to 90 ± 1% and 87 ± 1% for the FuRES and PLS-DA classifiers, respectively.

For the FOAM baseline correction along the intensity direction, the average classification accuracies of the two-way data objects were 85 ± 1% and 76 ± 1% for the FuRES and PLS-DA classifiers, respectively. After the FOAM baseline correction along the retention time direction, the average classification accuracies of the two-way data objects were 77 ± 1% and 30 ± 1% for the FuRES and PLS-DA classifiers, respectively. The FOAM reconstructions were underfitting the data. For the polynomial fitting baseline correction, the average classification accuracies of the two-way data objects were 83 ± 1% and 76 ± 1% for the FuRES and PLS-DA classifiers, respectively.

The MS measurements had a high dynamic range resulting in a large range of intensities. The square root transformation was applied to the GC/MS data before baseline correction. We suspected that this transformation would make the baseline variations much worse and have a negative effect on classification rates. However, the opposite trend occurred and the classification rates improved significantly, even without baseline correction. Many of the early and intense peaks were caused by PDMS fragment ions that arose from the SPME fiber and by decreasing the dynamic range of these spurious components, the classifiers performance improved.

Table 2 gives the average classification accuracies with 95% confidence intervals of FuRES and PLS-DA classifiers of the transformed data sets that yielded values of 93.8 ± 0.9% and 96.7 ± 0.8%, respectively. After the OB baseline correction with the reduced basis set (p of 5), the average classification rates were 96.0 ± 0.7% and 98.7 ± 0.5% for FuRES and PLS-DA, respectively. For the OB baseline correction with the full basis set (p of 20), no significant difference was observed for the classification rates from the uncorrected data.

For the FOAM baseline correction with the fuzzy function applied to the intensities, the average classification rates were 95.9 ± 0.7% and 97.5 ± 0.6%, respectively. For the FOAM baseline correction with the fuzzy function applied along the retention time direction, the average classification accuracies with 95% confidence intervals of FuRES and PLS-DA were 95.5 ± 0.8% and 98.2 ± 0.6%, respectively. For the PF baseline correction, the average classification accuracies with 95% confidence intervals of FuRES and PLS-DA were 92 ± 1% and 93 ± 1%, respectively.

For a 42 × 411 × 1801 sample data set and a 20 × 411 × 1801 blank data set, the execution times of baseline correction algorithms are 7.0 s for OB, 3237.2 s for FOAM, and 418.8 s for PF, on an Intel Core i7 2.93 GHz PC with 12 GB RAM running a Microsoft XP Professional x64 operation system, using the cuptime function on MATLAB R2013b.

OB Compared with Some Other Baseline Correction Methods. The orthogonal basis baseline correction method has been demonstrated to be effective in the present study. Orthogonal signal correction (OSC) was first introduced by Wold and orthogonal projection to latent structures (O-PLS) developed by Trygg and Wold was one of the most main OSC algorithms.[21−23] Projected orthogonal signal correction (POSC) can be seen as a special case of O-PLS. Apart from being an integrated part of the regular partial least-squares (PLS) modeling, the O-PLS method is also used as a preprocessing method.[22] It separates the correlated variation and the noncorrelated variation, and only removes the systematic orthogonal variation (i.e., orthogonal principal components) from a given data set.[22]

Our orthogonal basis method is more straightforward than O-PLS. For OB, additional solvent blank objects are collected under the same experimental condition as the sample objects. All the two-way sample objects constitute a three-way sample data set while all the two-way solvent blank objects constitute a three-way blank data set. Given that the majority part of systematic variations can be explained by the first 5 (sometimes more) principal components of the blank data matrix, background signals of the sample data matrix can be reconstructed from the first 5 (or more) eigenvectors of the blank data matrix. Background correction can be conducted thereafter. Because of the simplicity of OB, it can easily be implemented by nonchemometricians and its execution time on a PC is less than 10 s.

Another well-established baseline correction is penalized asymmetric least-squares (aLS) correction by tensor P-splines developed by Paul Eilers.[24] In Eilers's method, a Whittaker smoother is used to get an estimate of the baseline and P-splines can be extended to two and more dimensions with tensor products of B-splines and appropriate difference penalties.[24] However, there are two parameters that have to be tuned to the data at hand.[24] Besides, although this method

dx.doi.org/10.1021/ac501658k | Anal. Chem. XXXX, XXX, XXX−XXX

has been applied to various types of one-way data objects, including a GC chromatogram, a MALDI-TOF mass spectrum, Raman spectra, a FTIR spectrum, and NIR spectra as well, the results of two-way data object (e.g., a GC/MS data object) baseline correction have yet been provided in Eilers's paper.

On the contrary, the OB method in the present study does not require much tuning and it is directly applied to two-way GC/MS and LC-MS data objects. In addition, OB reconstructs the entire two-way backgrounds from solvent blank objects. In practice, solvent blank objects are always collected after running one or more sample objects to eliminate cross-contamination. Thus, using the OB method will not unnecessarily introduce extra labor into the data collection process.

A recently published baseline correction method named orthogonal spectral space projection has also attracted our attention.[25] This method can be successfully applied to two-way data objects, but the spectra of chromatographic background need to be estimated by using alternating trilinear decomposition (ATLD) because it is difficult to obtain the underlying spectra.[25] Comparatively the OB method in the present study is more robust in that it does not require high-quality spectra or mass spectra. It reconstructs background signals from ordinary blank objects.

CONCLUSIONS

In the present work, three baseline correction methods were investigated by using the two-way HPLC−MS and GC/MS data objects. SNR of major peaks and classification accuracy of FuRES and PLS-DA classifiers were used to evaluate the performance of different algorithms. The effectiveness of baseline correction was also visualized with the TIC chromatograms.

Among these baseline correction methods, OB and FOAM are two-way baseline correction methods which require a collection of blank objects. Besides, OB requires the predetermination of a proper component number for the basis set while FOAM requires the predetermination of a proper grid and a proper fuzzy membership function. PF is a pseudo-two-way baseline correction method that does not need blank objects but requires the predetermination of a proper polynomial order and different polynomial orders may be required for different ion chromatograms of a data object.

With improved sensitivity of modern instrumentation, baseline artifacts will become more prevalent and severe. Two-way baseline correction methods uses blank data objects to improve the analytical data quality by removing these artifacts. Performing two-way baseline correction prior to pattern recognition also improves the prediction accuracy. Therefore, baseline correction is important for the analysis of two-way chromatography/mass spectrometry data objects. Various hyphenated data objects can be subject to analysis in the future to fully exploit the baseline correction methods. Different baseline correction methods for two-way data objects will also be compared in another paper.

AUTHOR INFORMATION

Corresponding Author

*Phone: +01 740-994-0265. Fax: +01 740-593-0148. E-mail: harringp@ohio.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was financially supported by the Center for Intelligent Chemical Instrumentation and Department of Chemistry and Biochemistry at Ohio University. Dr. Longze Lin and Dr. Pei Chen at the United States Department of Agriculture are thanked for providing the celery seed sample.

REFERENCES

(1) Christensen, J. H.; Mortensen, J.; Hansen, A. B.; Andersen, O. *J. Chromatogr., A* **2005**, *1062*, 113−123.
(2) Matos, J. T. V.; Duarte, R. M. B. O.; Duarte, A. C. *J. Chromatogr., B* **2012**, *910*, 31−45.
(3) Schulze, G.; Jirasek, A.; Yu, M. M. L.; Lim, A.; Turner, R. F. B.; Blades, M. W. *Appl. Spectrosc.* **2005**, *59*, 545−574.
(4) Komsta, L. *Chromatographia* **2011**, *73*, 721−731.
(5) Zhang, Z.; Liang, Y. *Chromatographia* **2012**, *75*, 313−314.
(6) Gan, F.; Ruan, G.; Mo, J. *Chemom. Intell. Lab. Syst.* **2006**, *82*, 59−65.
(7) Zhang, Z.; Chen, S.; Liang, Y. *Analyst* **2010**, *135*, 1138−1146.
(8) Boelens, H. F. M.; Dijkstra, R. J.; Eilers, P. H. C.; Fitzpatrick, F.; Westerhuis, J. A. *J. Chromatogr., A* **2004**, *1057*, 21−30.
(9) Krishnan, S.; Vogels, J. T. W. E.; Coulier, L.; Bas, R. C.; Hendriks, M. W. B.; Hankemeier, T.; Thissen, U. *Anal. Chim. Acta* **2012**, *740*, 12−19.
(10) Lee, T. A.; Headley, L. M.; Hardy, J. K. *Anal. Chem.* **1991**, *63*, 357−360.
(11) Lu, Y.; Chen, P.; Harrington, P. B. *Anal. Bioanal. Chem.* **2009**, *394*, 2061−2067.
(12) Rearden, P.; Harrington, P. B. *Anal. Chem.* **2005**, *545*, 13−20.
(13) Wang, Z.; Chen, P.; Yu, L.; Harrington, P. B. *Anal. Chem.* **2013**, *85*, 2945−2953.
(14) Zhang, M.; Harrington, P. B. *Talanta* **2013**, *117*, 483−491.
(15) Wabuyele, B. Wa; Harrington, P. B. *Appl. Spectrosc.* **1996**, *50*, 35−42.
(16) Wang, Z.; Harrington, P. B. *Anal. Bioanal. Chem.* **2013**, *405*, 9219−9234.
(17) Harrington, P. d. B. *Anal. Chem.* **2014**, *86* (10), 4883−4892.
(18) Wang, Z.; Zhang, M.; Harrington, P. Reconstruction of mass spectra using fuzzy optimal associative memories (FOAMs). *Proceedings of the 62nd American Society for Mass Spectrometry Annual Conference on Mass Spectrometry and Allied Topics*, Baltimore, MD, June 15−16, 2014; American Society for Mass Spectrometry, Santa Fe, NM, 2014; WP03-026.
(19) Fredriksson, M.; Petersson, P.; Jörntén-Karlsson, M.; Axelsson, B.-O.; Bylund, D. *J. Chromatogr., A* **2007**, *1172*, 135−150.
(20) Harrington, P. B. *J. Chemom.* **1991**, *5*, 467−486.
(21) Svensson, O.; Kourti, T.; Macgregor, J. F. *J. Chemom.* **2002**, *16*, 176−188.
(22) Trygg, J.; Wold, S. *J. Chemom.* **2002**, *16*, 119−128.
(23) Wold, S.; Antti, H.; Lindgren, F.; Öhman, J. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 175−185.
(24) Eilers, P. H. C.; Boelens, H. F. M. Baseline correction with asymmetric least squares smoothing, 2005. http://www.science.uva.nl/~hboelens/publications/draftpub/Eilers_2005.pdf (accessed on July 20, 2014).
(25) Yu, Y.-J.; Wu, H.-L.; Fu, H.-Y.; Zhao, J.; Li, Y.-N.; Li, S.-F.; Kang, C.; Yu, R-Q. *J. Chromatogr., A* **2013**, *1302*, 72−80.