analytical chemistry

ARTICLE

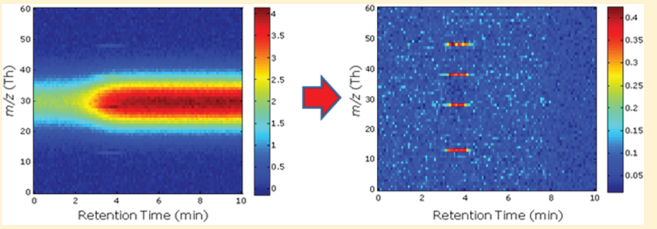
pubs.acs.org/ac

Baseline Correction Method Using an Orthogonal Basis for Gas Chromatography/Mass Spectrometry Data

Zhanfeng Xu, Xiaobo Sun, and Peter de B. Harrington*

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

ABSTRACT: A baseline correction method that uses basis set projection to estimate spectral backgrounds has been developed and applied to gas chromatography/mass spectrometry (GC/MS) data. An orthogonal basis was constructed using singular value decomposition (SVD) for each GC/MS two-way data object from a set of baseline mass spectra. A novel aspect of this baseline correction method is the regularization parameter that prevents overfitting that may produce negative peaks in the corrected mass spectra or ion chromatograms. The number of components in the basis, the regularization parameter, and the mass spectral range from which the spectra were sampled to construct the basis were optimized so that the projected difference resolution (PDR) or signal-to-noise ratio (SNR) was maximized. PDR is a metric similar to chromatographic resolution that indicates the separation of classes in a multivariate data space. This new baseline correction method was evaluated with two synthetic data sets and a real GC/MS data set. The prediction accuracies obtained by using the fuzzy rule-building expert system (FuRES) and partial least-squares−discriminant analysis (PLS-DA) as classifiers were compared and validated through bootstrapped Latin partition (BLP) between data before and after baseline correction. The results indicate that baseline correction of the two-way GC/MS data using the proposed methods resulted in a significant increase in average PDR values and prediction accuracies.



INTRODUCTION

For total ion current (TIC) chromatograms obtained by gas chromatography/mass spectrometry (GC/MS), baseline drift arises from column bleeding due to the thermal degradation or vaporization of the stationary phase when the temperature is elevated.[1] The solvent used and other sources may also cause column bleeding. The baseline could be simply corrected by subtracting a blank spectrum from the spectra if the baseline spectrum is constant. However, oftentimes baselines of chromatograms are not constant during chromatographic runs because of the use of temperature programs in GC and mobile-phase programs in liquid chromatography (LC). In addition, variations may arise from the detector as well. When different samples have common background components, the samples will appear similar to one another, and when the background components vary, replicate measurements will be less precisely defined. Both of these baseline effects may cause errors in classification or pattern recognition.

Several baseline correction methods for mass spectra have been reviewed by Hilario et al.;[2] however, there are still some situations such as small peaks sitting on a large broad peak for which the baseline could not be correctly removed. One approach of correcting a background is to construct an orthogonal basis of the background spectra and then project the spectra with signals and the background onto the basis. The spectra with signals are corrected by subtraction of the basis set projection from the spectra of the chromatographic run. The Gram−Schmidt orthogonalization is one such algorithm to construct bases. A detailed process of the Gram−Schmidt orthogonalization algorithm can be found in the literature.[3,4] Gram−Schmidt orthogonalization was successfully used to reconstruct gas chromatograms from interferograms collected by gas chromatography infrared spectrometry (GC-IR).[3] The Gram−Schmidt algorithm has been used to reconstruct GC-IR chromatograms for quantitative analysis[5] and in interferogram-based infrared search systems;[6] the Gram−Schmidt algorithm has also been coupled to cross-correlation to enhance the signal-to-noise ratio (SNR).[7] A comparison between a Gram−Schmidt orthogonalization and a low-resolution fast Fourier transform (FFT) integration method of gas chromatogram reconstruction demonstrated that the Gram−Schmidt orthogonalization procedure could obtain more sensitive total IR absorption signals if the parameters were optimized carefully.[8] In all of the above-mentioned work, the distances from the orthogonal bases were used to generate chromatograms from the collection of interferograms.

Parameters including the number of interferogram points, the amount of displacement from the interferogram centerburst, and the number of basis vector dimensions were used and were successfully optimized to construct chromatograms for GC-IR with a supermodified simplex process[9] by maximizing the signal-to-noise ratio in the chromatogram.[8] Using SNR as the response to optimize the Gram−Schmidt process is suitable for GC-IR

Received: June 29, 2011
Accepted: August 2, 2011
Published: August 08, 2011

data because the noise is noticeably large which is caused by the relatively low sensitivity of IR spectroscopy through a gold light pipe. However, using SNR as the optimization target may not be suitable for GC/MS data if the TIC chromatograms have higher SNR, which will be presented later in this paper.

Three new innovations were added in this work. First, singular value decomposition (SVD) was implemented instead of the Gram−Schmidt algorithm for constructing the orthogonal basis, which offers the advantages of numerical stability and the generation of an efficient basis. The Gram−Schmidt orthogonalization process may significantly lose the orthogonality[10] and is numerically unstable compared to SVD.[11] The SVD approach was then adapted for removing column bleeding components from GC/MS data objects. SVD has a long history[12] and has been widely used in signal processing[13−15] and chemometrics.[16−18] In this work instead of using the Gram−Schmidt algorithm, SVD was used to construct the basis for modeling and removing the background components. Another innovation was using projected difference resolution (PDR)[19] as the response to select the optimal parameters for baseline correction. Lastly, a direct regularization parameter is introduced to present overcorrection of the spectra during baseline removal that produces negative peaks in the total ion chromatogram or negative peaks in the mass spectrum.

Projected difference resolution is a method to calculate the resolution between classes or groups of objects in a multivariate data space. The PDR resembles chromatographic resolution. The larger the PDR value, the better the separation is among classes in the multidimensional data space. The PDR evaluation has been successfully applied to determine the optimized wavelet filter types and the compression level for the discrete wavelet transform,[20] to compare the performances of GC/MS and gas chromatography−differential mobility spectrometry (GC-DMS) for discriminating among ignitable liquids,[21] to measure the resolutions between bacterial classes using matrix assisted laser desorption/ionization-MS data,[22] and to optimize the data pretreatment steps for GC/MS data of jet fuels.[23] PDR is used to define the response surface for the basis set construction.

Partial least-squares-discriminant analysis (PLS-DA) is a standard multivariate classification method.[24,25] PLS-DA is a useful tool for validating newer classifiers such as fuzzy rule-building expert systems (FuRES) and support vector machines (SVM).[26,27] FuRES is based on the concept of classification trees with a minimal neural network at each branch.[28] The branches of the classification tree are multivariate fuzzy rules that generate fuzzy belief values that propagate through the entire tree. The classification tree provides a simple deductive structure that is amenable to interpretation by the path defined with the maximum belief value. FuRES is a powerful classifier that can generate reproducible models with high prediction accuracies and has been successfully used for several applications.[21,23,26,27,29,30]

In this work, a novel baseline correction method based on a PDR- or SNR-optimized algorithm for GC/MS data was developed. The parameters including the number of background spectra used, the number of basis vectors, and the error threshold were optimized through an experimental design and used for the further optimization of baseline correction algorithm by a simplex search. PDR and SNR were compared for forming a response surface for the simplex search. Two synthetic data sets and a real GC/MS data set of jet fuels were evaluated with the new baseline correction method, and the classification accuracy was evaluated by both FuRES and PLS-DA. The validation process was conducted via bootstrapped Latin partition (BLP), which combines resampling of Latin partitions by randomly dividing the data into calibration and prediction sets while maintaining equal class distributions.[31] Average prediction accuracies with confidence intervals were estimated from the bootstrapped results, providing a generalized measure of accuracy with precision bounds.

THEORY

A SVD process decomposes a matrix into a scores matrix, a diagonal matrix, and a loadings matrix. The decomposition[32] is defined by

$$\mathbf{X} = \mathbf{USV}^{\mathrm{T}} \tag{1}$$

for which $\mathbf{X}$ is a data matrix of background spectra sampled from regions of the chromatogram where there are no analytical signals, $\mathbf{U}$ is the normalized object scores matrix, $\mathbf{S}$ is the diagonal matrix with singular values, and $\mathbf{V}^{\mathrm{T}}$ is the transpose of the variable loadings matrix.

PDR is calculated as the resolution between the objects of two classes and used as the criterion to differentiate these two classes. Two matrices of objects $\mathbf{X}_a$ and $\mathbf{X}_b$ are defined by classes a and b, for which each matrix comprises unfolded two-way data objects in the rows. For GC/MS data, each two-way object (retention times by mass-to-charge ratios) is a matrix, which is unfolded into a vector. To calculate the PDR, the first step is to project the unfolded objects onto a vector defined by the difference between the means of classes a and b,

$$p_i = x_i(\overline{x_a} - \overline{x_b})^{\mathrm{T}} \tag{2}$$

for which p_i is a scalar projection of the ith unfolded two-way data object x_i onto the difference vector defined by the two class means $\overline{x_a}$ and $\overline{x_b}$. The PDR is calculated as

$$Rs_{a,b} = \frac{|\overline{p_a} - \overline{p_b}|}{2(s_a + s_b)} \tag{3}$$

for which $Rs_{a,b}$ is the PDR value for classes a and b, $\overline{p_a}$ and $\overline{p_b}$ are the averages of the projections for each class, and s_a and s_b are the standard deviations of the projections for each class. The geometric mean is also used for multiple classes and given by

$$\tilde{Rs} = \left(\prod_{i=1}^{n} Rs_i\right)^{1/n} \tag{4}$$

$$n = k(k-1)/2 \tag{5}$$

for which Rs_i is the PDR value between two classes obtained from eq 3 and n is the number of pairwise combinations of Rs values for k classes. The number of combinations n is obtained from eq 5. The PDR value provides a simple criterion to evaluate the overall separation of the classes in the data space that is analogous to chromatographic resolution, so it is easy for analytical chemists to understand.

Three parameters are optimized for the baseline correction, which are the number of background vectors, the number of basis vectors, and the error threshold for background subtraction. The background spectra are selected from retention time ranges where there are no peaks but with a significant background from column bleeding components. The number of basis vectors can range from zero to the number of background vectors. If zero basis vectors are selected no correction will occur. The error threshold is a regularization parameter to prevent the overcorrection of spectra that may occur from coincidental correlation between

the analyte spectra and the orthogonal basis. There are two different error threshold regularizations developed to correct the baseline. The first one is referred to as Smartbaseline1 that prevents any negative peaks from overfitting in the TIC caused by correction. The second one is referred to as Smartbaseline2 that prevents any negative peaks in mass spectrum during correction. Smartbaseline1 works as follows:

$$\lambda = \frac{\bar{x} - e}{\bar{x} - \bar{x_c}} \tag{6}$$

$$x_c = x - \lambda(x\mathbf{V})\mathbf{V}^T \tag{7}$$

for which $\bar{x}$ is the average of the mass peaks in the uncorrected mass spectrum x, $\bar{x_c}$ is the average of corrected mass spectrum, e is an error threshold that determines the smallest allowable negative intensity, and $\mathbf{V}$ is the orthogonal basis. Equations 6 and 7 are executed as long as $\bar{x_c}$ is less than the error threshold.

Smartbaseline2 is defined by the following equations:

$$x_b = (xV)V^T \tag{8}$$

$$i_{min} = min((x(i) - e)/x_b(i)) \tag{9}$$

$$s = x(i_{min}) - x_b(i_{min}) \tag{10}$$

$$\lambda = \frac{x(i_{min}) - e}{x_b(i_{min})} \tag{11}$$

for which x_b is the projection of mass spectrum on the orthogonal basis, i is the index of all the positive peaks in x_b, i_{min} is the index of the minimum that is based on the minimum calculation in eq 9, and s is the difference between the minimum positive peak in the original mass spectrum and its projected background, which is negative. Equations 11 and 7 are executed as long as s is less than the error threshold.

There are two responses used to determine the optimal parameters in baseline correction in this study. The first method determines the optimal parameters by maximizing the SNR for each two-way GC/MS object. The second method obtains the optimal parameters by maximizing the PDR for the whole data set. In the PDR-optimized baseline correction process, a collection of two-way data objects is individually corrected (i.e., a basis set is calculated for each object). Then, the two-way GC/MS data are unfolded into vectors, and the PDR value is calculated for the training set of objects.

When using SNR as the response to optimize the parameters, the SNR is defined in eq 12 [8] and is calculated from the total ion current chromatogram of two-way GC/MS data. The background vectors are split into two halves. The first half is used to calculate SNR, while the second half is used to build the orthogonal basis.

$$\mathrm{SNR} = \frac{x_{max}}{(\sum_{i=1}^{n} b_i^2/n)^{1/2}} \tag{12}$$

for which x_{max} is the largest chromatograms peak and b_i is the intensity of the background chromatogram i from the first set.

The optimization process of the background correction works as follows. A factorial experimental design of two levels and three factors with eight experiments is expanded to 15 with added experiments at center and faces of the experimental design cube. Then, these 15 background corrections of the training data set were implemented. The response surface is defined by either the resulting average SNR or the PDR for the data set. The response surface is modeled with a three-variable quadratic polynomial that includes interaction terms.

Table 1. Example of Expanded Factorial Experimental Design with Two Levels and Three Factors for the Synthetic Data Set with Two Classes[a]

	n_b	n_v	e	PDR
levels	[15, 23, 30]	[1, 8, 15]	[-1, −0.5, -1.0×10^{-6}]	
coded coefficient	−1	−1	−1	14.2
	1	−1	−1	13.0
	−1	1	−1	15.0
	1	1	−1	11.8
	−1	−1	1	1.8
	1	−1	1	1.4
	−1	1	1	1.3
	1	1	1	1.0
	0	0	−1	12.8
	0	0	1	1.0
	0	1	0	2.1
	0	−1	0	2.5
	1	0	0	4.2
	−1	0	0	2.7
	0	0	0	1.1

variables	b	variables	b	variables	b
1	8.65	$n_b \times n_v$	-4.41×10^{-3}	n_v^2	-2.59×10^{-3}
n_b	−0.752	$n_b \times e$	0.119	e^2	17.8
n_v	0.107	$n_v \times e$	−0.0188		
e	3.25	n_b^2	0.0178		

[a] n_b is the number of background vectors used, n_v is the number of basis vectors used, e is the error threshold, and b is the polynomial coefficients.

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1x_2 + b_5x_1x_3 + b_6x_2x_3 + b_7{x_1}^2 + b_8{x_2}^2 + b_9{x_3}^2 \tag{13}$$

for which y is the response, b_0, b_1, ..., b_9 are the polynomial coefficients, and x_1, x_2, and x_3 are the variables. The model is fit to the response surface. The coded experimental design and a sample calculation are given in Table 1. Then a simplex search is applied to the model to obtain the background correction parameters that yield the highest PDR or SNR. Once the three optimized parameters are obtained from the training set, they may be used to correct the prediction objects.

The PLS-DA algorithm was performed using the nonlinear iterative partial least-squares (NIPALS) algorithm and the PLS2 mode.[33] The dependent-variable matrix was a binary encoded matrix of classes. The algorithm was enhanced so that the number of latent variables is determined automatically.[34] Each component maximizes the covariance of the data set with the dependent matrix. Some latent variables with a low covariance will characterize noise in the spectra. Therefore, the number of latent variables used was optimized for each model. The bootstrapped Latin partition is used within the PLS-DA function to determine the number of latent variables. The training data set is split into two partitions and bootstrapped 10 times. One partition is used for model building, and the other is used for

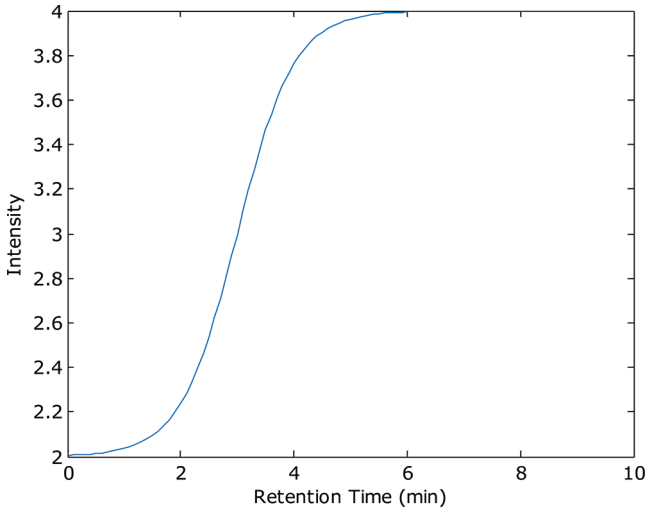


Figure 1. Plot of the function in eq 15 used to mimic the baseline variation.

prediction; then, the roles of the partitions are switched so that each partition is used one time for calibration and one time for prediction. The prediction errors are pooled between the two partitions. The number of latent variables that yielded the lowest average predicted residual error sum of squares (PRESS) within the training data set is selected for prediction.

With FuRES, a classification tree is obtained on the basis of the minimum entropy of classification. FuRES is based on the inductive dichotomizer 3 (ID3) algorithm[35] used by a rule-building expert system that was modified to develop multivariate rules and use the fuzzy set theory to optimize the logistic function of the rules.[36] Principal component transformation was performed to furnish lossless compression before applying FuRES to reduce the computation time. A detailed description of the FuRES algorithm has been described.[28] A key advantage of FuRES compared to PLS-DA is that there are no adjustable parameters to optimize.

EXPERIMENTAL SECTION

Two synthetic GC/MS data sets were created to initially evaluate the method. The retention time of the synthetic data sets ranged from 0 to 10 min with an increment of 0.1 min. The mass-to-charge ratio (m/z) of the synthetic data sets ranged from 0 to 60 Th with an increment of unit Th. Two classes comprised 10 objects in each class with added background noise for one synthetic data set and three classes with 10 objects for another synthetic data set. Each peak was a Gaussian peak created by eq 14 with different amplitudes. The Gaussian function in eq 14 was also used to mimic a background baseline in the synthetic GC/MS data sets. The background noise formed the baseline drift given in eq 14 with a position of 0.5 and a standard deviation of 0.1.

$$f(x) = a \exp\left(-\frac{(x-b)^2}{2c^2}\right) \tag{14}$$

for which a is the amplitude, b is the position or the center of the Gaussian peak, and c is the standard deviation that determines the shape of the Gaussian peak. The variation of baseline was created and added to the baseline by replacing a in eq 14 with eq 15 to mimic column bleeding.

$$f(t) = k + A\frac{1}{1 + \exp\big(-(t-B)/C\big)} \tag{15}$$

for which k is a constant, A is the amplitude, t is the retention time, B determines the center position, and C determines the slope of the bleeding. The plot of eq 15 with $k = 2$, $A = 2$, $B = 3$, and $C = 0.5$ is given in Figure 1. In the synthetic baseline, the column bleeding increases while the temperature of the GC oven increases and reaches a plateau at the end of the temperature ramp. There were 20 spectra in the first synthetic GC/MS data set and 30 spectra in the second one. Each spectrum was a 101 × 61 matrix with one spectrum in each row.

Jet fuel samples were analyzed by GC/MS to furnish the real data set. The jet fuel samples were provided by the Air Force Research Laboratory of Wright-Patterson Air Force Base (Dayton, OH). Five samples were chosen randomly and diluted 50 times with pentane of HPLC grade (Sigma Aldrich). All of the samples were freshly prepared and measured following a random block experimental design within a week. There were five samples including four JP-8 samples and one Jet A sample. Five replicates were measured for each sample.

All experimental data were collected on a Trace-GC 2000 gas chromatograph equipped with a Thermo Finnigan PolarisQ quadrupole ion trap mass spectrometer[37] (Thermo Electron Corp., San Francisco, CA, USA) as the detector. The Xcalibur software version 1.4 was used for the instrument control and data collection. A chemical ionization[38] source operated under the positive ion mode with isobutene (99.00%, Airgas) as the reagent gas at a flow rate of 0.6 mL/min was applied. A (5%-diphenyl)dimethylpolysiloxane (DB-5) column (Agilent Technologies) of 30 m long, a 0.25 mm i.d., and a coating thickness of 0.25 μm was used. The initial temperature was set at 50 °C, and it was held for 5 min. Then with a ramp of 10 °C/min the temperature was raised to 220 °C, and it was held for 5 min. A helium flow rate of 1.5 mL/min was used.

The retention time range of the GC/MS data was from 2.0 to 25.0 min with an increment of 0.033 min; a mass-to-change ratio range of 60−425 Th was selected for the full scan while each spectrum comprised a 2611 × 5475 (retention times by mass-to-charge ratios) matrix before the wavelet transformation. The real data set was compressed to 1/16 of its original size using two-way wavelet compression to reduce the computational load. The biorthogonal Villasenor wavelets with a decomposition level of two for each way (i.e., retention time and mass-to-charge ratio) were used. The compressed data set was 653 × 1369. The total ion current chromatograms constructed from the GC/MS runs of all the samples after wavelet transformation are given in Figure 2. The drift of baseline in each spectrum is apparent in the spectra.

All the calculations were performed with MATLAB R2010b Version 7.11.0.584 with Optimization Toolbox R2010b Version 5.1 (The Mathworks, Natick, MA). The PDR, PLS-DA, and FuRES algorithms were in-house-developed MATLAB programs. The wavelet transforms were implemented with the functions in Wavelab Toolbox Version 850.[39] All of the calculations were performed on an Intel Core i7 2.93 GHz personal computer with 12 GB RAM running a Microsoft Windows XP Professional x64 operation system with Service Pack 2 (Microsoft, Redmond, WA).

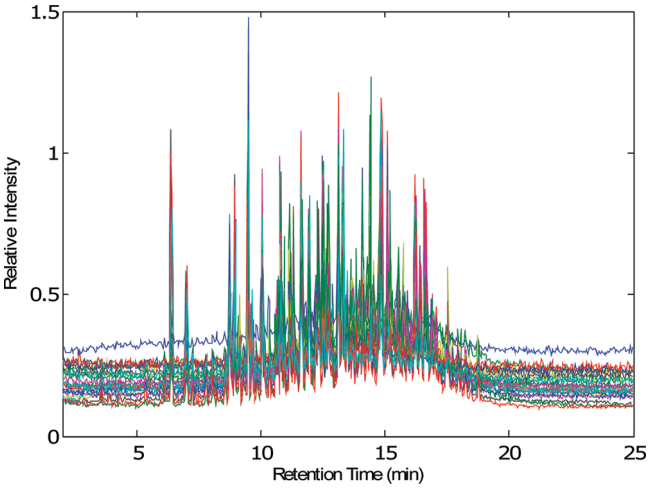


Figure 2. Reconstructed TIC chromatograms of the real GC/MS data set.

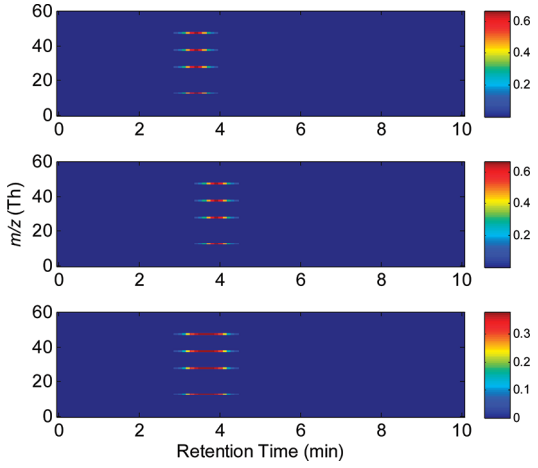


Figure 3. Heat map of the averaged synthetic GC/MS data set of class 1 (top), class 2 (middle), and mix of class 1 and class 2 (bottom) without baseline. Two classes are overlapped with a high similarity. The color bar gives the intensity scale of the MS peaks.

Each two-way data object was unfolded to a vector and normalized to unit length. The PDR values were calculated and bootstrapped 10 times so that the precision of each averaged PDR value could also be ascertained. The "fminsearch" function from the MATLAB Optimization Toolbox was used to locate the minimum of the negative PDR or SNR.

The PLS-DA and FuRES classifiers were compared according to their prediction accuracies through validation. All of the models built by using PLS-DA and FuRES were validated via bootstrapped Latin partitions. Ten bootstraps with three partitions were used, and each object was used once and only once for prediction in each bootstrap. The prediction results from the three partitions were then pooled within each bootstrap. The pooled prediction results obtained from the 10 bootstraps were averaged and the prediction accuracies of the different models were reported with 95% confidence intervals. Two-way analysis of variance (ANOVA) was applied to compare the different baseline correction methods and the different classifiers.

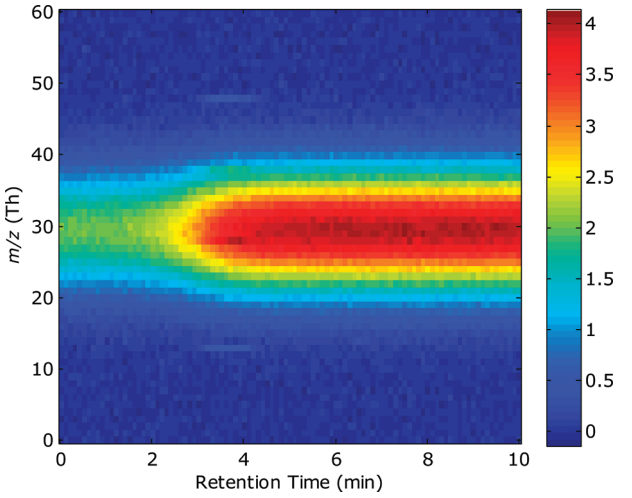


Figure 4. Heat map of the averaged synthetic GC/MS data set of two classes with added baseline. The color bar gives the intensity scale of the MS peaks.

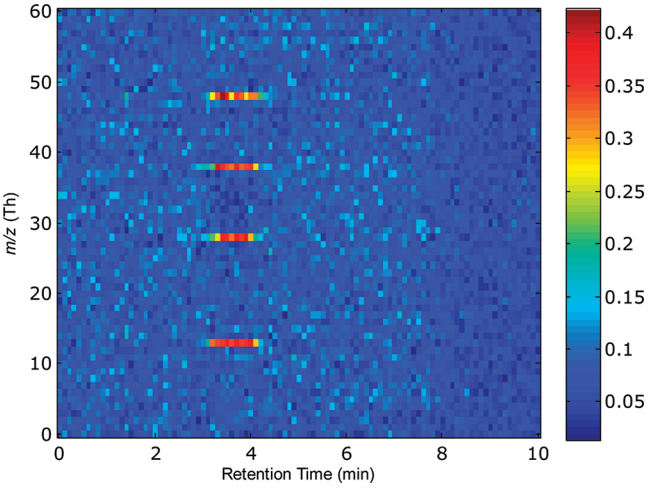


Figure 5. Average of the synthetic GC/MS data set with two classes after PDR-optimized baseline correction using the Smartbaseline1 method. The color bar gives the intensity scale of the MS peaks.

RESULTS AND DISCUSSION

The synthetic GC/MS data set with two classes was averaged and is given as a heat map in Figure 3 without baseline and in Figure 4 with baseline added. The two classes are obviously overlapped as given in Figure 3, and the peaks in the spectrum became less defined in Figure 4 with a PDR value of 0.75 ± 0.05 after the baseline components were added to the data. The average spectra after applying the PDR-optimized baseline correction using Smartbaseline1 are given in Figure 5, which displays visible and very sharp peaks after baseline correction with a PDR value of 11.2 ± 0.7. The average spectra after applying SNR-optimized baseline correction with a PDR value of 11.7 ± 0.7 (not given here) were similar to those in Figure 5.

The PDR values and prediction accuracies of PLS-DA and FuRES for the two class synthetic GC/MS data set with different processing methods are given in Table 2 for comparison. The PDR values were significantly improved after using baseline correction with either of the optimization method. The prediction accuracies obtained by both PLS-DA and FuRES gained significant increases

dx.doi.org/10.1021/ac2016745 |*Anal. Chem.* 2011, 83, 7464–7471

Table 2. Comparison of Prediction Accuracies Obtained by PLS-DA and FuRES for the Synthetic GC/MS Data Set with Two Classes[a]

	PDR	PLS-DA (%)	FuRES (%)
$\mathbf{X}$	0.75 ± 0.05	66 ± 10	74 ± 8
$\mathbf{X}^{n1}$	10.7 ± 0.7	95 ± 2	97 ± 2
$\mathbf{X}^{p1}$	11.2 ± 0.4	99 ± 2	98 ± 3
$\mathbf{X}^{s1}$	11.7 ± 0.7	100 ± 1	99 ± 2
$\mathbf{X}^{n2}$	10.7 ± 0.8	91 ± 5	96 ± 4
$\mathbf{X}^{p2}$	11.5 ± 0.3	99 ± 2	99 ± 2
$\mathbf{X}^{s2}$	10.7 ± 0.7	100 ± 0	100 ± 0

[a] $\mathbf{X}$ is the original synthetic data set; $\mathbf{X}^n$ is the data set after the baseline correction without parameter optimization; $\mathbf{X}^p$ is the data set after the baseline correction by using the PDR-optimized method; $\mathbf{X}^s$ is the data set after the baseline correction by using the SNR-optimized method; 1 and 2 in the superscript indicate the use of Smartbaseline1 and Smartbaseline2, respectively.

Table 3. Comparison of Prediction Accuracies Obtained by PLS-DA and FuRES for the Synthetic GC/MS Data Set with Three Classes[a]

	PDR	PLS-DA (%)	FuRES (%)
$\mathbf{X}$	0.81 ± 0.02	68 ± 2	89 ± 5
$\mathbf{X}^{n1}$	5.8 ± 0.2	84 ± 2	97 ± 1
$\mathbf{X}^{p1}$	7.6 ± 0.2	91 ± 3	100 ± 0
$\mathbf{X}^{s1}$	7.3 ± 0.2	93 ± 1	100 ± 0
$\mathbf{X}^{n2}$	5.1 ± 0.2	82 ± 2	96 ± 1
$\mathbf{X}^{p2}$	7.4 ± 0.2	93 ± 3	100 ± 0
$\mathbf{X}^{s2}$	7.0 ± 0.2	95 ± 1	100 ± 0

[a] $\mathbf{X}$ is the original synthetic data set; $\mathbf{X}^n$ is the data set after the baseline correction without parameter optimization; $\mathbf{X}^p$ is the data set after the baseline correction by using the PDR-optimized method; $\mathbf{X}^s$ is the data set after the baseline correction by using the SNR-optimized method; 1 and 2 in the superscript indicate the use of Smartbaseline1 and Smartbaseline2, respectively.

and reached a value equal or close to 100% after baseline correction with a SNR-optimized method through simplex search or a PDR-optimized method through simplex search in the calibration process. The result without optimization had the number of basis vectors equal to the number of background vectors and an error threshold of −1. Significant increases of prediction accuracies were achieved when the PDR-optimized method was compared to the results obtained without optimization with a p-value of 3.9×10^{-4}, and the SNR-optimized method was compared to the results obtained without optimization with a p-value of 1.4×10^{-5}. This result indicates that it is necessary to optimize the parameters for the baseline correction. The results obtained with a three-class synthetic data set are given in Table 3. A nominally large PDR value was obtained as displayed in both Tables 2 and 3 by using the optimal parameters obtained through simplex search.

The mean TIC chromatograms of the real GC/MS run before and after applying different baseline correction methods are given in Figure 6. As given in the figure, the baseline was successfully removed from the original TIC chromatograms without losing peak information, and both the PDR-optimized and SNR-optimized methods showed their utmost power to correct the baseline. The TIC chromatograms obtained by using the

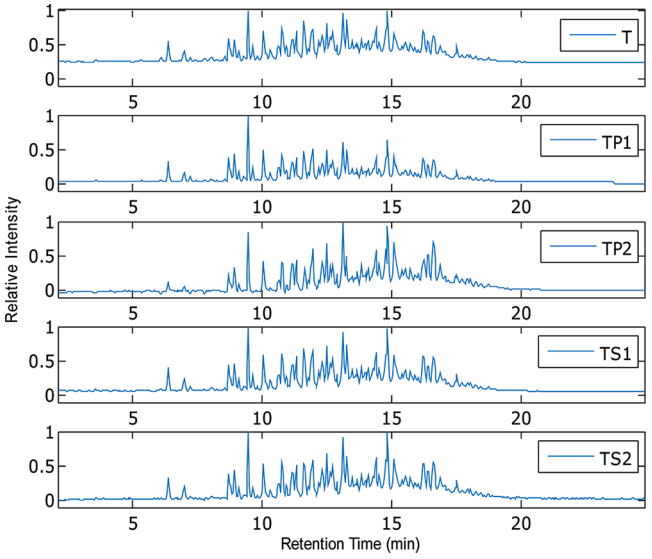


Figure 6. The mean TIC chromatogram of GC/MS run and the mean TIC chromatograms after applying the four different baseline correction methods. T is the original TIC chromatogram; TP1 is the TIC chromatogram with the baseline corrected by using the PDR-optimized Smartbaseline1 method; TP2 is the TIC chromatogram with the baseline corrected by the using PDR-optimized Smartbaseline2 method; TS1 is the TIC chromatogram with the baseline corrected by using the SNR-optimized Smartbaseline1 method; TS2 is the TIC chromatogram with the baseline corrected by the using SNR-optimized Smartbaseline2 method.

Table 4. Comparison of Prediction Accuracies Obtained by PLS-DA and FuRES for the Real GC/MS Data Set[a]

	PDR	PLS-DA (%)	FuRES (%)
$\mathbf{X}$	1.5 ± 0.2	86 ± 2	82 ± 3
$\mathbf{X}^{n1}$	4.6 ± 0.2	88 ± 2	92 ± 2
$\mathbf{X}^{p1}$	4.2 ± 0.3	98 ± 2	99.6 ± 0.9
$\mathbf{X}^{s1}$	4.5 ± 0.3	97 ± 3	99 ± 2
$\mathbf{X}^{n2}$	3.0 ± 0.3	89 ± 3	89 ± 3
$\mathbf{X}^{p2}$	4.1 ± 0.2	97 ± 3	99.6 ± 0.9
$\mathbf{X}^{s2}$	4.5 ± 0.4	98 ± 2	100 ± 0

[a] $\mathbf{X}$ is the original synthetic data set; $\mathbf{X}^n$ is the data set after the baseline correction without parameter optimization; $\mathbf{X}^p$ is the data set after the baseline correction by using the PDR-optimized method; $\mathbf{X}^s$ is the data set after the baseline correction by using the SNR-optimized method; 1 and 2 in the superscript indicate the using of Smartbaseline1 and Smartbaseline2, respectively.

SNR-optimized method and the PDR-optimized method were very similar except minor differences at the front and the back part of the chromatogram. This similarity is reflected in the PDR values and prediction accuracies obtained for both the synthetic data sets and the real data set. The prediction accuracies obtained by applying the PLS-DA and FuRES classifiers to the real GC/MS data set are given in Table 4. Significant increases in the PDR values and prediction accuracies were obtained with both the PLS-DA and FuRES classifiers. The significance of optimized baseline correction was confirmed by two-way ANOVA analysis with interaction between the uncorrected data and the data processed by using the optimized methods with a p-value of 8×10^{-16} for results obtained by PDR-optimized method and a p-value

dx.doi.org/10.1021/ac2016745 |Anal. Chem. 2011, 83, 7464–7471

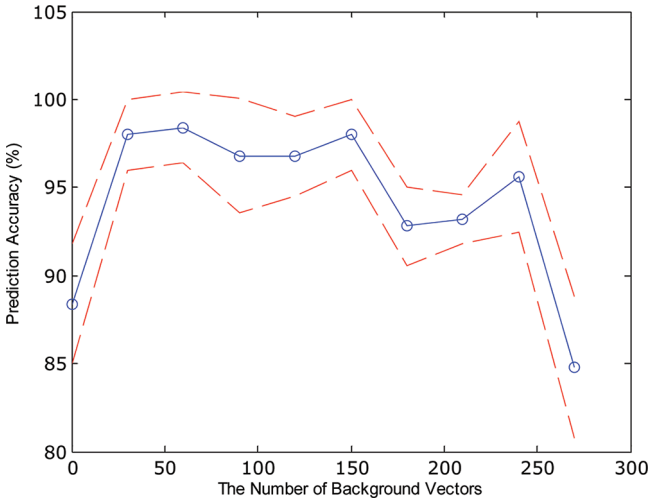


Figure 7. Prediction accuracy varies with the largest number of background vectors used for the real data set. The PDR-optimized Smartbaseline1 method was used. The classifier used is the PLS-DA. The 95% confidence intervals are given in dashed lines.

of 2×10^{-13} for results obtained by the SNR-optimized method. The average FuRES predictions for the PDR baseline corrected data were near perfect with one sample being misidentified from the 250 trials (25 samples $\times$10 bootstraps).

The number of background vectors needs to be selected very carefully. If the number of background vectors is too small the basis will inadequately model the background noise while an excessive number of background vectors will overcorrect the signal and cause the loss of useful information when spectra from analytical peaks are accidentally included in the set of backgrounds. Figure 7 gives the change of prediction accuracy with respect to the number of background vectors used. The classifier used was PLS-DA, and the parameters used for baseline correction were from the PDR-optimized Smartbaseline1 method. A significant decrease of prediction accuracy occurred when the number of background vectors used was greater than 180 (using the final 7.2 min), where spectra from GC peaks were included in the basis. The prediction accuracy did not vary significantly when the number of background spectra varied between 30 (the final 1.2 min) to 150 (the final 6 min). The reason may be because no GC peaks are included within these regions and the basis did not contain analyte mass spectra.

No significant difference between PLS-DA and FuRES prediction accuracies were obtained from the synthetic data set with two classes and the real data set. The prediction accuracies obtained by PLS-DA and FuRES were significantly different when they were applied to the synthetic data set with three classes. Two-way ANOVA was used to compare the prediction accuracies of real data set after baseline correction by using PLS-DA and FuRES across the bootstraps and yielded a p-value of 1.7×10^{-9} that indicates that FuRES performed significantly better than PLS-DA. Both methods using PDR or SNR as the response to optimize the baseline correction parameters could obtain consistently good prediction accuracies with FuRES classifiers. No significant differences were observed between Smartbaseline1 and Smartbaseline2.

CONCLUSION

The results demonstrate that the proposed PDR-optimized and SNR-optimized baseline correction methods work with both synthetic data and the real GC/MS data. Because PDR values vary with the number of background vectors, the number of basis vectors, and the error threshold, it is necessary and efficient to apply a simplex search to find the optimal parameters for the baseline correction. The baseline in both the synthetic data sets and the real GC/MS data set could be successfully removed by applying the PDR- and SNR-optimized methods. The prediction accuracies obtained by using the FuRES and PLS classifiers were both significantly improved along with the increases in the PDR values. Both error threshold regularizations, Smartbaseline1 and Smartbaseline2, have good performance without significant differences and effectively prevented negative peaks from occurring during the baseline correction. Compared to PLS-DA, the FuRES classifier had relatively constant performance regardless of data sets.

AUTHOR INFORMATION

Corresponding Author

*E-mail: Peter.Harrington@OHIO.edu.

ACKNOWLEDGMENT

Weiying Lu is thanked for his helpful comments. The Air Force Research Laboratory of Wright-Patterson Air Force Base is thanked for providing the jet fuels samples.

REFERENCES

(1) Gross, J. H. *Mass Spectrometry: A Textbook*; Springer: Heidelberg, Germany, 2004; p 483.
(2) Hilario, M.; Kalousis, A.; Pellegrini, C.; Müller, M. *Mass Spectrom. Rev.* **2006**, *25*, 409–449.
(3) Haseth, J. A. d.; Isenhour, T. L. *Anal. Chem.* **1977**, *49*, 1977–1981.
(4) Chow, T. L., *Mathematical Methods for Physicists: A Concise Introduction*; Cambridge University Press: Cambridge, U.K., 2000; p 209.
(5) Sparks, D. T.; Lam, R. B.; Isenhour, T. L. *Anal. Chem.* **1982**, *54*, 1922–1926.
(6) Haseth, J. A. d. *Anal. Chem.* **1981**, *53*, 2292–2296.
(7) Lam, R. B.; Sparks, D. T.; Isenhour, T. L. *Anal. Chem.* **1982**, *54*, 1927–1931.
(8) White, R. L.; Glss, G. N.; Brissey, G. M.; Wilkins, C. L. *Anal. Chem.* **1981**, *53*, 1778–1782.
(9) Routh, M. W.; Swartz, P. A.; Denton, M. B. *Anal. Chem.* **1977**, *49*, 1422–1428.
(10) Giraud, L.; Langou, J.; Rozloznik, M. *Comput. Math. Appl.* **2005**, *50*, 1069–1075.
(11) Knockaert, L.; Zutter, D. D. *Int. J. Electron. Commun.* **1999**, *53* (5), 254–260.
(12) Stewart, G. W. *SIAM Rev.* **1993**, *35* (4), 551–566.
(13) Deprettere, E. F. *SVD and Signal Processing: Algorithms, Applications and Architectures*; North-Holland: Amsterdam, 1988.
(14) Veen, A. J. V. D.; Deprettere, E. F.; Swindlehurst, A. L. *Proc. IEEE* **2002**, *81* (9), 1277–1308.
(15) Lathauwer, L. D.; Moor, B. D.; Vandewalle, J. *SIAM J. Matrix Anal. Appl.* **2000**, *21* (4), 1253–1278.
(16) Wold, S.; Geladi, P.; Esbensen, K.; Öhman, J. *J. Chemom.* **1987**, *1* (1), 41–56.
(17) Höskuldsson, A. *J. Chemom.* **1988**, *2* (3), 211–228.
(18) Bro, R.; Acar, E.; Kolda, T. G. *J. Chemom.* **2008**, *22* (2), 135–140.
(19) Cao, L. *Nonlinear Wavelet Compression Methods for Ion Analyses and Dynamic Modeling of Complex Systems*; Ohio University, Athens, OH, USA, 2004.
(20) Chen, P. *Applications of Chemometric Algorithms to Ion Mobility Spectrometry and Matrix Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry*; Ohio University: Athens, OH, USA, 2008.

(21) Lu, Y.; Chen, P.; Harrington, P. *Anal. Bioanal. Chem.* **2009**, *394*, 2061–2067.
(22) Chen, P.; Lu, Y.; Harrington, P. *Anal. Chem.* **2008**, *80*, 1474–1481.
(23) Sun, X.; Zimmermann, C. M.; Jackson, G. P.; Bunker, C. E.; Harrinton, P. B. *Talanta* **2011**, *83*, 1260–1268.
(24) Frank, I. E.; Kowalski, B. R. *Anal. Chim. Acta* **1984**, *162*, 241–251.
(25) Barker, M.; Rayens, W. *J. Chemom.* **2003**, *17*, 166–173.
(26) Harrington, P. d. B.; Laurent, C.; Levinson, D. F.; Levitt, P.; Markey, S. P. *Anal. Chim. Acta* **2007**, *599*, 219–231.
(27) Xu, Z.; Bunker, C. E.; Harrington, P. d. B. *Appl. Spectrosc.* **2010**, *64* (11), 1251–1258.
(28) Harrington, P. d. B. *J. Chemom.* **1991**, *5* (5), 467–486.
(29) Harrington, P. d. B.; Vieira, N. E.; Chen, P.; Espinoza, J.; Nien, J. K.; Romero, R.; Yergey, A. L. *Chemom. Intell. Lab. Syst.* **2006**, *82*, 283–293.
(30) Rearden, P.; Harrington, P. d. B.; Karnes, J. J.; Bunker, C. E. *Anal. Chem.* **2007**, *79*, 1485–1491.
(31) Harrington, P. d. B. *Trends Anal. Chem.* **2006**, *25* (11), 1112–1124.
(32) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; Jong, S. D.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics*: Part A; Elsevier: Amsterdam, 1997; Vol. 20A, p 867.
(33) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1–17.
(34) Harrington, P. d. B.; Kister, J.; Artaud, J.; Dupuy, N. *Anal. Chem.* **2009**, *81*, 7160–7169.
(35) Quinlan, J. R. *Mach. Learn.* **1986**, *1*, 81–106.
(36) Zadeh, L. A. *Inf. Control* **1965**, *8* (3), 338–253.
(37) Williams, P.; Norris, K. *Near-Infrared Technology in the Agricultural and Food Industries*; American Association of Cereal Chemists: St. Paul, MN, USA, 1987; p 330.
(38) Pereira, C. F.; Pimentel, M. F.; Galvão, R. K. H.; Honorato, F. A.; Stragevitch, L.; Martins, M. N. *Anal. Chim. Acta* **2008**, *611*, 41–47.
(39) Donoho, D.; Maleki, A.; Shahram, M. WAVELAB 850, http://www-stat.stanford.edu/~wavelab/Wavelab_850/index_wavelab850.html; accessed in 2010.