analytical chemistry

Article

pubs.acs.org/ac

Data Size Reduction Strategy for the Classification of Breath and Air Samples Using Multicapillary Column-Ion Mobility Spectrometry

Ewa Szymańska,*[,†,‡] Emma Brodrick,[§] Mark Williams,[§] Antony N. Davies,[§,∥] Henk-Jan van Manen,[∥] and Lutgarde M. C. Buydens[‡]

[†]TI-COAST, Science Park 904, 1098 XH Amsterdam, The Netherlands

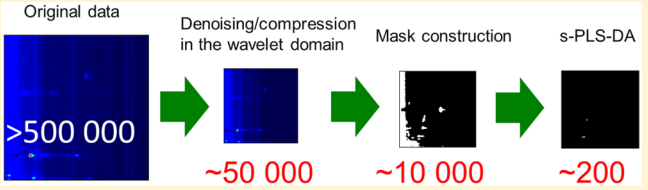
[‡]Radboud University Nijmegen, Institute for Molecules and Materials (IMM), P.O. Box 9010, 6500 GL Nijmegen, The Netherlands

[§]School of Applied Sciences, Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, CF37 1DL, United Kingdom

[∥]AkzoNobel N.V., Supply Chain, Research and Development, Strategic Research Group - Measurement & Analytical Science, P.O. Box 10, 7400 AA, Deventer, The Netherlands

Supporting Information

ABSTRACT: Ion mobility spectrometry combined with multicapillary column separation (MCC-IMS) is a well-known technology for detecting volatile organic compounds (VOCs) in gaseous samples. Due to their large data size, processing of MCC-IMS spectra is still the main bottleneck of data analysis, and there is an increasing need for data analysis strategies in which the size of MCC-IMS data is reduced to enable further analysis. In our study, the first untargeted chemometric strategy is developed and employed in the analysis of MCC-IMS spectra from 264 breath and ambient air samples. This strategy does not comprise identification of compounds as a primary step but includes several preprocessing steps and a discriminant analysis. Data size is significantly reduced in three steps. Wavelet transform, mask construction, and sparse-partial least squares-discriminant analysis (s-PLS-DA) allow data size reduction with down to 50 variables relevant to the goal of analysis. The influence and compatibility of the data reduction tools are studied by applying different settings of the developed strategy. Loss of information after preprocessing is evaluated, e.g., by comparing the performance of classification models for different classes of samples. Finally, the interpretability of the classification models is evaluated, and regions of spectra that are related to the identification of potential analytical biomarkers are successfully determined. This work will greatly enable the standardization of analytical procedures across different instrumentation types promoting the adoption of MCC-IMS technology in a wide range of diverse application fields.



Ion mobility spectrometry (IMS) is increasingly in demand for medical applications and process and environmental control, as well as food quality and safety.[1,2] In all these applications, the breath or air samples are extremely complex mixtures and include numerous volatile organic compounds (VOCs). Through a combination of IMS with a multicapillary column (MCC-IMS) separation, a fast, robust, noninvasive, and easy-to-use system for qualitative and quantitative analyses of VOCs at low ppb and ppt levels is now available.[3,4]

MCC-IMS spectra present chemometric challenges due to their large size and megavariate nature (more than a million variables per sample) as well as redundancy of information (several variables associated with a single VOC).[1,5] Computational challenges such as "out of memory" problems and extensively long computation times often occur when data is exported from the analytical equipment and analyzed on the standard PC. Moreover, overfitting and false positive associations commonly take place when applying standard multivariate data analysis tools, which cannot handle properly megavariate data sets with collinearities and noisy data. Therefore, the main idea presented in this work is to develop an efficient strategy for size reduction of MCC-IMS data sets taking their characteristics into account.

Currently, MCC-IMS data size reduction is addressed with approaches including clustering of characteristic peak structures[6] and peak picking with specialized software and databases such as VisualNow.[1,7] These targeted approaches select peak regions in MCC-IMS spectra and quantify known, identified compounds using their maximal peak height. An untargeted approach, i.e., when compound identification is not a primary step of data analysis, allowing a reduction to 25% of the MCC-IMS spectra data points, was recently introduced in the work of Bader et al.[8] In this approach, wavelet transformation with *Daubechies 8* wavelet is successfully applied to both dimensions of MCC-IMS spectra. Due to only one human breath sample included in this work, no benefits of size reduction to further

Received: June 4, 2014
Accepted: December 18, 2014
Published: December 18, 2014

DOI: 10.1021/ac503857y
Anal. Chem. 2015, 87, 869−875

analysis were demonstrated. The aim of our paper is to develop an untargeted data analysis strategy in which the size of the MCC-IMS spectra is reduced without significant information loss so further analysis is enabled, e.g., the classification of different samples and the selection of spectral regions important for their classification.

The proposed strategy includes several preprocessing steps and a discriminant analysis as shown in Figure 1. Data size

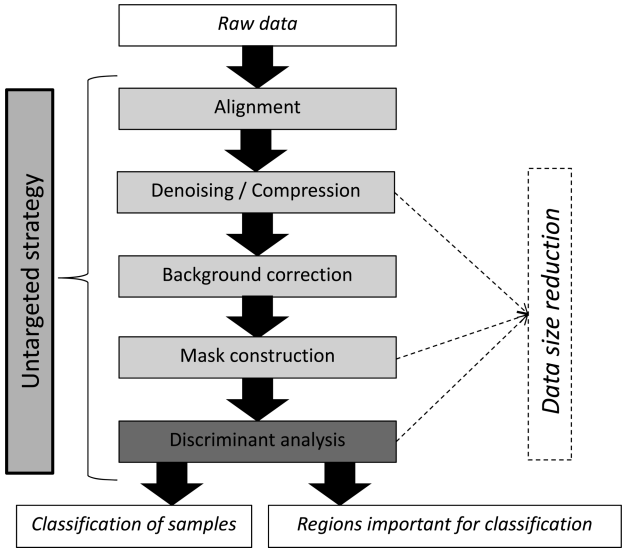


Figure 1. Workflow of the developed data processing strategy. Steps contributing to the data size reduction are marked.

reduction is achieved in three steps: (a) by compression with wavelet transform, (b) by mask construction, and (c) by variable selection during discriminant analysis with sparse-partial least squares-discriminant analysis (s-PLS-DA). The theoretical background of these essential size reduction steps is briefly summarized in the Theory section. Next, the developed strategy is successfully implemented and validated in the analysis of a large multiclass MCC-IMS data set with 264 sample spectra of human breath and ambient air. The performance of the strategy is described in the Results and Discussion section and includes the effects of different preprocessing steps on data quality and information loss as well as the optimization of settings for the best classification of results. Advantages of the proposed strategy in reference to currently used strategies are apparent and presented in the Results and Discussion section.

THEORY

This section gives a brief theoretical background of the methods used in the data size reduction steps.

Wavelet Transform. Wavelet transform is established as a data processing method in analytical chemistry.[9,10] The main fields of application are related to denoising, compression, variable reduction, and signal suppression. Wavelet transform is also used to denoise and compress ion mobility data,[11,12] MCC-IMS data,[8] and data obtained with other hyphenated techniques such as GC-DMS, GC/MS, LC-MS, and EC-MS.[13−16] Either the two-dimensional wavelet basis function or the one-dimensional wavelets, applied along the two axes of the decomposed signal,[8,13] can be used.

Wavelet transform is a mathematical transformation for hierarchically decomposing functions, i.e., signals.[9,10] In a single level decomposition, a signal goes through two complementary low-pass and high-pass filters. The output of the low-pass filter gives the signal approximation (A) coefficients, while the high-pass filter gives the signal detail (D) coefficients. Low- and high-pass filters are defined by basis wavelets which are specified by wavelet filter coefficients and are selected on the basis of original signal characteristics.

During data denoising, thresholding of the wavelet detail coefficients (D) is applied after wavelet transform. Therefore, noisy components below the threshold are removed, and the denoised signal is reconstructed. Signal compression can be achieved by discarding the detail coefficients and retaining the approximation coefficients after wavelet transform. Thus, 50% of data compression is obtained. Additional compression can be obtained by further applying wavelet transform to the kept approximation coefficients. At the nth level of compression, $1/2^n$ of the original data size will be saved. The optimum data compression level depends on the data and the wavelet filter type.

Mask Construction. Mask construction is well-known in image analysis to define a region of interest (ROI),[17,18] which can also be applied for feature selection in analytical chemistry, for example, in two-dimensional gel electrophoretic data.[13] During mask construction, there are predefined criteria; e.g., thresholds are applied, and only features (variables) meeting those criteria are selected and included in further analysis. The criteria are based on the data characteristics and the goal of data analysis, for example, the threshold based on the mean signal.[13] Mask construction not only significantly reduces data dimensionality but also has other advantages: selected features often exclude background features and include numerous peaks, and each peak is described by many correlated variables that can help in the final interpretation.[13]

Sparse-PLS-DA. Sparse-partial least-squares-discriminant analysis was introduced by Lê Cao et al.[19] as a natural extension of s-PLS proposed by the same authors.[20] A sparse version of PLS-DA aims at combining variable (feature) selection and classification in a one-step procedure. Variable selection is performed per latent component by applying penalties l_1 (Lasso penalties) on the loading vectors of the X data set and is user-controlled by the penalization parameter λ. For practical reasons, the number of variables selected in the Le Cao et al.[19] algorithm are chosen, opposed to the penalization parameter λ. Two parameters have to be tuned before s-PLS-DA is carried out: the number of latent variables (the model complexity, number of dimensions H) and the number of selected variables for each latent variable. For multiclass problems with K classes, $H = K - 1$ is usually advised for generating the most stable models.[19] The number of selected variables is related to the complexity of data and is optimized on the basis of model performance and stability.

EXPERIMENTAL SECTION

Samples. 110 human breath samples and 154 ambient room air samples are analyzed in this study. The breath data consist of human breath samples with and without ingestion of sweets, 57 and 53 samples, respectively. Room air samples are collected at three sites: 21 samples at site 1, 41 samples at site 2, and 92 samples at site 3.

Instrumental Analysis. The ion mobility spectrometer used in this study is coupled with a multicapillary column

DOI: 10.1021/ac503857y
Anal. Chem. 2015, 87, 869−875

Table 1. Denoising/Compression in the Wavelet Domain[a]

	ion mobility dimension		retention time dimension					
setting	denoising	compression	denoising	compression	number of variables	efficiency (%)	RMSE	RMSEM
s0	−	−	−	−	521 280	0	0	0
s1	+	−	+	−	521 280	0	1.93	0.27
s2	+	2×	+	−	260 640	50	1.76	0.19
s3	+	4×	+	−	130 560	75	1.77	0.20
s4	+	8×	+	−	65 280	87.5	2.25	0.90
s5	+	16×	+	−	32 640	93.75	10.27	7.68
s6	+	−	+	2×	260 640	50	1.93	0.27
s7	+	2×	+	2×	130 320	75	2.07	0.29
s8	+	4×	+	2×	65 280	87.5	2.08	0.30
s9	+	8×	+	2×	32 640	93.75	2.50	0.93
s10	+	16×	+	2×	16 320	96.87	10.32	7.69
s11	+	−	+	4×	130 320	75	1.94	0.28
s12	+	2×	+	4×	65 160	87.5	2.52	0.72
s13	+	4×	+	4×	32 640	93.75	2.52	0.72
s14	+	8×	+	4×	16 320	96.87	2.88	1.14
s15	+	16×	+	4×	8160	98.44	10.42	7.71
s16	+	4×	+	2×	65 280	87.5	2.08	0.30

[a]Settings s0−s16 include different denoising and compression parameters where + indicates application of denoising and − indicates no denoising/compression. 2×, 4×, 8×, and 16× indicates the level of applied compression. Settings s8 and s16 have the same parameters in denoising and compression. The number of variables retained after denoising/compression and region selection is listed against the efficiency of compression (%), root mean square errors (RMSE), and root mean square error of the mean spectrum (RMSEM) of the reconstructed data sets vs original data set (setting s0).

(MCC) for preseparation. The instrument is a BreathDiscovery using VOCan v 1.4 operating software (B&S Analytik, Dortmund, Germany) and has a SpiroScout inlet attachment (Ganhorn Medizin Electronic, Niederlauer, Germany). The main parameters of the instrument are detailed elsewhere.[21−30] The IMS uses a 550 MBq ^{63}Ni ß-radiation source for ionization and synthetic air as a carrier gas. For preseparation of the sample, the IMS utilizes a nonpolar multicapillary column (MCC, type OV-5, Multichrom Ltd., Novosibirsk, Russia) with 1000 parallel capillaries, each with an inner diameter of 40 μm and a film thickness of 200 nm. The total diameter of the separation column is 3 mm.

Original MCC-IMS Data Set. The full data set is composed of 264 samples spectra including ca. 1.5 million points per sample: 2499 ion mobility vs 600 retention time variables. The size of the data set is 2.5 GB.

An example heatmap of a breath sample is presented in Supplementary Figure 1, Supporting Information. A single peak spans ca. 150 points: 15 in ion mobility and 10 in retention time dimension. For each sample point, the ion mobility (IM) scale is also expressed as reduced inverse ion mobility ($1/K_0$) values in a range of 0−1.5 Vs/cm^2 and retention time points correspond to retention time (RT) in seconds in a range of 0−300 s. The inverse reduced ion mobility is calculated using the ion mobility coefficient and is adjusted for temperature and pressure as determined by the IMS device during operation.

Software. Matlab 2013a with the Wavelet design and Analysis toolbox (The Mathworks Inc., Natick, Massachusetts, USA) is used in all steps of the data processing excluding sparse-PLS-DA for which R software version 3.0.1 with the mixOmics package[31] is employed (available at http://cran.r-project.org/web/packages/mixOmics/index.html).

Steps of Developed Chemometric Strategy. *Step 1: Alignment.* The ion mobility scale is adjusted so all samples begin at the same value of inverse reduced ion mobility ($1/K_0$). No alignment in retention time dimension is employed.

Settings. The maximum $1/K_0$ value (0.43 Vs/cm^2), used as a shifting target, is found by sorting in descending order the 700th value of the $1/K_0$ scale from all samples. The 700th point was arbitrarily selected as no peaks were observed in the spectra before this point. In a single sample, the 700th point in the ion mobility dimension (out of 2499) corresponds to ca. $1/K_0$ = 0.4 Vs/cm^2 (see Supplementary Figure 1, Supporting Information). The sample spectra are then shifted accordingly in the $1/K_0$ dimension. There are no areas of interest after 240 s in the retention time dimension; therefore, the spectra used in further analysis has the $1/K_0$ range of 0.43−1.2 Vs/cm^2 and retention time dimension range of 0−240 s (1300 × 480 points per sample spectra) (Supplementary Figure 1, region A, Supporting Information).

Step 2: Denoising and Compression in the Wavelet Domain. In this step, denoising and compression of the data set is obtained by wavelet transformation, first in ion mobility (IM) dimension and then in retention time (RT) dimension.

Settings. Fifteen different settings of denoising and compression are applied as presented in Table 1. Denoising is carried out in both dimensions using discrete wavelet transform with interval dependent thresholding[32] (*cmddenoise* function in Matlab). Set thresholds are based on Donoho thresholds,[33] and for denoising, the maximum number of intervals are set to 6 and 3 in the IM and RT dimensions, respectively, with levels set to 5 and 2 in the IM and RT dimensions, respectively. *Daubechies 8* wavelet and hard thresholding is used in denoising for both dimensions. After denoising in the ion mobility dimension, data is denoised in the retention time dimension (setting s1) or compressed in ion mobility dimension. In the latter, decomposition of data by multilevel wavelet decomposition (*wavedec* function in Matlab) is applied with 5 levels and the *Daubechies 8* wavelet. Compressed data are obtained by wavelet reconstruction using the settings of the wavelet decomposition (*waverec* function in Matlab) and a selected level of deconstruction

DOI: 10.1021/ac503857y
Anal. Chem. 2015, 87, 869−875

(from 1 to 4). Obtained data are either further denoised in the retention time dimension (settings s2−s5) or denoised and compressed in the retention time dimension (settings s6−s15). Analogously to compression in the ion mobility dimension, decomposition of data by multilevel wavelet decomposition with 3 levels and wavelet deconstruction with 2 levels are used.

Figures of Merit. The efficiency and efficacy of compression is evaluated on the whole data set. The efficiency of compression is assessed by the number of points/variables retained after compression, relative to the original data size, subtracted from unity and expressed in %. The efficacy of compression is a measure of similarity of the reconstructed data compared to the original data. This is estimated by reconstruction errors, i.e., the root-mean-square error (RMSE) and the root-mean-square error of the mean spectrum (RMSEM) as recommended by Urbas and Harrington.[12] RMSE and RMSEM errors of compressed data sets are calculated on the reconstructed data (inverse wavelet transform, *idwt* function in Matlab) in reference to the data before compression (setting s0).

Step 3: Background Elimination. A minimum spectrum is subtracted from every spectrum of each sample to compensate for the reactant ion peak (RIP) and baseline drift caused by RIP tailing. The minimum spectrum is obtained by ascendingly sorting all values of spectra obtained at the same inversed reduced ion mobility (sorting in retention time dimension).[34]

Step 4: Mask Construction with Region Selection. First, a region of spectra excluding RIP is selected (see Supplementary Figure 1, region B, Supporting Information). Second, variables of spectra lower than a certain threshold are eliminated, i.e., by mask construction. Masks are constructed for sample classes that are included in the discriminant analysis. The mask construction consists of two steps: (1) Construction of a mask per sample class. (2) Sum of the masks of different classes included in the discriminant analysis.

Settings. The selected region of spectra consists of inversed reduced ion mobility range of 0.56−1.16 Vs/cm^2 and retention time range of 0−240 s. Three different masks are constructed for three classification problems considered in the discriminant analysis (Table 2). In each case, the mask threshold is based on the limit of detection (LOD), i.e., mean + 3 × standard deviation of a blank region of the spectra from all samples included in the discriminant analysis. The blank region is defined as a region where no peaks are observed, i.e., a region between 0.94 and 1.16 Vs/cm^2 and between 200 and 240 s. Masks per sample class are constructed by eliminating variables that are lower than the set threshold in more than 50% of the samples. Next, a sum mask is created by including all variables included in at least one of the masks per sample class.

Table 2. Classification Problems Considered in the Discriminant Analysis

classification problem	description	number of classes	total number of samples	samples per class
1st	breath vs air	2	264	110:154
2nd	air 1 vs air 2 vs air 3	3	154	21:41:92
3rd	breath with sweet vs breath without sweet vs air 1 vs air 2 vs air 3	5	264	57:53:21:41:92

Step 5: Discriminant Analysis. Partial least squares-discriminant analysis and its sparse version (s-PLS-DA) are used. The data is unfolded before discriminant analysis. The response matrix Y of size ($n \times K$), where n is the number of samples and K is the number of classes, is created for each of the three classification problems presented in Table 2. In the response matrix dummy, variables 0 and 1 indicate the class membership of each sample.

Settings. The number of latent variables kept in the (s-)PLS-DA model are set to 1, 2, and 3 latent variables (dimensions) for the 1st, 2nd, and 3rd problem, respectively, following the recommendation by Lê Cao et al.[19] Six different sparsity levels are tested and contain all (PLS-DA) and 800, 400, 200, 100, and 50 variables (s-PLS-DA) with nonzero loadings for each latent variable.

All s-PLS-DA and PLS-DA models are cross-validated (7-fold cross-validation repeated 20 times). For class prediction of the test samples, the maximum distance approach is used.[19] This approach is based on the predicted matrix, which can be seen as a probability matrix to assign each test data to a class. The class with the largest class value is the predicted class.

Figures of Merit. The performance of each model is assessed by calculating the average percentage of misclassified samples in all 20 repetitions.[35] For each classification problem, (s-)PLS-DA models with different sparsity levels (6 levels) and denoising/compression settings (15 settings presented in Table 1) are compared and ranked by performance. The optimal sparsity and the optimal wavelet transform settings are subsequently selected on the basis of a sum of ranks obtained for the three classification problems. Performances of s-PLS-DA models on data without denoising/compression (setting s0) and data without mask construction (setting s16) are also reported.

Variable importance is evaluated by a frequency of variable selection in different cross-validations and repetitions. If a variable is selected in at least 80% of cases of all cross-validations and repetitions (114 times out of 140: 7 cross-validations and 20 repetitions), it is assumed to be important and therefore included in further interpretations. In the case of models with more than one latent variable, variable importance is assessed per latent variable and a summary of selected variables is presented for further interpretation.

RESULTS AND DISCUSSION

Strategy at Work. In this study, a new chemometric strategy for the analysis of MCC-IMS data sets is developed and validated. This strategy consists of five steps as shown in Figure 1 and is applied to a MCC-IMS data set to classify different breath and air samples as shown in Table 2 and described in the Experimental Section.

In the first step of the developed strategy (alignment), spectra of each sample are aligned by their reduced inverse ion mobility scale to a reference spectrum. This step reduces the shift in ion mobility dimension significantly (visual inspection). An alternative can be the normalization of reduced inverse ion mobility by the RIP.[36]

In the second step (wavelet transform), spectra are denoised and compressed. Fifteen different settings of wavelet transform are tested (Table 1) and compared with data sets without denoising and compression (setting s0). The number of points kept after wavelet transform, compression efficiency, and compression efficacy (RMSE and RMSEM) are reported in Table 1.

DOI: 10.1021/ac503857y
Anal. Chem. 2015, 87, 869−875

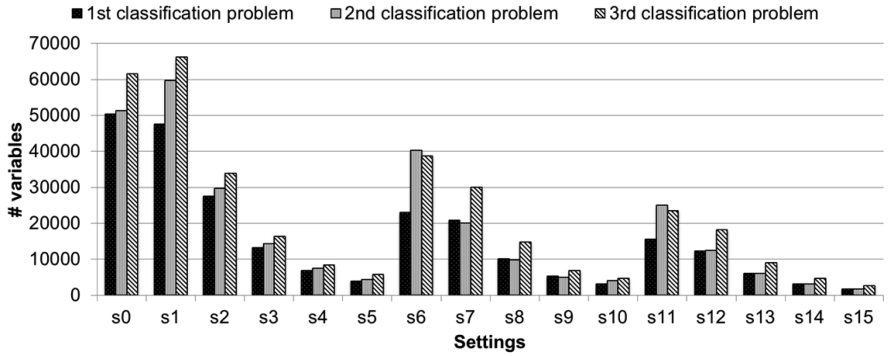


Figure 2. Numbers of variables kept for further analysis after mask construction (4th step of the developed data processing strategy) for different settings of wavelet transform (see Table 1) and classification problems (see Table 2).

The compression efficiency ranges from 50% for setting s2 (s1 consists of only denoising) to 98.44% for setting s15, which is a compression of 64 times (16 times in the ion mobility dimension and 4 times in the retention time dimension). Both RMSE and RMSEM values show an increase in reconstruction error with an increased compression level. Values of RMSE and RMSEM for settings with 2× and 4× compression in the ion mobility dimension are similar (s2−s3, s7−s8, s12−s13) in contrast to values for settings with 8× and 16× compression (s4−s5, s9−s10, s14−s15). This may indicate a great information loss during 16× compression in the ion mobility dimension. No clear indication of information loss is observed when RMSE and RMSEM are compared for the retention time dimension.

Differences in values of RMSE and RMSEM are related to the noise level of the data. RMSE is largely affected by noise in contrast to RMSEM. Therefore, RMSE is more useful for evaluation if the noise is removed, e.g., after denoising/compression. RMSEM is used to evaluate any misleading artifacts introduced to the data.[12]

In the third step (background correction), RIP peak tailing is reduced and data size remains unchanged. A minimum spectrum is subtracted from every spectrum of each sample MCC-IMS spectra to compensate for RIP and baseline drift caused by RIP tailing. However, some RIP tailing is still present after this step (Supplementary Figure 2, Supporting Information) and is included in regions selected in the mask construction.

In the fourth step (region selection and mask construction), first a region of spectra without RIP is selected yielding a number of points/variables ranging from 8160 (setting s15) to 521 280 (setting s0 and s1) (as reported in Table 1). Data size is further reduced by the mask construction. In this step, points/variables of spectra are selected separately for each of the three classification problems considered in the discriminant analysis (Table 2). The number of points and efficiency of mask construction for each classification problem is presented in Figure 2 and Supplementary Table 1, Supporting Information, respectively. An example of a mask for the first classification problem is presented in Figure 3.

On average, the efficiency of the mask construction is ca. 80% but it is highly dependent on the classification problem and data set obtained after one of 15 different settings of the wavelet transform. Efficiency of the constructed mask decreases with the increase in number of classes (1st > 2nd > 3rd classification problem) and the level of compression. The number of retained variables ranges from 66 222 for setting s1

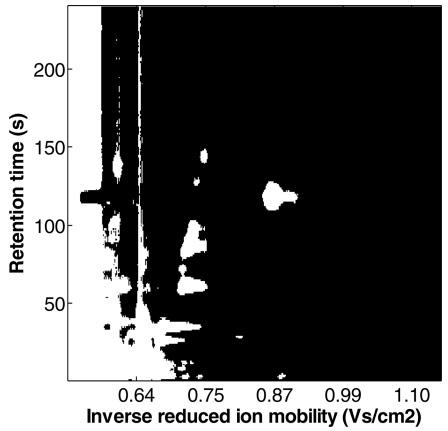


Figure 3. Example mask constructed for the 1st classification problem and setting s3 in denoising/compression in wavelet domain.

for the third classification problem and 1802 for setting s15 for the first classification problem.

In the final step, discriminant analysis takes place. The results of discriminant analysis including the performances of PLS-DA and sparse-PLS-DA models for the different settings of wavelet transform and different classification problems are shown in Figure 4 and Supplementary Figures 3 and 4, Supporting Information.

The considered classification problems differ by the number of classes, the number of samples in each class, and the difficulty of discrimination (see Table 2). This is reflected in the performances of the PLS-DA and s-PLS-DA models. PLS-DA models including all variables have a mean misclassification rate of ca. 25%, 16%, and 20% for 1st, 2nd, and 3rd classification problems, respectively. The performance of the s-PLS-DA models including only 50−800 variables is consistently better than any of the PLS-DA models, and in the case of the second classification problem, improvement is 4-fold.

There is no difference between the denoised and undenoised data set (settings s0 and s1) in regard to model performance. A decrease in the model performance can usually be observed when no mask construction step is applied during the data processing (setting s16 and s8). When the performances of models are compared between the 15 settings of denoising/compression by wavelet transform, two main trends can be observed. The first trend is related to compression in the retention time dimension. When three sets of five settings, s1−s5, s6−s10 and s11−s15, with the same compression in the retention time dimension are compared, the performance

DOI: 10.1021/ac503857y
Anal. Chem. 2015, 87, 869−875

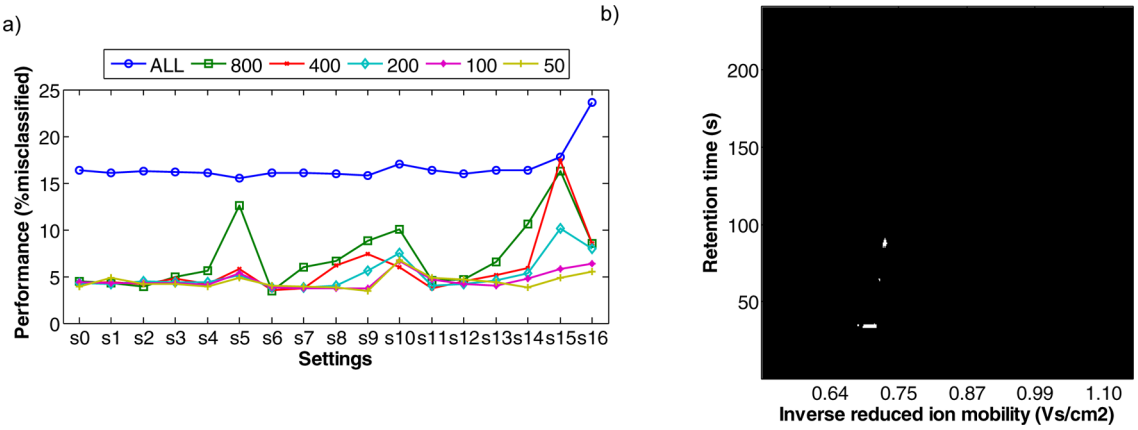


Figure 4. Discriminant analysis results for the 2nd classification problem. (a) Performance of (s)-PLS-DA models expressed as % of misclassified samples for different settings of denoising/compression (see Table 1); 6 sparsity levels: 50, 100, 200, 400, 800, and all variables. (b) Regions of spectra selected as the most relevant by the optimal (s)-PLS-DA model.

decreases from the first to the last set. The second trend is a consistent behavior within the same set of five settings: performance is usually the best for the second and the third setting (s2 and s3, s7 and s8, s12 and s13). The ranking of models is given in Supplementary Table 2, Supporting Information. The model performance of data processed with settings including 2×, 4×, and 8× total compression in both dimensions (settings s2, s3, s4, s6, s7, s8, s11, s12) is in the same range as the model performance of uncompressed data sets (settings s0 and s1). The best performance was obtained with settings s2, s6, and s2 for the 1st, 2nd, and 3rd classification problem, respectively.

Finally, when the performances of models with different levels of sparsity are compared, models with 100 variables are consequently ranked as the best or the second best (Supplementary Table 3, Supporting Information). The variables important for optimal s-PLS-DA of the second classification problem (based on wavelet transform settings s6 with 100 variables) are studied and presented in Figure 4b (1st and 3rd classification problem are shown in Supplementary Figure 3 and 4, Supporting Information). Selected variables consist of 45, 63, and 740 variables for the 1st, 2nd, and 3rd classification problems, respectively. Grouping of selected variables in regions resembling peaks is observed, but their careful interpretation is still required.

Added Value of the Developed Strategy. The developed strategy is untargeted and includes three steps in which data size is significantly reduced. No expert or software-based peak detection or peak fitting is required in contrast to current processing strategies for MCC-IMS data.[7,37] The reduced size of the data sets leads to valuable results, i.e., the discrimination of different classes of samples and the identification of spectra regions—potential biomarker features by sparse-PLS-DA. The strategy has been tested on real, large MCC-IMS data sets with multiple classes of air and breath samples.

The most important steps of the developed strategy include the compression of data by wavelet transform, mask construction, and sparse-PLS-DA. Data size is reduced up to 64 times during wavelet transform and again ca. 5 times by mask construction with up to 50 variables (less than 0.01% of original size). After wavelet transform, redundant information is removed from spectra including noise; thus, the size of data is efficiently reduced so further analysis is computationally feasible. Moreover, this study has shown that wavelet transform compression of data size up to 8 times yields similar discriminant analysis results to these obtained on uncompressed data so no significant information loss can be reported.

Mask construction and sparse-PLS-DA are implemented in the analysis of MCC-IMS data for the first time. The mask construction is applied as a supervised step where the threshold, which is used to select variables/points of spectra, is based on the presence of a considered variable above a background level in a specified class of samples. After mask construction irrelevant information, i.e., variables below background level or with inconsistent behavior within a class of samples are removed. Therefore, the reduction of data size during mask construction is dependent on the classification problem and makes mask construction a very attractive tool for the visualization and interpretation of results. Sparse-PLS-DA for multiple class problems, with adequate cross-validation procedures, allows for further reduction in data size, thus selecting variables directly related to each classification problem, i.e., identified biomarker features. These variables may be directly used as a valid set of variables on new samples within each class of air and breath samples.

The optimal settings of the developed strategy are selected on the basis of optimal performance of the validated classification models, e.g., by a fit-for-use approach.[38] The performance of s-PLS-DA models (based on different settings of the preprocessing strategy) is validated with 7-fold cross-validation and 20 repetitions. It is stressed that such an extensive validation approach of the processing strategies, for MCC-IMS data sets, is not performed or available elsewhere. Recent work on preprocessing strategies for MCC-IMS data sets only includes the visual inspection of preprocessed data and the evaluation of differences in intensity of the referenced features of the studied spectra.[5,8]

Outlook. The developed strategy can be used to analyze other large two-dimensional data sets. These may include gel electrophoretic analysis, gas chromatography-gas chromatography (GC-GC), liquid chromatography-liquid chromatography (LC-LC), or image analysis by other analytical techniques. The developed strategy offers an efficient combination of three data size reduction steps and is suited to multiclass problems, and the settings of the methods can be easily adjusted. Further research is required to optimize the settings of this strategy to a

DOI: 10.1021/ac503857y
Anal. Chem. 2015, 87, 869−875

specific data set and compare the developed strategy with other currently used strategies.

CONCLUSIONS

The developed strategy reported in this Article is the first untargeted approach for the analysis of MCC-IMS data sets. This is a novel alternative and complements targeted approaches including expert or software-based peak picking. It consists of three steps in which data size is significantly reduced. Wavelet transform, mask construction, and sparse-PLS-DA allow data size reduction with up to 50 variables relevant to the goal of analysis. The strategy is successfully implemented and validated on a large MCC-IMS data set on multiple classes of breath and air samples. The reduced size of the data set allows the discrimination of classes of samples and the identification of spectral regions—potential biomarker features.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: Szymanska.ewcia@gmail.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge all academic and industrial partners of the Analysis of Large data sets By Enhanced Robust Techniques (ALBERT) project for valuable input and discussion. This research received funding from The Netherlands Organisation for Scientific Research (NWO) in the framework of the Technology Area COAST.

REFERENCES

(1) Hauschild, A.; Baumbach, J.; Baumbach, J. *Genet Mol. Res.* **2012**, *11*, 2733−2744.
(2) Armenta, S.; Alcala, M.; Blanco, M. *Anal. Chim. Acta* **2011**, *703*, 114−123.
(3) Ruzsanyi, V.; Baumbach, J. I.; Sielemann, S.; Litterst, P.; Westhoff, M.; Freitag, L. *J. Chromatogr., A* **2005**, *1084*, 145−151.
(4) Jünger, M.; Vautz, W.; Kuhns, M.; Hofmann, L.; Ulbricht, S.; Baumbach, J. I.; Quintel, M.; Perl, T. *Appl. Microbiol. Biotechnol.* **2012**, *93*, 2603−2614.
(5) Hauschild, A.-C.; Schneider, T.; Pauling, J.; Rupp, K.; Jang, M.; Baumbach, J.; Baumbach, J. *Metabolites* **2012**, *2*, 733−755.
(6) Bader, S.; Urfer, W.; Baumbach, J. I. *J. Chemom.* **2006**, *20*, 128−135.
(7) Bödeker, B.; Vautz, W.; Baumbach, J. I. *Int. J. Ion Mobility Spectrom.* **2008**, *11*, 89−93.
(8) Bader, S.; Urfer, W.; Baumbach, J. *Int. J. Ion Mobility Spectrom.* **2008**, *11*, 43−49.
(9) Walczak, B.; Massart, D. *TrAC, Trends Anal. Chem.* **1997**, *16*, 451−463.
(10) Ehrentreich, F. *Anal. Bioanal. Chem.* **2002**, *372*, 115−121.
(11) Chen, G.; Harrington, P. B. *Anal. Chim. Acta* **2003**, *490*, 59−69.
(12) Urbas, A. A.; Harrington, P. B. *Anal. Chim. Acta* **2001**, *446*, 393−412.
(13) Daszykowski, M.; Stanimirova, I.; Bodzon-Kulakowska, A.; Silberring, J.; Lubec, G.; Walczak, B. *J. Chromatogr., A* **2007**, *1158*, 306−317.
(14) Smolinska, A.; Hauschild, A. C.; Fijten, R. R. R.; Dallinga, J. W.; Baumbach, J.; van Schooten, F. J. *J. Breath Res.* **2014**, *8*, 027105.
(15) Chen, P.; Lu, Y.; Harrington, P. B. *Anal. Chem.* **2008**, *80*, 7218−7225.
(16) Cao, L.; Harrington, P. d. B.; Liu, C. *Anal. Chem.* **2004**, *76*, 2859−2868.
(17) Kontos, D.; Megalooikonomou, V. *Proc. SPIE* **2004**, *5370*, 1324−1331.
(18) Privitera, C. M.; Stark, L. W. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 970−982.
(19) Lê Cao, K.-A.; Boitard, S.; Besse, P. *BMC Bioinf.* **2011**, *12*, 253.
(20) Lê Cao, K.-A.; Rossouw, D.; Robert-Granié, C.; Besse, P. *Stat. Appl. Genet. Mol. Biol.* **2008**, *7*; DOI: 10.2202/1544-6115.1390.
(21) Bunkowski, A.; Maddula, S.; Davies, A. N.; Westhoff, M.; Litterst, P.; Boedeker, B.; Baumbach, J. I. *Int. J. Ion Mobility Spectrom.* **2010**, *13*, 141−148.
(22) Juenger, M.; Boedeker, B.; Baumbach, J. I. *Anal. Bioanal. Chem.* **2010**, *396*, 471−482.
(23) Westhoff, M.; Litterst, P.; Freitag, L.; Urfer, W.; Bader, S.; Baumbach, J. I. *Thorax* **2009**, *64*, 744−748.
(24) Bunkowski, A.; Boedeker, B.; Bader, S.; Westhoff, M.; Litterst, P.; Baumbach, J. I. *J. Breath Res.* **2009**, *3*, 046001/046001−046001/046010.
(25) Westhoff, M.; Litterst, P.; Freitag, L.; Baumbach, J. I. *J. Physiol. Pharmacol.* **2007**, *58*, 739−751.
(26) Baumbach, J.; Westhoff, M. *Spectrosc. Eur.* **2006**, *18*, 22−27.
(27) Baumbach, J. I. *Anal. Bioanal. Chem.* **2006**, *384*, 1059−1070.
(28) Westhoff, M.; Litterst, P.; Maddula, S.; Bödeker, B.; Rahmann, S.; Davies, A.; Baumbach, J. *Int. J. Ion Mobility Spectrom.* **2010**, *13*, 131−139.
(29) Boedeker, B.; Davies, A. N.; Maddula, S.; Baumbach, J. I. *Int. J. Ion Mobility Spectrom.* **2010**, *13*, 177−184.
(30) Maddula, S.; Blank, L.; Schmid, A.; Baumbach, J. *Anal. Bioanal. Chem.* **2009**, *394*, 791−800.
(31) Dejean, S.; Gonzalez, I.; Kim-Anh L. C.; http://cran.r-project.org/web/packages/mixOmics/index.html (Accessed Nov 13, 2013).
(32) Praveen, A.; Vijayarekha, K.; Venkatraman, B. *Int. J. Comput. Appl.* **2013**, *72*, 1−5.
(33) Donoho, D. L. *IEEE Trans. Inf. Theory* **1995**, *41*, 613−627.
(34) Bunkowski, A. *MCC-IMS data analysis using automated spectra processing and explorative visualization methods*, Ph.D. Thesis, Bielefeld University, Bielefeld, Germany, 2011.
(35) Szymańska, E.; Saccenti, E.; Smilde, A. K.; Westerhuis, J. A. *Metabolomics* **2012**, *8*, 3−16.
(36) Vautz, W.; Bödeker, B.; Baumbach, J.; Bader, S.; Westhoff, M.; Perl, T. *Int. J. Ion Mobility Spectrom.* **2009**, *12*, 47−57.
(37) Bödeker, B.; Vautz, W.; Baumbach, J. I. *Int. J. Ion Mobility Spectrom.* **2008**, *11*, 77−81.
(38) Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J. J.; Downey, G.; Blanchet, L.; Buydens, L. M. C. *Trends Anal. Chem.* **2013**, *50*, 96−106.

DOI: 10.1021/ac503857y
Anal. Chem. 2015, 87, 869−875