Towards batch correction for GC-IMS data

Luis Fernandez*†, Arnau Blanco†, Celia Mallafré-Muro*†, Santiago Marco*†
Email: lfernandez@ub.edu, ablancob@ibecbarcelona.eu, cmallafre@ub.edu, santiago.marco@ub.edu
*Department of Electronic and Biomedical Engineering, University of Barcelona, Barcelona, Spain.
†The Institute for Bioengineering of Catalonia, Barcelona, Spain.

***Abstract*—Gas Chromatography Ion Mobility Spectrometry (GC-IMS) is a fast, non-expensive analytical technique that allows obtaining relevant chemical information from vapor mixtures. However, the technique presents some difficulties that should be solved to ensure reliable and reproducible results, namely: 1) data exhibits simultaneously high dimensionality and sparsity on their chemical information content, 2) data samples must usually be corrected even within a batch because of baseline and misalignment problems, 3) additional data corrections must be performed to prevent from chemical fingerprinting variations among batches. In this work, we have acquired data from two different batches (A and B) of ketone mixtures (2-Butanone, 2-Pentanone, 2-Hexanone, and 2-Heptanone). The analytical method for batch A and B was the same, except for the value of carrier gas flow parameter, which was approximately doubled for batch B. We have addressed problems 1) and 2) independently for each batch, obtaining as a result two peak tables. 3). Common peaks present in batches A and B were found after scaling the retention time axis of batch B and perform k-medoids clustering. Using this information, test data from batch B has been corrected through a linear transformation.**

***Keywords*—** *GC-IMS, batch effect, batch correction*

I. Introduction

Gas Chromatography – Ion Mobility Spectrometry (GC-IMS) is an analytical technique widely used in food analysis to distinguish different varieties and/or product qualities [1][2][3]. The technique combines the ability of gas chromatographs to separate mixtures of volatiles with the capacity of ion mobility spectrometers to generate chemical fingerprints characteristic of the different volatile compounds in the mixture.

The main limitation of the technique lies in managing the high amount of complex data that the instrument produces, fact that hinders data processing automation. GC-IMS readings are arranged in matrices, in which, for each retention and drift time coordinates, the instrument provides an intensity value. Interestingly, although GC-IMS data are highly dimensional (thousands of features per sample), their chemical information content is sparsely distributed in the form of peaks. Consequently, a feature extraction process to separate relevant from spurious information is practically required in GC-IMS data. In addition to this, slight modifications in environmental conditions, such as temperature and atmospheric pressure, and/or instrumental tolerances in GC-IMS method parameters' cause baseline and misalignment problems, both in retention and drift time axes. Therefore, intrabatch corrections are needed. Finally, systematic technical differences in GC-IMS analysis between batches can produce variations in data unrelated to any chemical cause. The latter problem is known as a batch effect, and it is a major issue in -omics sciences since data becomes separable per batch, but not because of its biological information content [4].

In this work, we have studied how to minimize batch effect problems in GC-IMS data. To do it, we have prepared a mixture of four ketones to be analysed using GC-IMS technique. These data were acquired in two measuring campaigns (batches A and B). We intentionally modified one of the parameters of the GC-IMS on batch B to force batch effect. Intrabatch corrections were performed on the data before feature extraction process. To correct interbatch differences we have found some correspondence among peaks present in samples of both batches. Features of batch B were transformed through a linear mapping to become similar to the ones in batch A.

II. Materials and Methods

A. Experimental

We have used mixtures of ketones (2-Butanone, 2-Pentanone, 2-Hexanone, and 2-Heptanone) to create two batches of samples, A and B. Both batches were nominally equal in terms of ketone mixture preparation: First, we added 1 mL of each of the ketones to generate a stock solution. Then, we pipetted 20 µL of the stock solution into a graduated flask containing 20 mL milli-Q water. We took 1 mL of this second solution and solved it into 50 mL of milli-Q water. Finally, each sample in a batch was generated filling a 20 mL glass grass vial with 1 mL of the latter solution. Batches A and B consisted of 10 and 13 samples, respectively. The stock solution was generated independently for batches A and B.

Data acquisition was performed by headspace sampling using a commercial GC-IMS unit (*FlavourSpec*, G.A.S., Dortmund, Germany). This instrument includes its own autosampler (PAL3, CTC Analytics). The GC-IMS method for batches A and B was identical (drift gas flow = 200 ml/m, column temperature = 60ºC, drift tube temperature = 60 ºC, injector temperature = 80ºC, transfer line temperature = 80 ºC) with the exception of a single parameter, the carrier gas flow (carrier gas flow A = 11 mL/m, carrier gas flow B = 20 mL/m).

B. Intrabatch Correction

We have corrected data within batches applying three signal pre-processing steps, first in drift time, and after that in retention time axes: 1) digital smoothing, 2) baseline removal, 3) peak alignment. All the pre-processing steps were performed using MATLAB 2019b (MathWorks, USA). Digital smoothing was carried out using *Savitky-Golay* filters, while baseline removal employing *Psalsa*, an upgrade of the standard asymmetric least algorithm [5]. A discussion regarding the values of the parameters for these techniques can be found in our previous work [3]). Different methods for peak alignment were used depending on the time axis. For drift time axis correction, we utilized *Correlation Optimized Warping* (segment length I = 25, slack parameter t = 10) [6].

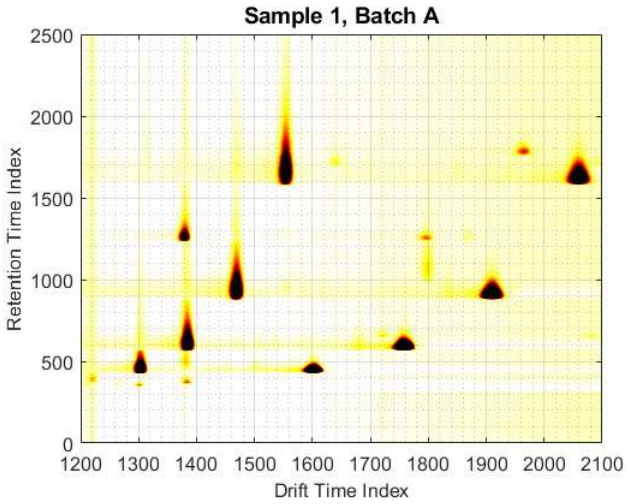

Fig. 1. Data corresponding to the first sample acquired in batch A. The chemical inforation in the image is found in peaks. Observe that the higher the mollecular mass of the ketone, the latter elutes from the chromatographic column. So the order of elution is the following: 2-Butanone, 2-Pentanone, 2-Hexanone and 2-Heptanone. Also, that for each ketone in the micture, the instrument returns 2 peaks: a monomer and a dimer. Note that data suffers from baseline problems in both dritft and retention time axes. The reactant ion peak has not been included in the figure to enhance the contrasts of of the other peaks.

Similarly, to correct retention time axis, we used *Parametric Time Warping*. This process was performed in two steps: First, a linear warping was applied globally to all samples in a batch. Next, an individual quadratic warping was computed for each sample.

C. Feature Extraction

To reduce data dimensionality keeping as much chemical information as possible, we have performed a 4) 2-dimensional peak picking followed by 5) 2-dimensional clustering among samples of the same batch. Our peak picking strategy was based on the computation of the numerical first and second derivatives on a data matrix. Zero-crossing points for the first and second derivative provided, respectively, the coordinates for the peak maximum and their inflexion points. Note that combining the positions of the inflexion points in retention and drift time, a rectangle can be formed. This rectangle defines the position where the peak is placed, and it is usually called region of interest (ROI). Note also that two or more peaks are close enough in the drift time – retention time space that they become convoluted in the same ROI. The intensity in one ROI, was defined as the sum of intensities of all features within a ROI. To cluster ROIS from different samples, we have applied the *k-medoids* algorithm [7] on the coordinates of the detected peaks, for all samples, and using as distance for the clustering the squared Euclidean distance weighted by variance. After finishing this stage, a peak table was obtained, where each row represented a different sample and each column the intensity corresponding to a different ROI.

D. Interbatch Correction

To correct differences between batches A and B, we have used the following approach: 6) We have scaled the retention time axis of batch B to match with the retention time axis of batch A. Then, 7) we have found some correspondence among peaks of batches A and B. Finally, 8) we have transformed drift and retention time axes of batch B to batch A using a linear mapping. To obtain the linear mapping between batches A and B, we used 10 samples per batch. The remaining samples of batch B (3 samples) were used as a test samples. Since both batches were previously corrected individually, we used the median peak coordinates within cluster as a cluster representative to perform interbatch correction. Retention time scaling consisted in multiplying the retention time axis of batch B by a factor c. To find its optimum value, we have swept c from 0 to 4 in steps of 10^{-2}. For each of these values, we have computed the area of the convex hull using the peak positions of batch B. Then, we have also calculated the area of the convex hull for the peak positions of batch A. The value for c that made the difference of convex hull areas between batches A and B closest to zero was selected for scaling the retention time axis of batch B. The rationale for that is that the convex hull for a set of points in a plane is the smallest convex polygon that contains them, so the optimum value for c is the one that makes the areas of the polygons for batch A and B equal. After scaling, we have clustered together the peaks of batches A and B to find a correspondence of peaks between batches. The selected algorithm to perform the clustering was *k-medoids*, [7] using as a distance the squared Euclidean distance weighted by variance. The most intense peak in a cluster belonging to the same batch was selected as a cluster representative in its batch. Consequently, the correspondence of peaks of batches A and B was one to one for each cluster. The original batch B data (that is, before scaling the retention time axis) and data from batch A were employed to compute the linear transformation that related A and B spaces:

$$B^{+}A = X \tag{1}$$

Where B^{+} is the Moore-Penrose pseudoinverse of B. Note the previous mapping could not have been done before knowing the correspondence among peaks from batches A and B. Once X was computed, we used it to correct drift and retention axes, for new data from batch B (B_{new}):

$$A_{new} = B_{new}X \tag{2}$$

Where A_{new} is B_{new} seen from the space of batch A

III. Results

Fig. 1 shows the raw data for the first sample acquired in batch A. Any peak in the image represents an ion detected by the Ion Mobility Spectrometer. Each of the ketones present in the mixture generated two peaks, a monomer and dimer (8 peaks). Additional peaks corresponding to unknow contaminants can also be found in the image. From the figure, it is evident that raw data exhibit baseline problems in both retention and drift time axes. The position of the ketone peaks slightly varied from sample within the same batch (not shown). We have pre-processed and extracted data features by batch. As a result, we have obtained two different peak tables. Fig. 2a show the coordinates, in drift and retention time indexes, of the median centroid positions for peak clusters in batches A (training data) and B (test data). Peaks belonging to batches A and B are represented, respectively, using black circle and red triangle markers. The distribution of peaks along the drift time axis is very similar for the two batches.

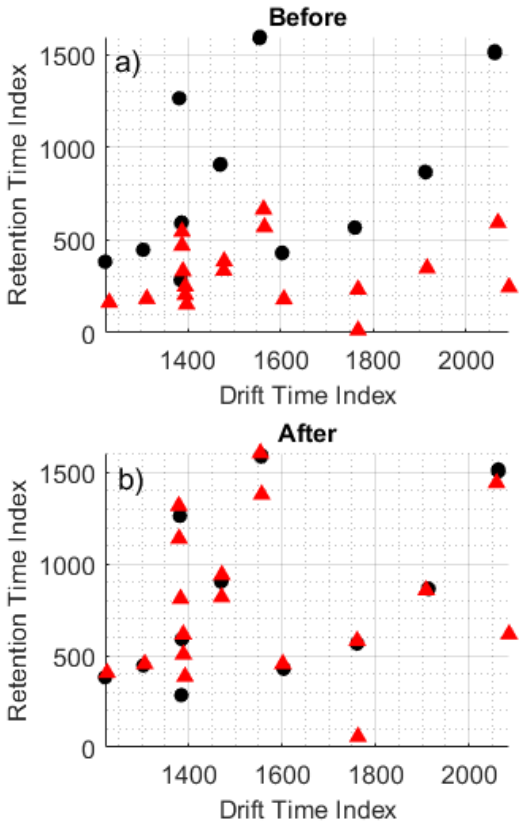

Fig. 2. Median position for the peaks detected in batches A and B, a) before batch correction, and b) after batch correction. In both scatter plots, peaks from batches A and B are represented using respectively, black circles and red triangles.

However, that is not the case when the same operation is performed along the retention time axis: Peak distribution is compressed for samples belonging to batch B. This result is agreement with the fact that GC-IMS methods for batches A and B only differed in the value for the carrier gas flow: the higher the carrier gas flow, the lower the elution time of the different ketones on the mixture.

To compute the retention time scaling factor and cluster the peaks of batches A and B, we have used the training set data. The optimum value for retention time scaling in batch B was 2.1. After clustering, only 9 clusters included ROIs from both clusters. We used these 9 clusters to perform the linear mapping between batches. The application of the linear mapping to the test data of batch B, led to the result shown in Fig. 2b. Colour and marker and codifications are the same as in Fig. 2a. The correction works fine for the ketones present in both batches.

IV. Conclusions

In this work, we have dealt with GC-IMS data, which provides complex, high-dimensional, sparse chemical information from samples. The data consisted in mixtures of four ketones, acquired in two different measurement campaigns (batches A and B). We have corrected a batch effect produced by an intentional modification of the carrier gas flow introduced between the two batches. Our strategy consisted in matching peaks that were present in both batches, and after that, applying a linear mapping that transformed the drift – retention time space of batch B to A. The correction successively worked for the data tested. Further investigation is required to confirm that the presented method generalizes properly when applied to compounds not considered during the linear mapping training stage.

Acknowledgement

We would like to acknowledge the Departament d'Universitats, Recerca i Societat de la Informació de la Generaralitat de Catalunya (expedient 2017 SGR 1721), the Comissionat per a Universitats i Recerca del DIUE de la Generalitat de Catalunya, and the European Social Fund (ESF). This research was also supported by the Spanish MINECO coordinated project TensorChrom (Total2SChrom, ref. RTI2018-098577-B-C21 and TENSOMICS RTI2018-098577-B-C22). Additional financial support has been provided by the Institut de Bioengienyeria de Catalunya (IBEC). IBEC is a member of the CERCA Programmes.

References

[1] R. Garrido-Delgado, L. Arce, A. V. Guamán, A. Pardo, S. Marco, and M. Valcárcel. "Direct coupling of a gas–liquid separator to an ion mobility spectrometer for the classification of different white wines using chemometrics tools". Talanta, 2011, 84, no. 2, pp. 471-479.

[2] R. Garrido-Delgado, M. del Mar Dobao-Prieto, L. Arce, M. Valcárcel, "Determination of volatile compounds by GC–IMS to 634 assign the quality of virgin olive oil". Food Chem., 2015, 187, pp. 572–579.

[3] R. Freire, L. Fernandez, C. Mallafré-Muro, A. Martín-Gómez, F. Madrid-Gambin, L. Oliveira, A. Pardo L. Arce and S. Marco, "Full workflows for the analysis of Gas Chromatography - Ion Mobility Spectrometry in Foodomics: Application to the Analysis of Iberian Ham aroma". Sensors, 2021, 18, pp. 6156.

[4] W.W.B. Goh, W. Wang, and L., Wong, 2017. "Why batch effects matter in omics data, and how to avoid them". Trends Biotechnol., 2017, 35(6), pp. 498-507.

[5] S. Oller-Moreno, A. Pardo, J. M. Jiménez-Soto, J. Samitier, and S. Marco. "Adaptive Asymmetric Least Squares baseline estimation for analytical instruments". In 2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14) (pp. 1-5). IEEE.

[6] G. Tomasi, F. Van Den Berg, and C. Andersson. "Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data". J. Chemom, 2004, 18(5), pp.231-241.

[7] H.S. Park, and C.H. Jun, 2009. "A simple and fast algorithm for K-medoids clustering". Expert Syst. Appl 2009, 36(2), pp.3336-3341.