LC/MS metabolomic signals normalisation through a Joint Diagonalisation of Covariance matrices

Francesc Fernández-Albert
Department d'Enginyeria de Sistemes
Automàtica i Informàtica Industrial
Universitat Politècnica de Catalunya
Barcelona, Spain

Rafael Llorach
Biomarkers and Nutrimetabolomics
Nutrition and Food Science Department
University of Barcelona
Barcelona, Spain

Mar Garcia Aloy
Biomarkers and Nutrimetabolomics
Nutrition and Food Science Department
University of Barcelona
Barcelona, Spain

Andrey Ziyatdinov
Department d'Enginyeria de Sistemes
Automàtica i Informàtica Industrial
Universitat Politècnica de Catalunya
Barcelona, Spain

Cristina Andrés-Lacueva
Biomarkers and Nutrimetabolomics
Nutrition and Food Science Department
University of Barcelona
Barcelona, Spain

&

Alexandre Perera
Department d'Enginyeria de Sistemes
Automàtica i Informàtica Industrial
Universitat Politècnica de Catalunya
Barcelona, Spain

***Abstract*—The computational processing and analysis of Liquid Chromatography/Mass Spectrometry signals in Metabolomics has several issues including ion suppression, carryover or changes in the sensitivity and intensity. Among these problems, the peak intensity might suffer from important drift effects that may even constitute the main source of variance in the data, which might lead to misleading statistical results. We propose a methodology based on a joint diagonalisation of covariance matrices prior to a data normalisation to tackle this problem. We compared our methodology with four other methods by calculating the Silhouette and Dunn clustering indices as a quality metric. In this comparison it was found that our methodology had better performance than any of the other four tested methods.**

I. Introduction

Metabolomics aims to asses the metabolic changes in a global way to infer biological functions and provide the detailed biochemical responses of cellular systems [1]. Liquid Chromatography/Mass Spectrometry (LC/MS) devices are among the most-used experimental setups in metabolomics. LC/MS analyses of biological samples such as urine or plasma give high-throughput data having a three index scheme: retention time, mass/charge ratio and intensity values [2]. In metabolomic data, the intensity values of the variables might be biased or might suffer from variations due to external factors. Among these factors is a contribution from the drift of the experimental devices, due to various causes such as column ageing in the case of LC/MS, temperature variations or contamination effects [3]. The presence of peak intensity drift in the data is an important issue, as its effects can be important enough to mask the real statistical behaviour of the data and may indeed be the largest source of variance in the data [4]. In most LC/MS protocols, quality control (QC) samples are regularly injected to ensure good analytical device performance [5]. In LC/MS metabolomics studies the quality controls have been carried out using pools of biological samples, spikes with standards or Milli-Q water samples [6]. These quality control samples consist either of a pooling of all the samples in the study or of a spike-in of some known metabolites (several classes having different types of QC samples might be injected). In the data preprocessing stage, one may distinguish two different steps: data normalisation and data equalisation. Data normalisation step is the mathematical process which makes the *variables* in the data set comparable, whereas the data equalisation step, makes the *samples* from the data set comparable. In the literature, many normalisation and equalisation methods may be found. Regarding equalisation methods, an approach using the injected samples for internal control (i.e. QCs) to fit a smoothed model for the intensity levels of certain features, and then to correct all the biological samples accordingly [5]. The R package sva includes the ComBat function which compensates the batch effects on microarray data using an empirical Bayes approach [7]. Equalisation methods based on a sample-wise correction for LC/MS metabolomic data have also been tested and compared by Veselkov et al. [4]. Their results suggest that a variance stabilisation transformation of the data, followed by a median fold change normalisation, gives the best performance as compared to three other methods. Among the equalisation methods, the one proposed by Artursson et al., based on component correction (CC), was developed in the sensor array field [8]. This method is based on the assumption that, in multivariate data, the drift direction is the first Principal Component (PC) of a PCA decomposition for a class consisting of measurements of the same samples. Such samples are known as technical replicates (i.e. there is no biological or chemical variation in addition to the variability of the technical replication of the measure). Once the drift direction is computed, the drift is removed from the data by subtracting the data projection on the drift direction from the original data. However, if some between-class variability is aligned with the drift direction, it will also be subtracted and some non-drift variability will be removed. A natural extension of the CC method is the one proposed by Ziyatdinov et al. which is based in a Common

Principal Component Analysis (CPCA) decomposition [9]. This method proposes modelling the drift contribution in the data as the direction capturing maximum variance that simultaneously diagonalises the covariance matrices of a set of classes. All the variability of the samples in that particular direction is considered to be drift-induced variability, and the projection of the data on that direction is subtracted from the data as in the CC method. In this paper, to find the drift model, we state the hypothesis that the intensity drift of the chromatograms is the common variance direction of all the QC classes that captures the maximum variance. In this context, we propose a preprocessing method based on a two-step approach by first equalising the data through a CPCA, and then normalising the data using a median fold change step.

II. Materials and Methods

A. Description of the Data

The samples were analysed by liquid chromatography coupled with a hybrid quadrupole time-of-flight (LC-q-TOF, Hybrid quadrupole TOF QSTAR Elite, AB/MDS Sciex) in positive mode using the protocol proposed by Tulipani et al. (Tulipani et al. 2011). Throughout all the analysis, data process quality control (QC) samples were analysed in order to monitor the stability and functionality of the system. The sample collecting span was of 18 days and there was a replacement of the chromatographic column in the process on day 14. There were 994 study samples and 182 QC samples. Three classes of QC samples were used for each batch:

- Water: Milli-Q water samples (n=96 samples).
- Spikes: Standard mixture solution (n=48 samples) consisting of 12 metabolites at the final concentration of 5ppm for all of them except for indole-3-acetic-2,2-d2 acid whose final concentration was of 10 ppm.
- Reference: Urine sample belonging to the one volunteer. (n=38 samples).

B. Preprocessing

All the methods were applied to the chromatograms without any prior feature detection. The R package XCMS was used to read the chromatograms of the mzXML files containing the sample data [10]. The chromatograms were aligned using an in-house developed R package (UB/UPC). The chromatographic data of all the files read were merged, creating an $n \times m$ chromatogram matrix X. This step required the binning of the retention time in m bins that were given by the XCMS package. Therefore, the chromatogram matrix had samples as rows and retention time as columns (in our case, n = 1176 samples and m = 441 retention time points). Thus, the i-th row of this matrix corresponds to the chromatogram of the i-th sample. From here on, the variable j refers to the retention time bins in the chromatogram matrix. A class-wise outlier detection and removal procedure was applied to the QC classes. 9 outlier samples were detected (4 samples in class *reference*, 3 in class *water* and 2 in class *spikes*).

C. Methods

The five methods compared in this paper (CPCA, CC, Median fold change, ComBat and our CPCA+Median Fold Change) have different input parameters. The methods based on a CPCA decomposition or the CC method involve a class (or classes) selection step to use them for the drift modelling. These methods also need as input the number of components of the drift decomposition which are supposed to be captured. The ComBat method needs the batch relation for each class, whereas the Median fold change method does not need any specific input parameter in addition to the dataset

1) Component Correction: The hypothesis underlying this method is that the drift direction is found in the first PC of a reference class. The methodology used to normalise this data is described in Artursson et. al. [8]. As the feature pattern of the QC samples was more complex than that of the other two QC classes, the reference class was selected to generate the PCA model. Because of this higher complexity, this class is better able to capture the drift in the data than would a class with a simpler feature pattern. The methodology proposed by Artursson et al removes one PC, but the method can be generalised to remove as many PCs as can be found in the data. We tested this method removing 1, 2 and 3 PCs.

2) Median Fold Change: The Median Fold Change method is not focussed on finding the drift direction. Its objective is to rescale the data to make the median fold changes of the variables close to zero. The methodology followed in applying this method is the one of Veselkov et.al., based on a sample-wise approach [4]. The first step of this method is to compute the median for each variable, thus obtaining a vector (equation (1)). This vector is used to rescale the original data set Y into a new one, $\hat{Y}$ (see equation (1)).

$$\hat{Y_{ij}} = \frac{Y_{ij}}{\hat{y_i}} \text{ where } \hat{y_i} = median_i(Y_{ij}) \tag{1}$$

To obtain the normalised data set Z^M, the data set Y is divided by the sample median of the matrix $\hat{Y}$ (defined as $\hat{w_j}$) as shown in equation (2)).

$$Z^M_{ij} = \frac{Y_{ij}}{\hat{w_j}} \text{ where } \hat{w_j} = median_j(\hat{Y_{ij}}) \tag{2}$$

where the superscript *M* refers to Median Fold Change.

3) ComBat: The ComBat method is a function of the R package sva. This function aims to correct the batch effects, which are known to be a source of bias, in gene expression experiments using an empirical Bayes approach; its extension to LC/MS metabolomic datasets is both natural and straightforward.

4) CPCA: CPCA is a generalisation of the PCA decomposition for different classes first introduced by Flury et. al. [11]. Say we have *k* classes and Σ_k are the set of their covariance matrices, then CPCA aims at finding a space such as the one defined by the V matrix shown in equation (3). In the space spanned by V, the covariance matrices for all the classes involved Σ_k are diagonal.

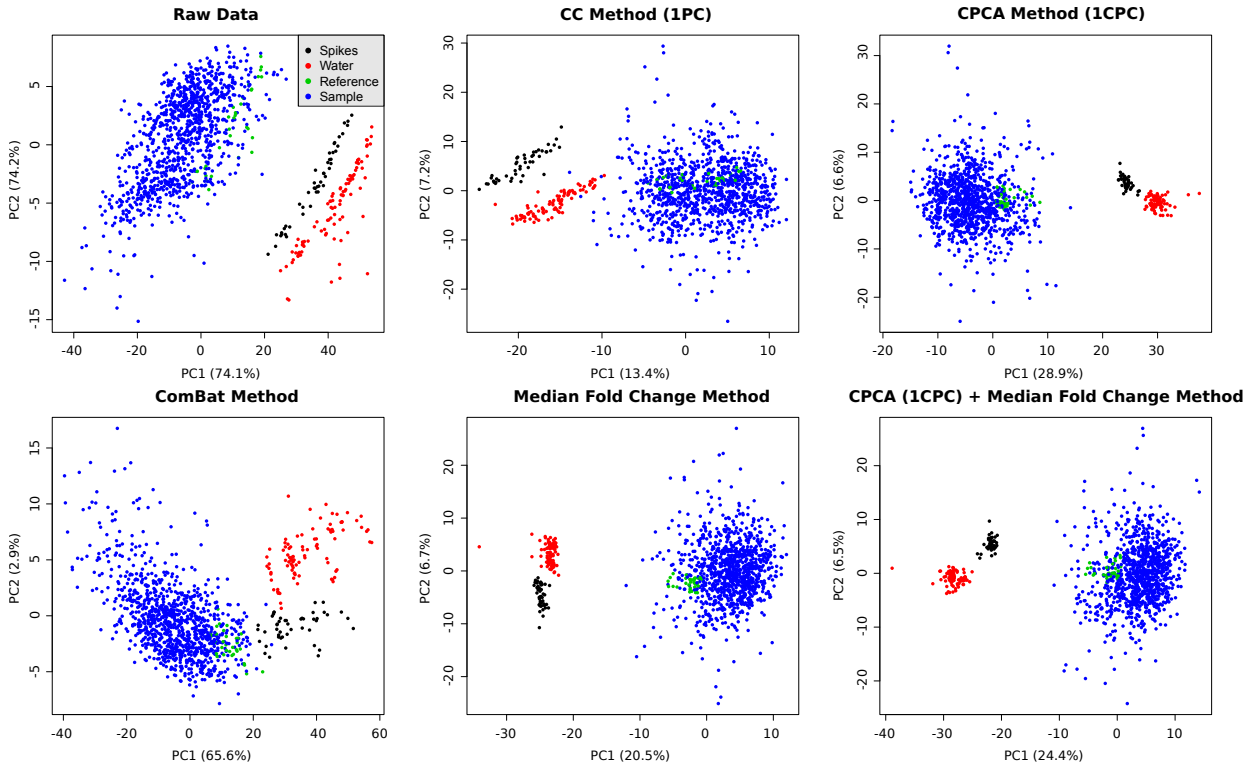

Fig. 1. Set of PCA Scoreplots showing the raw data and the effect on data for each method. The class labelled as sample is the study class. The numbers in brackets on the axes of the plots refer to the estimated variance for that particular direction in the data.

$$\Lambda_k = V^T \cdot \Sigma_k \cdot V \tag{3}$$

where Λ_k is the diagonalised covariance matrix for class k. Each one of the dimensions of this space is called a Common Principal Component (CPC). The hypothesis underlying the CPCA method for drift correction is that the drift direction is contained in the CPC capturing the largest variance. The CPC will be computed by using the Y_{QC} data set (i.e. there are three expressions like equation (3), using the different covariance matrices for the QC classes: Σ_r, Σ_{water}, Σ_{spikes}). In a similar way as in a PCA decomposition, given the desired number of CPCs and following a stepwise algorithm, it is possible to compute the number of CPCs one by one [12]. We have tested the values $Ncomps = 1, 2, 3$ separately for this method. Once the CPCs are found, the data set is projected onto this space as shown in (4)

$$Y_d^{CPCA} = (Y \cdot V) \cdot V^T \tag{4}$$

Y_d^{CPCA} contains the drift component in the data. To eliminate the drift from the data, the last step is to subtract this drift from the data (equation 5)

$$Z^{CPCA} = Y - Y_d^{CPCA} \tag{5}$$

where Z^{CPCA} is the corrected data set through CPCA.

5) CPCA + Median Fold Change: The method we propose consists of a two-step approach. Firstly, the data is equalised by removing the drift using CPCA and, in the second step, the data is normalised by applying the Median Fold Change method. As the CPCA method was applied three times with different number of extracted CPCs ($Ncomps$ in previous subsection), the proposed method will be computed for the same number of components ($Ncomps = 1, 2, 3$).

D. Validation

From the class definition in section II-A, it follows that a PCA score plot of all the classes should have the classes clearly separated in different clusters. We propose a quality measure for peak intensity drift correction methods based on the standard clustering internal measures Dunn and Silhouette for the QC classes in the principal plane (the plane explaining maximum variability of the data) score plot of all the classes (including the study class). The clustering technique used was k-means. The R package clValid was used to compute the quality indices [13]. In general, the greater the Dunn and Silhouette indices, the better the clustering, meaning that the QC classes are more easily separable in the principal plane and that the intra-class variance is lower.

III. Results and Conclusions

The top left plot in Figure 1 depicts a PCA score plot for the raw data using all the classes. The Figure shows that one of the main sources of variance is the interclass variability with

a similar tendency in all the technical replicates (QC classes). Therefore, we conclude that there is a clear drift component (having different sources) that is causing an important drift of the QC classes and which, in all likelihood, affects the samples in the study class as well. Table I contains the Dunn and Silhouette values for all the methods used, whereas Figure 1 depicts the PCA Scoreplots for the same methods. The CPCA+Median fold change method shows the highest clustering values (highest Dunn index when two components are removed and highest Silhouette index when one component is removed) and it has a slight advantage over the CPCA and the Median Fold Change methods. The Silhouette index (Table I) for the different CPCA methods applied suggests that the drift seems to be contained in just the first CPC, as the quality measures go down as more CPCs are removed from the data. The CC method corrects some of the drift in the data although a large drift component is still to be found in the data (Figure 1). The larger Dunn index value for the CC method as compared to the raw data value is evidence for the drift correction (Table I). However, this improvement is not validated by the Silhouette index which remains practically unchanged as compared to the raw Silhouette value. The ComBat method seems not to be the most suitable method for correcting LC/MS metabolomic data despite being used widely and successfully in the field of gene expression and methylation data. Although it corrects some batch effects in the study samples, the batch effects are still important in the QC classes after the correction (see lower left plot of Figure 1). The Median Fold Change method considerably improves both the Dunn and the Silhouette indices. A visual inspection of the resulting PCA score plot for the Median Fold Change method confirms this improvement (Figure 1). Nevertheless, the PCA score plot also shows that the *spikes* and *water* classes have similar shapes and these long shapes turn out to be caused by residual uncorrected drift effects. This fact suggests that, as the Median Fold Change method normalises the data without specifically trying to remove the drift, there may still be a source of variance in the data caused by the drift of the experimental device. On the other hand, because the methods based in the CPCA approach (CPCA and CPCA+Median Fold Change methods) are developed to model the drift direction, their resultant datasets show less residual drift in their corresponding principal plane score plot. Overall, in the context of LC/MS drift correction, the proposed two-step methodology shows better clustering properties of the QC samples for large metabolomic studies than the median fold change method. The method also shows a robust behaviour under small sample size conditions. Furthermore, unlike the median fold change method, the two-step method is able to capture intensity drifts that covariate with the retention time.

TABLE I
DUNN AND SILHOUETTE VALUES FOR ALL THE TESTED METHODS. THE NUMBER IN BRACKETS REFER TO THE NUMBER OF REMOVED COMPONENTS. THE HIGHEST CLUSTERING INDICES ARE SHOWN IN BOLD.

Method / Index	Dunn	Silhouette
None	0.029	0.560
CPCA (1CPC)	0.159	0.749
ComBat	0.074	0.553
CC (2PC)	0.249	0.600
Median Fold Change	0.171	0.719
CPCA (1CPC)+Median	0.208	**0.794**
CPCA (2CPC)+Median	**0.344**	0.690

ACKNOWLEDGMENT

This research was funded by Spanish national grants AGL2009-13906-C02-01/ALI, AGL2010-10084-E, 2014 SGR 1063, 2014 SGR 1566, the CONSOLIDER INGENIO 2010 Programme, FUN-C-FOOD (CSD2007-063) from the MICINN and Merck Serono 2010 Research Grants (Fundación Salud 2000). R. Llorach thanks the MICINN and the European Social Funds for their financial contribution to the R. L. Ramón y Cajal contract (Ramón y Cajal Programme, MICINN-RYC). This work was partially funded by the Spanish Ministerio de Ciencia y Tecnología through the TEC2010-20886-C02-02 and TEC2010-20886-C02-01 grants, and the Ramón y Cajal programme. A. Perera is part of the 2009SGR-1395 consolidated research group of the Generalitat de Catalunya, Spain. CIBER-BBN is an initiative of the Spanish ISCIII. F. Fernández-Albert thanks EVALXARTA-UB and Agència de Gestió d'Ajuts Universitaris I de Recerca, AGAUR (Generalitat de Catalunya) for their financial support. M.G.-A. thanks the Generalitat de Catalunyas Agency for Management of University and Research Grants (AGAUR) for the predoctoral FI-DGR 2011 fellowship.

REFERENCES

[1] O. Fiehn, B. Kristal, B. van Ommen, L. W. Sumner, S.-A. Sansone, C. Taylor, N. Hardy, and R. Kaddurah-Daouk, "Establishing reporting standards for metabolomic and metabonomic studies: a call for participation." *Omics : a journal of integrative biology*, vol. 10, no. 2, pp. 158–163, 2006.

[2] L. Xin, Z. Xinjie, B. Changmin, Z. Chunxia, L. Guo, and X. Guowang, "Lc-ms-based metabonomics analysis," *Journal of Chromatography B*, vol. 866, pp. 64–76, 2007.

[3] L. Burton, G. Ivosev, S. Tate, G. Impey, J. Wingate, and R. Bonner, "Instrumental and experimental effects in LC-MS-based metabolomics." *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, vol. 871, no. 2, pp. 227–235, 2008.

[4] K. A. Veselkov, L. K. Vingara, P. Masson, S. L. Robinette, E. Want, J. V. Li, R. H. Barton, C. Boursier-Neyret, B. Walther, T. M. Ebbels, and et al., "Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery," *Analytical Chemistry*, vol. 83, no. 15, pp. 5864–5872, 2011.

[5] W. B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J. D. Knowles, A. Halsall, J. N. Haselden, and et al., "Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry." *Nature Protocols*, vol. 6, no. 7, pp. 1060–1083, 2011.

[6] R. Llorach, M. Urpi-Sarda, O. Jauregui, M. Monagas, and C. Andres-Lacueva, "An LC-MS-based metabolomics approach for exploring urinary metabolome modifications after cocoa consumption." *Journal of proteome research*, vol. 8, no. 11, pp. 5060–5068, 2009.

[7] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods." *Biostatistics (Oxford, England)*, vol. 8, no. 1, pp. 118–27, 2007.

[8] T. Artursson, T. Eklöv, I. Lundström, P. Mårtensson, M. Sjöström, and M. Holmberg, "Drift correction for gas sensors using multivariate methods," *Journal of Chemometrics*, vol. 14, no. 5-6, pp. 711–723, 2000.

[9] A. Ziyatdinov, S. Marco, A. Chaudry, K. Persaud, P. Caminal, and A. Perera, "Drift compensation of gas sensor array data by common principal component analysis," *Sensors and Actuators B: Chemical*, vol. 146, no. 2, pp. 460–465, 2010.

[10] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification," *Analytical Chemistry*, vol. 78, no. 3, pp. 779–787, 2006.

[11] B. N. Flury, "Common Principal Components in K Groups," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 892–898, 1984.

[12] N. T. Trendafilov, "Stepwise estimation of common principal components," *Computational Statistics & Data Analysis*, vol. 54, no. 12, pp. 3446 – 3457, 2010.

[13] G. Brock, V. Pihur, S. Datta, and S. Datta, *clValid: Validation of Clustering Results*, 2011, r package version 0.6-4.