UNIVERSITAT DE BARCELONA

Multivariate Signal Processing for Quantitative and Qualitative Analysis of Ion Mobility Spectrometry data, applied to Biomedical Applications and Food Related Applications

Ana Verónica Guamán Novillo

FACULTAT DE FÍSICA

Departament d'Electrònica

MEMÒRIA PER OPTAR AL TÍTOL DE DOCTOR PER LA UNIVERSITAT DE BARCELONA

Doctorat en Enginyeria i Tecnologies Avançades (RD 99/2011)

Multivariate Signal Processing for Quantitative and Qualitative Analysis of Ion Mobility Spectrometry data, applied to Biomedical Applications and Food Related Applications

by

Ana Verónica Guamán Novillo

Director:

Dr. Antonio Pardo

Codirector:

Dr. Josep Samitier

Tutor:

Dr. Antonio Pardo

Chapter Four
Qualitative Analysis of IMS

4.1. Introduction

Qualitative models are deeply used for solving many research problems (Dixon-Woods et al., 2001, Patton, 1999). Among of them, discriminate groups(Stevenson et al., 2010), detecting the presence or not presence of a substance in an organism(Cen and He, 2007), enhancing the data analysis (Chen et al., 2010, Razifar et al., 2009, Statheropoulos et al., 1999), analyzing the evolution of some medication or vaccines in the organism (Webster and Bertoletti, 2001), and many other research interest. These research problems can be solve by using variety of algorithms with different grade of difficulty from a single principal component analysis to cluster algorithm(Wang et al., 2004) (Krause, 1998, Tsoukias, 2008).

Applications about the presence of explosives was one of the initial uses of IMS (Ewing et al., 2001). Later on, the interest as analytical technique was also moved on to discover which compounds or analytes are involved in some specific class into a clustering model. Hence, the development of classification models is one of the interests in the IMS field.

As soon as, bio-related applications have emerged, the need to develop strategies for building reliable qualitative models has also appeared. Certainly, there are exploratory techniques, which are commonly used for visually inspect the data, such as PCA (Bishop, 2006). However, there are other alternatives that are useful for tackling the common problems in IMS, such as the use of multivariate curve resolution algorithms (de Juan and Tauler, 2006, de Juan et al., 2000). Some of them will be discussed in the course of this chapter. Surely, it will be necessary to compare the results with a reference analytical technique in order to confirm them. In this thesis GC analysis has been used as reference technique for contrasting the IMS results.

The content of this chapter is split in two main parts. The first one seeks to enhance signal to noise ratio of IMS spectra trough pre-processing techniques, which are also used for the analysis present in chapter five. The second one presents two different alternatives to resolve a classification problem that can be used in IMS field. The classification problem is proposed to be solved either using the whole spectra information, or using a multivariate curve resolution algorithm and use the pure compounds which have more discriminant information.

4.2. Pre-processing of IMS spectra

Pre-processing seeks both to enhance the signal to noise ratio and prepare the spectra for a later data analysis in order to guarantee a certain degree of success in further quantitative or qualitative analysis. Pre-processing can be done in consecutive steps, even though there is no rule to be followed. In this work, we propose to perform the pre-processing in the following way: (i) noise reduction or signal smoothing, (ii) baseline remove, (iii) peak alignment of the same sample or same group of samples (same class), and (iv) a posterior alignment of samples from different classes.

4.2.1. Noise reduction or smoothing

It is well known that the performance of the noise reduction processes depends on the kind and characteristics of the noise present in the spectra. During the development of this thesis, two different kinds of noise have been identified which are close related with the specific hardware development of the IMS instruments. Figure 4.1 shows raw spectra of two different spectrometers: GDA2 and UV-IMS. Note, how different is the noise present in spectra and how the peaks of interest are been affected by the noise. In one hand, spectra from GDA2 present peaks that are really well resolved. It can be seen that the spectra has a baseline that need to be subtracted with a kind of noise that is not really affecting the peak information. Figure 4.1(b) shows a zoom in a region in which no information exist and it is appreciable that baseline behave is not lineal and noise seems to be high frequency noise. On the other hand, Figure 4.1(c) depict a really different scenario, quite different from GDA2, the peaks are wider than GDA2 and the noise is really interfering with the signal information, whereas the baseline has a lineal behave. Figure 4.1(d) shows also the artifacts of the spectra, which may probably due to some experimental manoeuvre. In addition, it is clear that UV-IMS has noise of both high and low frequencies.

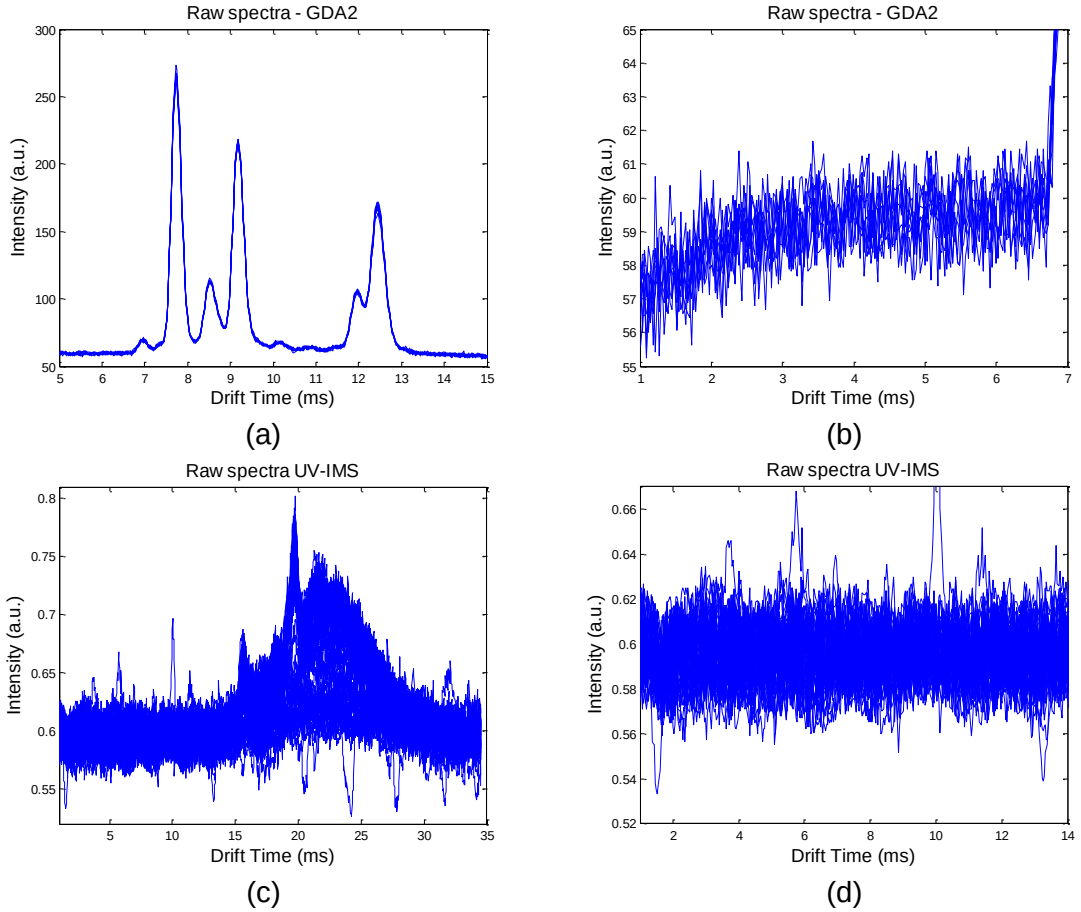


Figure 4.1 Raw spectra of a single measurement of two spectrometers. (a) GDA2 raw spectra (b) zoom of the tail without peaks information of the same spectra as (a), (c) UV-IMS raw spectra and (d) zoom of the tail with not relevant information.

Certainly, it is necessary to use different strategies for both kinds of problems. The simplest and easiest one is to tackle high frequency noise in GDA2. The information is not affected by the noise and the high frequency noise can be tackled using smoothing

algorithms such as savitzky and golay filter (Savitzky and Golay, 1964). In order to use this algorithm, two crucial parameters have to be set up which are the width of the filter, and the polynomial order that the algorithm need to fit. The order of the polynomial might be one or two, since polynomials of higher order do not significantly improve the baseline correction. So, using order one or two, either a straight line or quadratic function will be fitted into the signal. The width of the filter can be based on the peak resolution of the spectrometer, hence avoiding any distortion of the peak information. In the case of GDA2, the peak resolution is 32 (see chapter two, table 2.1), and several values, which were proportional to the peak resolution, were used to test the filter and observe the results. Figure 4.2 (a) and (b) shows the effects of applying a filter of order 2 with different widths in a single spectrum. It can be seen that the peak starts to be distorted when the width has a value proper to the peak resolution and peaks that are overlapped (11-13 ms) becomes broad.

On the other hand, the peaks get distorted while the width is large. Surely, there is a trade-off between noise reduction and holding the relevant information. In our case, a value above 7 and bellow 15 should be a good option. Actually, Figure 4.2 (c) and (d) shows the final result after being applied the filter of order 2 and width 15 to a single measurement, which consist of several spectra. On the one hand, noise has been diminished, but the height peaks also is attenuated. Moreover, the application of the filter improves de signal to noise ratio (SNR) (see Table 4.1).

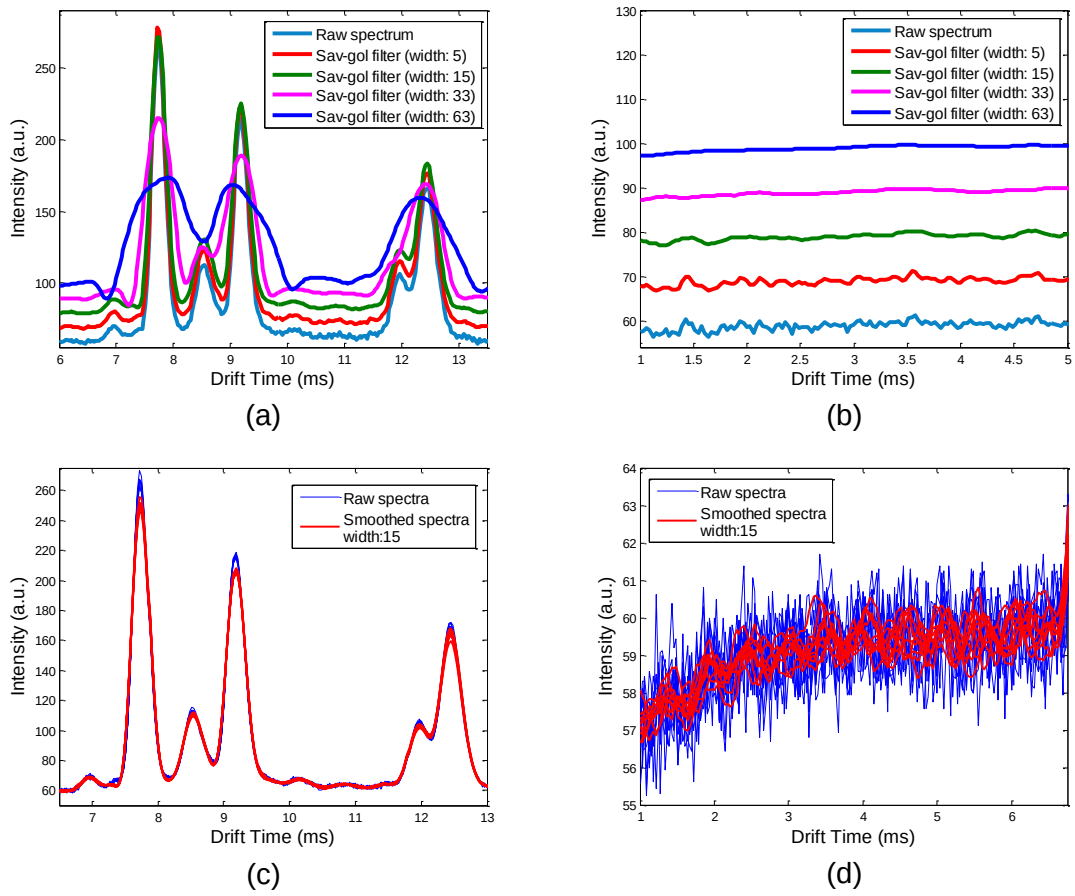


Figure 4.2 (a) Smoothing using savitzky-golay filter of order 2 using different width sizes, (b) Smoothing using savitzky-golay filter of order 2 using different width sizes (region with no peaks), (c)) Smoothed spectra using savitzky-golay filter of order 2 and width of 15, (d) Smoothed spectra using savitzky-golay filter of order 2 and width of 15 (region of no peaks).

The same filter was applied to UV-IMS spectra and the results are shown in Figure 4.3. It can be seen that the high frequency noise is smoothed, but the low frequency noise becomes ever clearer. SNR of raw spectra was calculated giving as result 18 dB and after filtering 24dB (see Table 4.1). Obviously, a slight enhancement was obtained, but it seems not enough since spectra is still noisy.

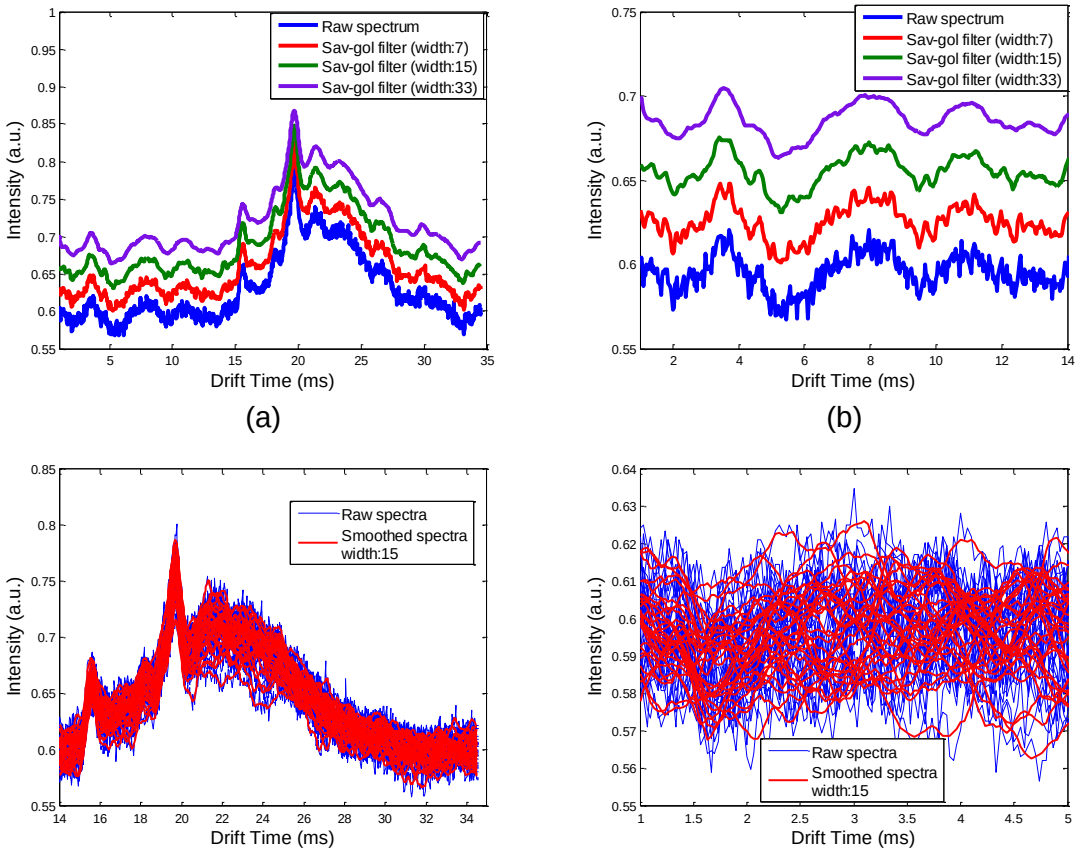


Figure 4.3 Savistzky and golay filter applied to UV-IMS spectra (a) Filter of order 2 using different width sizes applied to one spectrum (b) Filter of order 2 using different width sizes applied to one spectrum (region with no peak information), (c) Smoothed spectra using savitzky-golay filter of order 2 and width of 15, and (d) Smoothed spectra using savitzky-golay filter of order 2 and width of 15 (region with no peak information)

In this particular measurement, a low frequency noise was coupled to the signal and the source of this noise was already known, therefore the rejection of this noise becomes quite feasible. Conventional filtering techniques such as low-pass filters cannot be used due to they are likely to seriously distort the peaks shape. Therefore, two different approaches are detailed next. The first approach consist to use principal component analysis as filtering technique (Statheropoulos et al., 1999). The procedure consists of rebuilding the raw data eliminating sinusoidal contributions (periodic signals). Coming up, the steps are explained in detail.

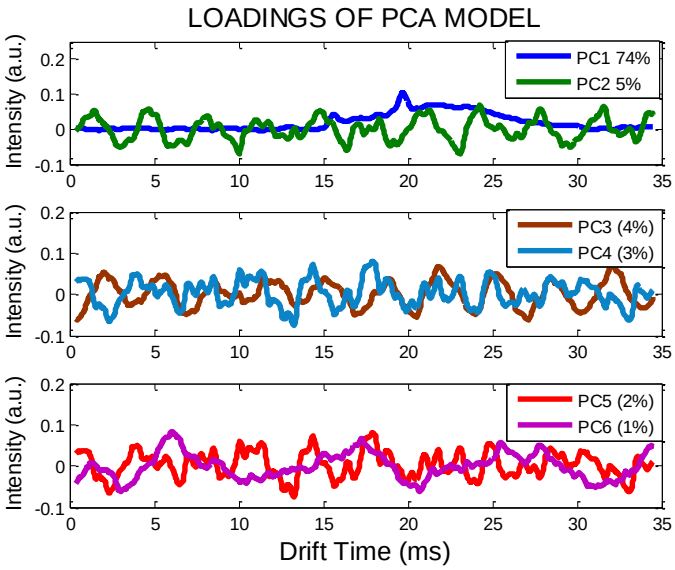
i. Mean center the dataset.
ii. Build a PCA model with as much principal components as cumulative variance is captured.
iii. Examine the loadings of the PCA model and select ones that have a periodic behavior, or through a Fast Fourier transform analysis that have the noisy frequency.
iv. Rebuild the data with PCs that do not have the noisy information.

$$X = TW^T$$

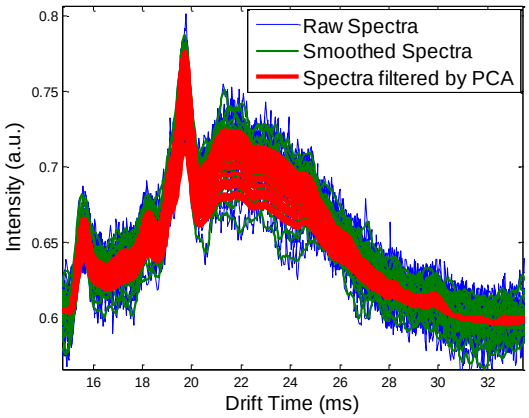
X is data , *T* scores and *W* loadings of the model.

v. Sum to *X* the mean center of (i).

The procedure explained above was applied to UV-IMS spectra after being filtered with savitzky-golay filter. The first six loadings of PCA model is shown in Figure 4.4(a) which cumulative variance is around 90%. It can be seen that the first principal component (PC), which recover 74% of variance, is the only one that do not have any periodical behave and the other loadings are likely to be a sinusoidal signal. In this case, rebuilding the data became really easy because the first component is the only one that has the main information. Nevertheless, there is a 26% of the information that is going to be lost which might totally be correlated with the noise. The final spectra after applying PCA is shown in Figure 4.4(b) and (c), and a clear improvement is observable. Actually, the SNR after applying PCA as filter was 58dB. Surely, it can be seen that the height of peaks diminish a little bit if it is compare to raw data, but the improvement is more important when the comparison is done in the tail where no information is located Figure 4.4(d).



(a)



(b)

Figure 4.4 PCA used as filter. (a) Loadings of PCA model, (b) UV-IMS spectra before and after filtering.

The second approach is the use of independent component analysis (ICA) which seeks to separate independent sources linearly mixed (Comon, 1994). The sources have to be statistically independent in order to be able to use this algorithm. In our case, the noise comes from the engine of a chemical hood located near to the spectrometer, so presumably both signals are independent. ICA has been extended used in biomedical signal processing, for instance, when recording electroencephalograms (EEG) on the scalp, ICA can separate out artifacts embedded in the data (Ren et al., 2006, Saruwatari et al., 2006). The procedure is quite similar to the PCA.

i. Mean center the dataset.
ii. Build ICA model, the number of independent components can be selected by using the information of PCA model.
iii. Examine the independent components (IC) and select ones that have a periodic behavior, or through a Fast Fourier transform analysis that have the noisy frequency
iv. Rebuild the data with ICs that do not have the noisy information.

$$S = WX,$$

where *X* is the data , *S* independent components and *W* linear statistic transformation of S.
v. Sum to *X* the mean center of (i)

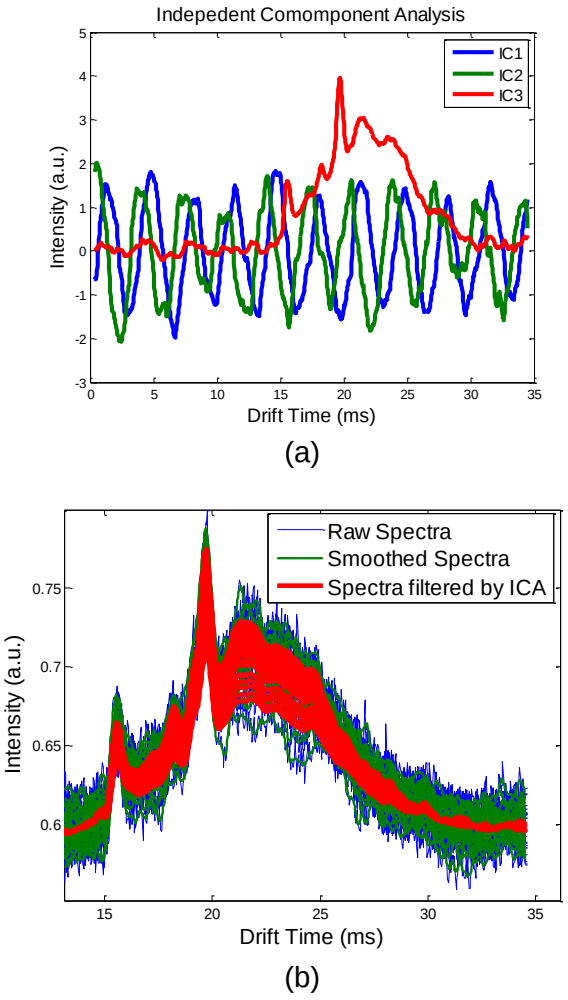


Figure 4.5 ICA used as filter. (a)Independent components, (b) UV-IMS spectra before and after filtering.

Figure 4.5 (a) depicts the independent components as result of ICA model. The two first independent components are clear sinusoidal signals and the third one is quite similar to the mean spectrum of the data. Thus, the spectra are reconstructed using only the third component and the results are shown in Figure 4.5 (b) and (c). Again, a really good improvement can be seen with a final SNR of 67dB. A final comparison in terms of SNR is shown in Table 4.1 and it is obvious that the best option is to use either PCA or ICA as filter in case of UV-IMS dataset.

Signal to noise ratio	Raw spectra	Savitzky-golay filter order 2 width 15	PCA	ICA
GDA2	82 dB ± 2dB	97 dB ± 3dB	-	-
UV-IMS	18 dB ± 5dB	24 dB ± 9dB	58 dB ± 3dB	67 dB ± 7dB

Table 4.1 Signal to noise ratio before and after using different filtering algorithms.

A single spectrum that shows the differences between the uses of the different filters is shown in Figure 4.6 (a). It can be seen that the first smoothing filter reduced the high frequency noise and an enhancement is evident. Then, the use of PCA and ICA as filter strategy improve significantly the SNR. SNR that was calculated using all filtering techniques and raw spectra (see Table 4.1), shows that the best approach is the use of ICA. Certainly, ICA is intended to separate independent sources of the signal, thus since the goal is to separate the coupling noise that come from another equipment, ICA is likely to be the best option. On the other hand, PCA require that the noise have to be orthogonal to the information in order to separate both signals. In this example, the principal components from 2 to 6 have noisy signals, which are orthogonal to the first PC that contains the main information of the data, thus reject the noise became feasible. That will unfortunately still not means that part of the noise is not orthogonal and it is present in the first PC. In addition, there is 26% of the data which are excluded and information can be part of this percentage as well as noise.

Two considerations have to be taking into account for using both strategies. The first one is the number of principal and independent components need to be chosen. In the PCA, it was hardly by chance that first component has the main information because the large amount of analyte present in the sample allows high peak intensity. Nevertheless, the information does not have to be in the first PC but may be distributed in other PCs. Therefore, it is important to preserve the most information in PCs as it is possible. The same idea has to be used when ICA is applied. The second consideration is how to known which PCs or ICs are noise. In this example, the selection was done by visual inspection, but it is not feasible if the number of samples is in order of hundreds or thousands. Thus, an automatic algorithm should be implemented. The algorithm can perform the fft of each independent or principal component and choose ones that have the fundamental frequencies of the coupled noise (only if it is known). An example is shown in Figure 4.6 (c), and you can see the fft of the two first independent components has a main fundamental frequency around 300 Hz and the bandwidth of the third independent component are between 0 to 100 Hz. The fft of raw spectrum and spectrum filtered by PCA and ICA are shown in Figure 4.6 (d), which shows that the main information is conserved and frequencies of noise are attenuated.

The results show that the filtering methodology works quite well, especially when the noise is orthogonal to the signal and therefore can be perfectly separated in one or more components. However, when the noise is not orthogonal to the signal, the performance of the algorithm decreases. For instance, the noise of VG-Test comes from the internal engine of the drift pump. Figure 4.6 (b) shows an example of VG-Test spectra before and after filtering in which ICA was used to eliminate the noise. The SNR of the raw spectra was 70dB and after filtering 91dB, even though a sinusoidal

noise can be seen in the final result showing that this technique was not feasible to completely eliminate it.

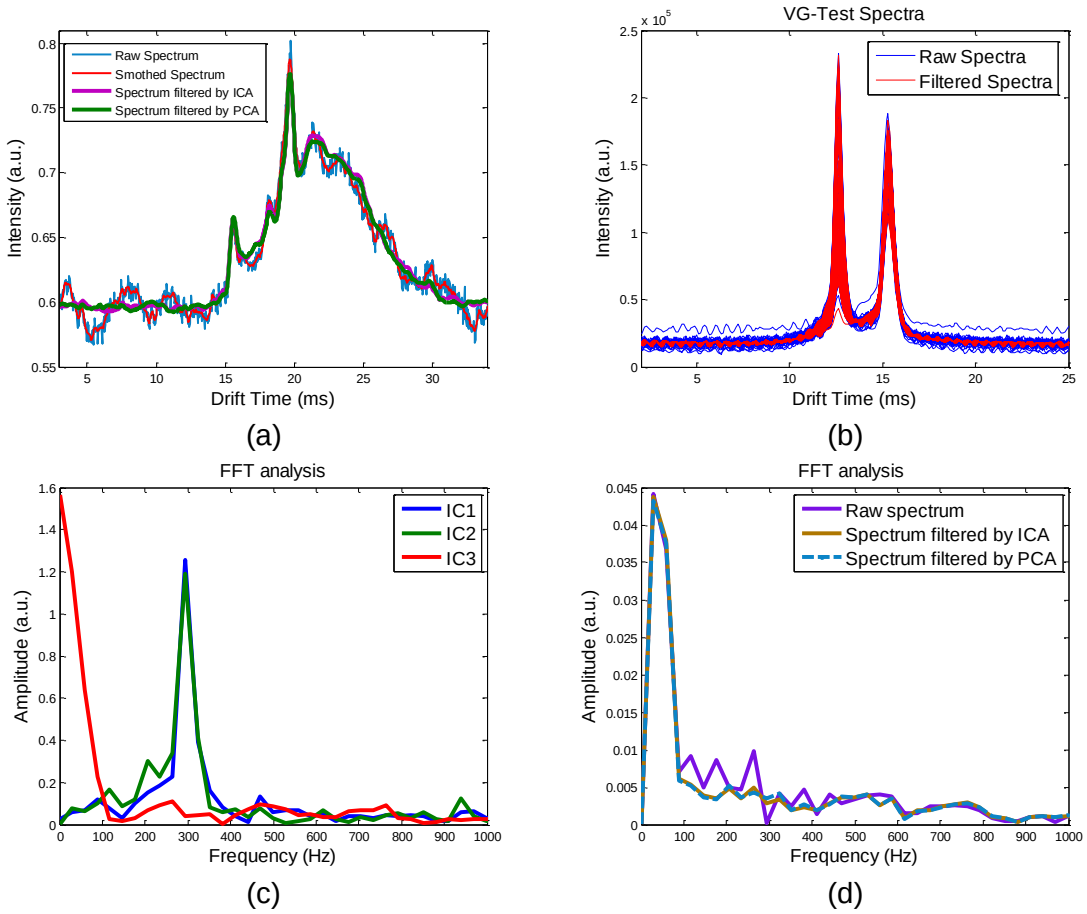


Figure 4.6 (a) Single spectrum before and after noise reduction , (b) VG-Test spectra before and after noise reduction (c) Fast fourier transformation of each independe component of ICA, and (d) fft of single spectrum before and after filtering.

4.2.2. Baseline subtraction

Baseline in analytical chemistry is a serious problem because if it is not done properly many problems might occur during the data analysis; among of them, peak detection, calculate the area of peaks of interest, and calibration problems. In IMS, as well as other analytical techniques, this problem needs to be solved. The main objective in analytical chemistry is to extract reliable information above background noise of the sample. One of the steps is to get a list of peaks and analyse them for selecting the most representative ones. This procedure is sometimes performed manually, thus there is a dependency on analyst expertise. The manually procedure is to fit a polynomial curve between the tails of the peak of interest. Nevertheless, this procedure is time consuming and there is likely to get errors on the determination. Therefore, developing and using automatic algorithms for fitting the baseline is worthy, especially in high dimensional data with hundreds of compounds.

The baseline correction in stand-alone IMS tends to be simpler than other analytical techniques such as GC/MS or IMS/MS. For instance, a chromatogram lasts several minutes and the baseline changes from time to time, thus determine a unique polynomial for the whole chromatogram results unavailable. On the contrary, a spectrum of a stand-alone IMS lasts miliseconds, thus there is not an important change of the baseline.

In this work the baseline correction is addressed with polynomial fitting, and the order of the polynomial will depend on the specific IMS instrument. The tails of the IMS spectrum, where no significative information is presented, are used to perform the fitting. In this thesis, the polynomial order was estimated manually, testing different polynomial order and calculating the final error when the baseline was subtracted.

Figure 4.7 depict the baseline counteraction process for the three spectrometers used in this thesis. In case of GDA2, spectrum from 1 to 5.5 ms and 18 to 27 ms, which no peak are located, were used for fitting a polynomial of fourth order. Note, at least the first period of drift time is constant due to the GDA2 has the RIP and it is unusual to have peaks before of it.

A polynomial of first order was used for UV-IMS spectra; the section used to fit the baseline in this example was from 5 to 13 ms and 31 to 34 ms. However, this values can significantly vary because peaks from analytes will appears at different drift times.

A polynomial of order 3 was used for VG-test and the polynomial was fitted using spectra from 3 to 10 ms and 20 to 23ms. In contrast to GDA, the last part of spectrum is more stable because there are not likely to appear peaks after the dopant TEP (peak that appears around 15 ms). During the estimation of the baseline, different order of polynomial was tested, and it was seen that higher orders did not provide better results than lower ones. It is important remark that the baseline does not change significantly in presence of different compounds, so that the polynomial was set up and used independently of the application.

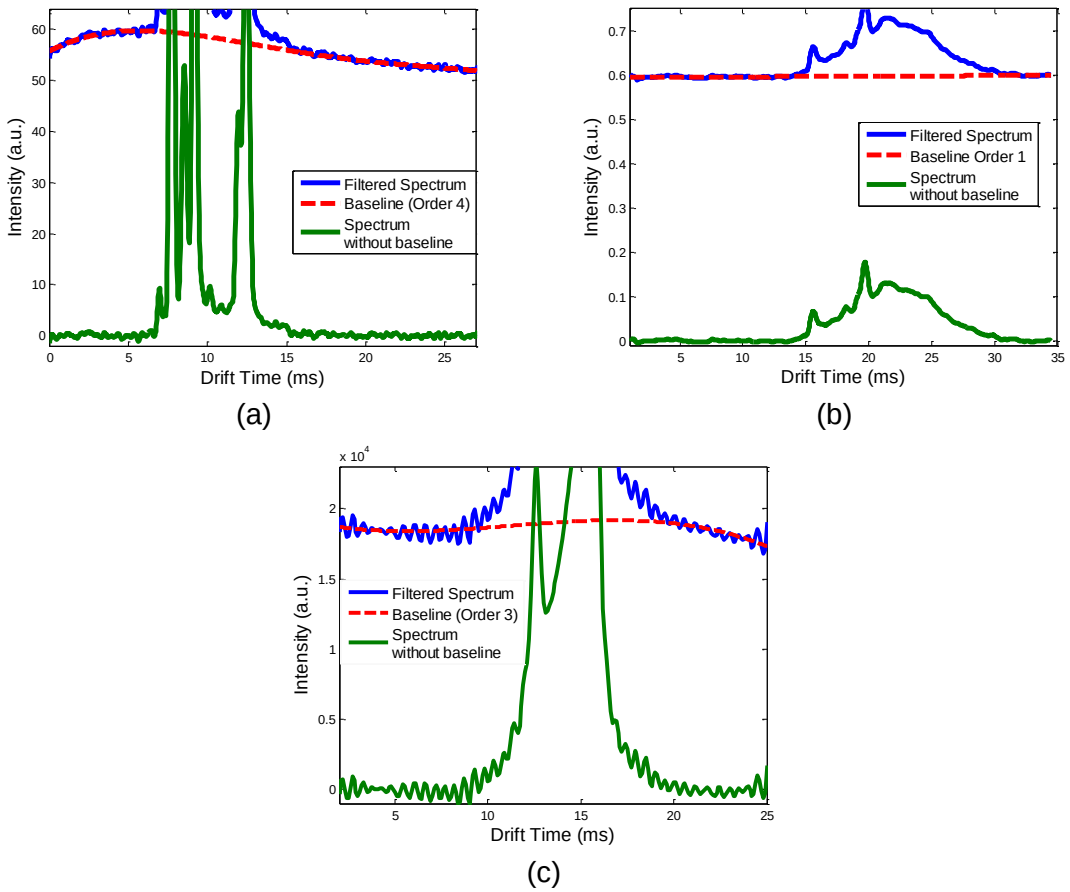


Figure 4.7 Baseline subtraction (a) GDA2, (b) UV-IMS and (c) VG-Test.

4.2.3. Peak alignment

The peak misalignment in IMS comes from changes of temperature, pressure, humidity, external and uncontrolled conditions. This misalignment usually appears as a shift of the main peaks. When the IMS has a reference peak such as reactant ion peak in GDA, which come from the radioactive ionization of the water and ammonia, the misalignment becomes evident. Since the misalignment may occur in few seconds of measurement, this problem became really important in the IMS field. Despite of the fact, there is not too much research about the kinds of misalignment in stand-alone IMS, the misalignment can be divided by misalignment additive, and multiplicative. The common one is additive misalignment where it is expected that a constant shift affects the whole spectrum. There is not a deep study that confirms stand-alone IMS has a multiplicative misalignment, but it is expected that shifts will depend on drift time position of each peak.

In this thesis, it was assumed that spectrum of stand-alone IMS has additive misalignment, and was focused in how to solve misalignment within spectra and between different samples. Thus, peak alignment problem can be divided into (i) alignment of spectra from a single measurement and (ii) alignment of spectra from different samples that contain same information (analyte). The first one is that a single measurement can last minutes, thus slight misalignments can happen. It is recommendable to fix it before trying to align different measurements or samples. An example is shown in Figure 4. 8 (a) and (b). This is a single measurement of a same compound that last around 12 minutes. It can be seen that from scan 200 (~3 minutes)

the main peak (9-9.5 ms) starts to shift little by little of the original drift time. This shift is mainly due to temperature and environmental changes and this behavior might appear even in few seconds of analysis. Actually, when measurements are taken during a day, further misalignment can be happen such as is seen in Figure 4. 8 (c). In this experiment, three different measurements were done in intervals of hours, thus the misalignment is larger than it was in a single measurement. This is the reason why it is advisable to align first each measurement separately and then do it with all samples.

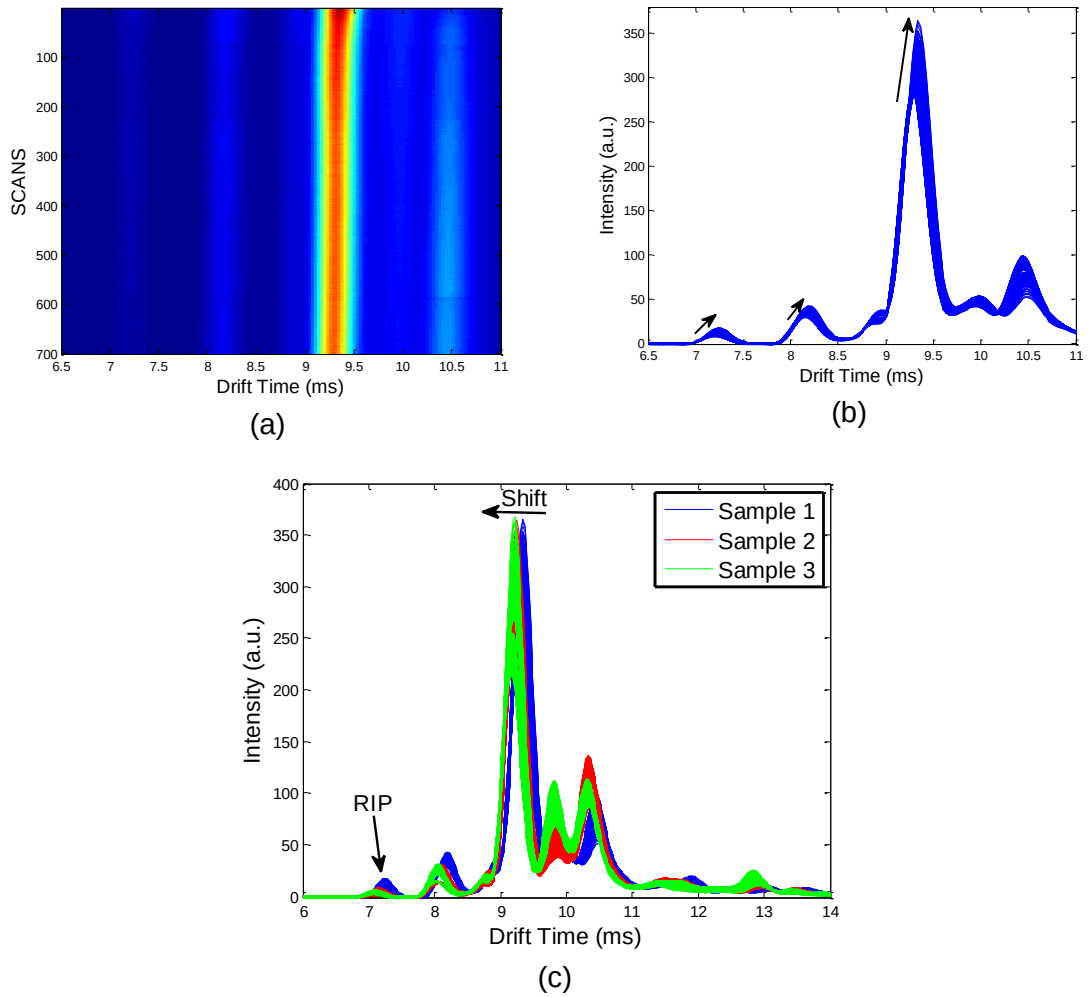


Figure 4. 8 GDA2 spectra (a) measurement of a single analyte during 12 minutes in which a slight misalignment is observable, (b) spectra of a single measurement that last 12 minutes, (c) different measurements of the same analyte in which a misalignment should be fix it.

Certainly, how to perform an alignment will depend on many factors such as having or not reference peaks, and to analyse if the misalignment is additive or multiplicative. On the other hand, when different samples need to be aligning, it is important to be sure that the compound of one sample is exactly the same compound in the others.

The best case is to have reference peaks of known compounds in order to perform the proper alignment. It is the case of GDA2 and VG-Test spectrometer that have a reference peak reactant ion peak (RIP) and TEP respectively. Consequently, the precision of the alignment can be tested or performed based on these peaks. Note that knowing the compound, the coefficient of reduced mobility (K_0) is also known; thereby the alignment can be done in basis of K_0 instead of doing in drift time axes. The

alignment is recommendable to be performed in terms of reduced mobility, since it can be corrected by temperature and pressure of each measurement using Eq. 3.2. When there is not an internal peak in every measurement, it is recommendable to add a dopant into the measurement or even better a calibrant. However and depending on the complexity of the sample, it should be not practicable.

For instance, Figure 4. 8 shows an example when the IMS has a reference peak, and an additive correction can be performed. It can be seen that all the set of peaks drift in the same direction, thus the drift of RIP can be taken as reference to correct the other peaks. In this case the resultant spectra is shown in Figure 4. 9(a) and (b) where no drift is observable. Exactly the same procedure can be applied when different samples need to be aligned. For align different sample, one sample (Figure 4. 9 (b)) is taken as reference and sample 2 and 3 are being corrected in terms of a reference spectrum. Since, these samples are replicates of the sample 1, there are more than one common peak to be aligned. Thus, alignment will consider all possible information. Again, an additive correction was performed and the result is shown in Figure 4. 9 (c) in which all samples are correctly aligned.

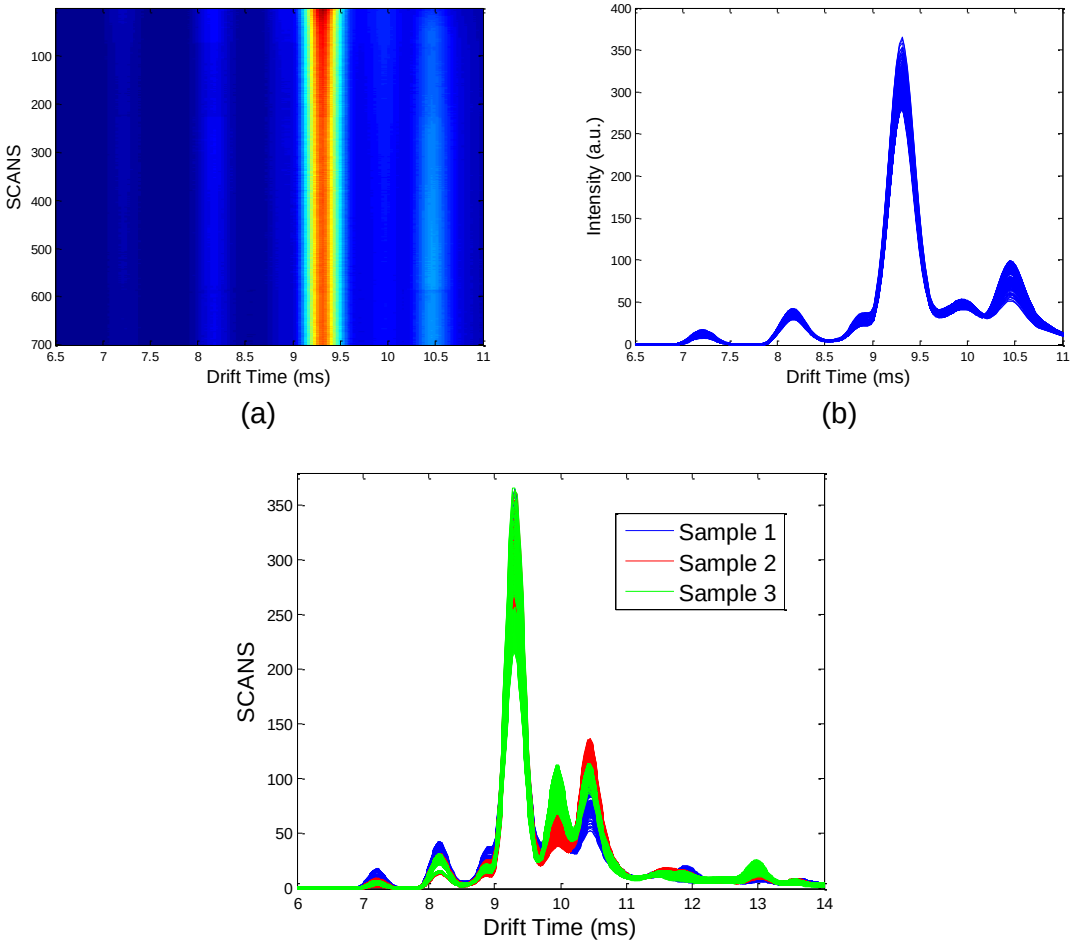


Figure 4. 9 Alignment of peaks using a reference peak (RIP). (a) measurement of a single analyte during 12 minutes, (b) aligned spectra of a single measurement that last 12 minutes, (c) different measurements of the same analyte.

The other scenario is when there is no reference peaks such as experiments done with UV-IMS. In this case, the best option is to have a reference spectrum and try to align using either common peaks or the whole spectrum. In this cases the alignment can be done either performing an additive shift or mix a shift together with a warping technique. Icoshift (Tomasi et al., 2011) is a warping technique which main principle is warp the spectra according to a reference spectrum. Icoshift is part of the warping techniques such as dynamic time warping (DTW) and correlation time warping (CTW)(Tomasi et al., 2004), but with the main advantage that works in the frequency domain allowing align a huge amount of data in few time.

Figure 4.10(a) is an example of four different measurements of wines sample, which are from different origins, analyzed with UV-IMS. Despite of the fact there is not any calibrant, it can be seen there are common peaks between samples. Therefore, the main idea is to choose one spectrum as reference and align the rest to the reference. In this example spectra of Montilla Moriles were used as reference and the other classes were aligned. The final result is depicted in Figure 4.10 (b) where the alignment is not yet perfect, but a signicant improvement was achieved. Note that we assume that there is common information (peaks) that should be align, but this is a particular case because all samples comes from the wine. Nevertheless, in qualitative analysis some assumptions should be taken in order to perform an accurate analysis. It is advisable for choosing reference peaks either to add a calibrant in the sample, or use an expert opinion for the identification.

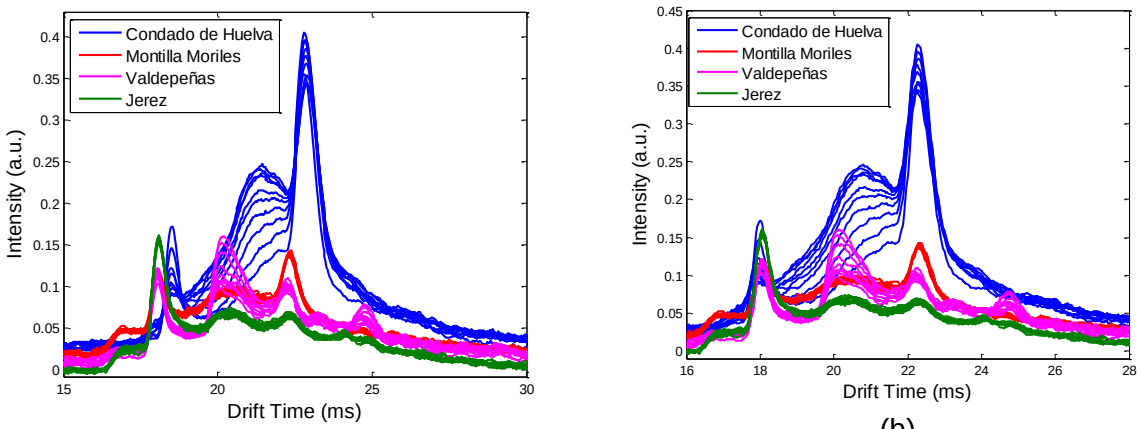


Figure 4.10 Wine spectra of different origins. (a) Preprocessed spectra, (b) Preprocessed spectra after aligment.

4.3. Discrimination of wines using Multivariate Analysis based on the information from whole spectra

Wine consumption is widely spread around the world. According with the International organization of wine, the world wine consumption in 2013 has been of 238,7 millions of hectolitres (iCEX, 2014). This huge figure has, of course, a very important impact on the economy of wine producer countries, such as Spain, France and Italy. Due to this direct impact on the economy, the fight against wine fraud is becoming an important issue. In recent years, public administrations and wine industry have driven research for quality improvement and fraud detection, which has been carried out within the chemical science(Holmberg, 2010).The quality of the wine has a direct relationship with the wine composition. Many factors have influence on the composition as, for example, grape variety, soil and climate, culture, yeast, winemaking practices, transport and storage (Bisson et al., 2002, Arvanitoyannis et al., 1999).

This work studies the viability of IMS as analytical technique for differentiate wine from different origin. A full optimization of the acquisition system, called CFS-GPS-UV-IMS (Garrido-Delgado et al., 2011), was firstly carried out to extract the volatile compounds present in wine samples and in-line analysis by UV-IMS equipment. Subsequently, a signal and data processing study was applied to classify wines according to their origin from the IMS recorded spectra. In addition, the analysis of the same set of wine samples by GC-FID were carried out so to compare the obtained results by IMS to demonstrate the potential of this technique.

The dataset consist of 56 wine samples from different origins analyzed with UV-IMS. Figure 4.10 (b) depict a single measurement of each kind of wine after being preprocessing as was explained before. Figure 4.11 depict the mean spectra of each sample in which can be seen main differences between classes. Note that a peak around 17ms can cluster the wines samples in two groups: (i) Condado de Huelva and Jerez, (ii) Valdepeñas and Montilla Moriles. Nonetheless, this difference seems to be attributing to ethanol content and not to the origin of them. Therefore, it is expected that the other peaks contribute more to discrimination according to the origin.

Each measurement comprised 50 spectra but only 34 spectra (15-49) were used for data evaluation because the first part and last part of the analysis do not provide relevant information. The relevant information in all cases was included only in the spectral region between 15.4 to 27 ms (351 variables). The dimension of the data matrix is 1904 x 351, corresponding to the 56 wine samples by 34 spectra for each sample and the 351 useful variables.

As it was mentioned before (section 4.2) data must be carefully pre-processed, since any inaccuracy introduced at this stage can cause significant errors in the statistical analysis. First a smoothing Savitzky-Golay filter of order 3 was used to improve the signal to noise ratio of all spectra of wine samples. Later, the baseline from each spectrum was corrected subtracting the mean value of an empty area of peaks (between 0 to 15 ms), common to all the original spectra of wine. Additionally, all spectra were aligned with a shift in x-axis based on a polynomial function fitted to a reference peak. New positions of the peaks are maximally close among the different spectra.

Finally, the precision of the method proposed was assessed by analyzing the same wine sample on the same or three different days under identical testing conditions. The within-day precision was obtained in eleven replicates on the same day and the between-day precision was obtained in three replicates within three consecutive days. The within-day and between-day precision values were obtained by using all data of the range selected in each sample (mean value of 34 spectra per sample from 15.4 to 27 ms). The within-day and between-day precision values obtained were 2.2% and 3.1% respectively calculated as relative standard deviation (RSD).

In order to find possible disturbing outliers for the pattern recognition analysis, a Hotelling's T (Bishop, 2006) square statistic test using a confidence interval of 95% has been implemented. As a result of this test, three samples were discarded from *Montilla-Moriles* wines, one sample was discarded from *Jerez* wines and four samples were discarded from *Valdepeñas* wines. Therefore, after the discarded samples, the final data base for the analysis is composed of 48 samples, 12 samples per class. The dimension of the new data matrix was 1632(48x34) x 351.
Table 4.2 summarizes the data and variables per each type of wine sample analyzed.

	Samples	Montilla-Moriles	Jerez	Condado de Huelva	Valdepeñas
		Number of spectra per sample			
	1	34	32	–	33
	2	34	–	34	34
	3	34	34	34	–
	4	29	34	34	–
	5	34	34	34	34
	6	–	31	34	34
	7	–	32	34	34
	8	34	34	34	29
	9	34	32	34	33
	10	34	34	–	34
	11	33	34	34	31
	12	28	31	–	34
	13	34	–		30
	14	34			–
	Number of samples	12	11	9	11
	Number of spectra	396	362	306	360
Data set	**Training**	9 Samples	8 Samples	6 Samples	8 Samples
	Validation	3 Samples	3 Samples	3 Samples	3 Samples

Table 4.2 Summary of wine dataset analyzed by UV-IMS.

As it was seen in Figure 4.11(a) spectra present some differences between classes, so that it is to be hoped that exist a pattern able to discriminate the four wines. Additionally, there may be a set of compounds (peaks) more discriminative of each class. Note that dataset (Table 4.2) has a high dimensionality and also note that the number of features is very much larger than number of samples. Thus a dimensional reduction technique is required prior to build any classification model. As it has been commented in chapter 2, both peaks height and area, are commonly calculated to tackle this problem. However the univariate strategy is naïve and not useful on applications where IMS is used with water-chemistry configuration, which means that it is non-selective. This is the case of the wine application, where spectra also present peaks overlapping and surely there is a grade of uncertainty on the correspondence of each peak with an specific compound. Thus a better option is to work with the whole

spectra and, using the proper techniques, extract the most informative information which can give a feasible discrimination. In this work PCA together with LDA was proposed as dimensionality reduction technique with discriminatory trait, thus PCA is used to reduce the data dimension and new projection of LDA is going to maximize the discrimination between classes. This strategy has been used before by other authors, in order to overcome the known tendency of this algorithm to overfitting in small-sample-size problems, where the dimensionality is higher than the number of vectors in the training set (Westerhuis et al., 2008, Smit et al., 2007).

The data was initially splitted in two subsets (Table 4.2): training and validation data. The first one, training set, was used to estimate the calibration model which contains 1224 samples that correspond to 75% of the total amount. The remaining data (408 samples) is used to validate the model. Principal components were used on training dataset in which 3 PCs were selected that jointly explained 95.7% of the total variance. However, by this approach, the different wine samples could only be separated by their alcohol content which it was not the objective of this work (Figure 4.11 (b)). It can be seen that there are two main clusters (i) Condado de Huelva and Jerez and (ii) Valdepeñas and Montilla Moriles. Both clusters depict that there is a separation by the alcohol content and it can be demonstrated in the loadings (Figure 4.11 (c)). The loadings of the two first principal components show that the peak around 17 ms, which peak is related with alcohol, has more importance than the others.

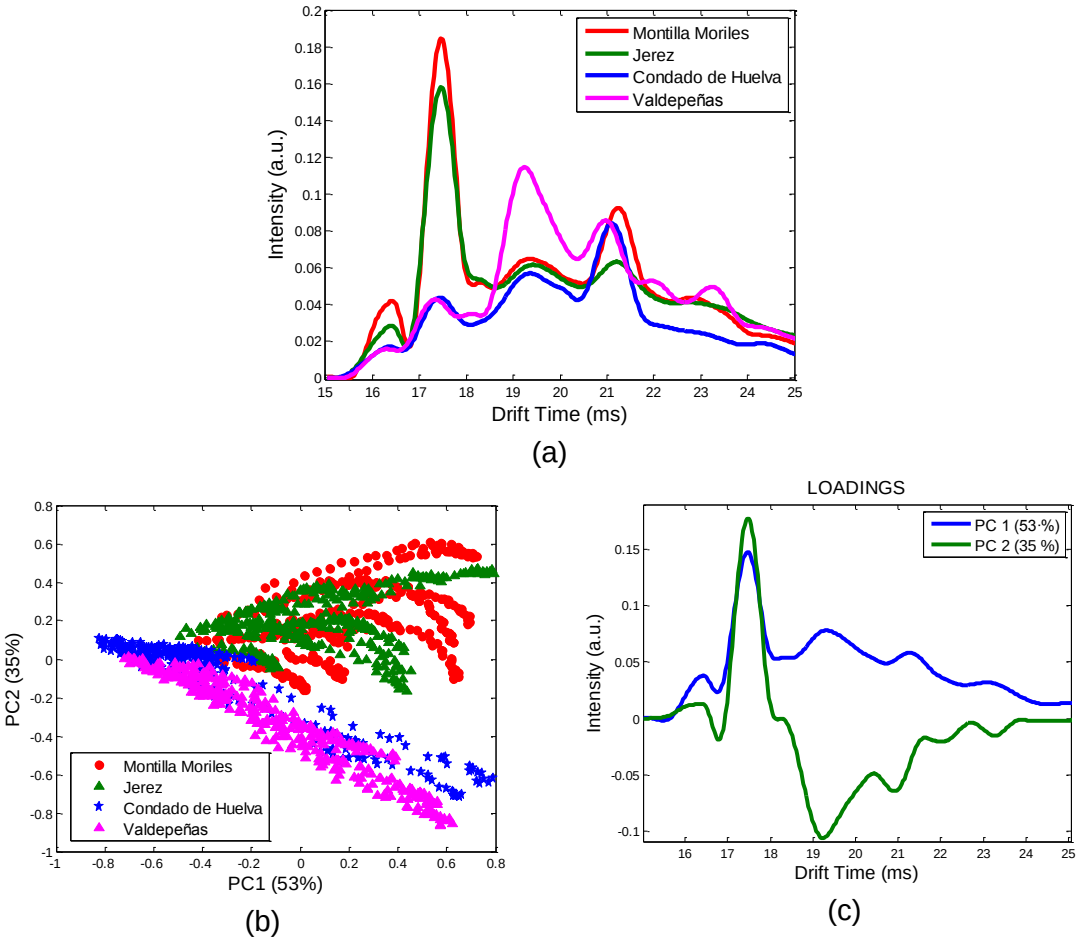


Figure 4.11 (a) Mean spectrum of each wine sample (b) Scores of the PCA model (c) loadings of the PCA model

Knowing that the main objective is to perform wine clustering avoiding the alcohol content, performing a LDA can help to achieve this goal. The main idea is to build a LDA model in the PCA space and use a classifier to test the discrimination. The kNN classifier was used to get the final accuracy of PCA-LDA model. Despite of the fact the model was built using each spectrum, the classification was done by each sample. Thus, a sample is assigned to a wine class through a majority vote procedure, i.e. if the majority of its spectra belong to that wine class. In addition, the number of principal components was determined by cross-validation using bootstrap algorithm (Felsenstein, 1985, Efron, 1979). Under the Bootstrap validation procedure, the training set is randomly selected (with replacement) over the total number of data and the remaining samples that were not selected for training are used for the validation. This procedure is repeated for a specific number of folds ($B = 100$). It must be again highlighted that selection has been done over the samples, not over the spectra, i.e. to select a wine sample means to select all its corresponding spectra. For every step in the procedure, PCA and LDA combination is built using the information of the training set. Then the validation set is projected over the model and a kNN classifier ($k = 3$) is used for estimate the classification rate of the model.

A scanning from 4 until 20 PCA dimensions has been done in order to test the performance of the PCA–LDA strategy. Figure 4.12 shows the evolution of the overall classification rate with error bars representing the confidence interval at 95%

confidence level. It can be seen that the best classification rate is achieved with 16 PCs but on the other hand, taking into account the statistical significance, the classification rate of 8 PCs is comparable to the best solution, with a 92.0% classification rate value with confidence interval (89.0%, 95.0%) at P = 0.05 confidence level.

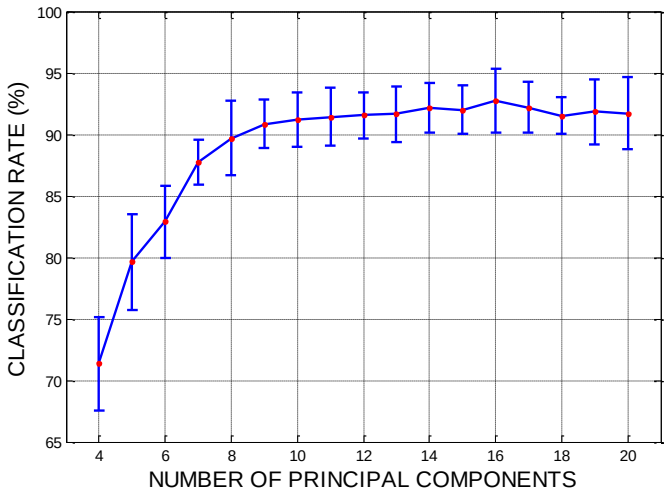
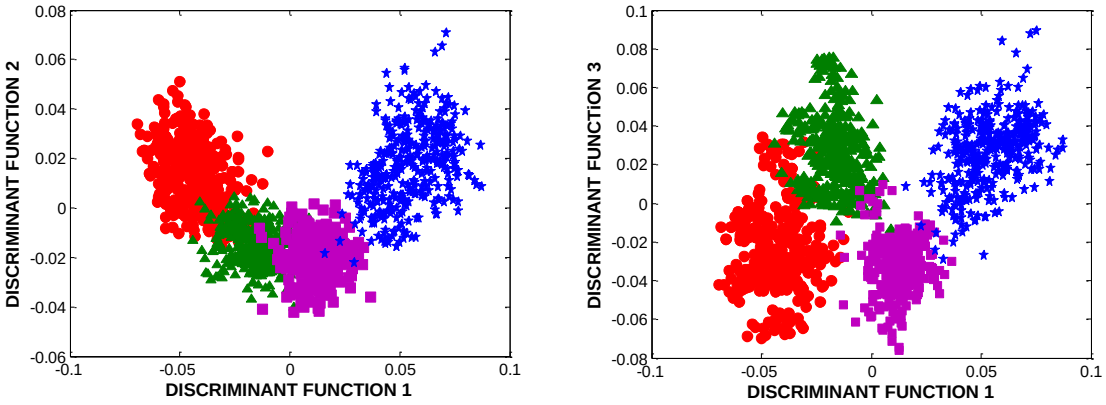


Figure 4.12 Scanning plot of PCA-LDA strategy from 4 to 20 principal components.

Figure 4. 13 (a) and (b) depict the LDA model that was built using the 16 PCs, which shows the best classification rate using boostrap (Figure 4.12). It can be seen that the first discriminant function is still separating the sample based on the alcohol content. However, the two remaining discriminant functions are able to separate the samples by their origins. The classification rate using 16 PC was 93 %(90.5-95.5%).



(a) (b)

Figure 4. 13 Scatter plot for the LDA obtained using 16Pcs on training set from IMS data. Montilla-Moriles (red circle), Jerez (green triangle), Valdepeñas (liliac square) and Huelva (blue star).

It was seen in Figure 4.12**¡Error! No se encuentra el origen de la referencia.** that from 8 PCs the improvement of the model is slightly different from 16PCs. The final model was done using the 8PCs and the classification results are shown in Table 4.3. It can be seen that the model can classify the four wine classes with an accuracy of 92%. That shows that IMS can be potentially used as analytical technique for discriminante wine origins avoiding the alcohol content.

Confusion Matrix (8 PCs)		PREDICTED			
		Montilla-Moriles	Jerez	Condado de Huelva	Valdepeñas
REAL	Montilla-Moriles	93%	7%	0%	0%
	Jerez	10%	87%	3%	0%
	Condado de Huelva	3%	6%	87%	4%
	Valdepeñas	0%	0%	1%	99%

Classification Performance	
Classification Rate	**92% (89%-95%)**
Montilla-Moriles	93% (90%-95%)
Jerez	87% (83%-90%)
Condado de Huelva	87% (83%-91%)
Valdepeñas	99% (97%-99.5%)

Table 4.3 Classification performace of PCA-LDA model using boostrap validation.

The same samples were analyzed with a reference analytical technique in order to compare and confirm the IMS results. Despite of the fact, the GC analysis is out of the scope of this thesis; a briefly summary of the main results will be presented. The area of 36 compounds was integrated from the chromatogram, 11 of which were identified as acetaldehyde, methyl acetate, ethyl acetate, methanol, 2-butanol, 1-propanol, isobutanol, 1-butanol, 2-methyl-butanol, and 3-methyl-butanol. It was observed that acetaldehyde, methyl acetate, 1-propanol, isobutane, 1-butanol, 2-methyl-butanol and acetoin classify to the wines in two groups and the area of 3-methyl-butanol was as different as the wine origin.

A signal processing strategy similar to IMS was applied to GC dataset. Figure 4.14 shows the resultant PCA model from GC dataset. The first PC discriminate the samples by the alcohol content similar to IMS data set. The second PC separate by the wine origin. Using kNN classifier (k = 3) with hold out validation, a percentage of 96.5% of good classification is obtained, with a confidence interval (88.2%, 99.9) at P = 0.05 confidence level. Just a single sample, corresponding to *Valdepeñas* wine was bad labeled.

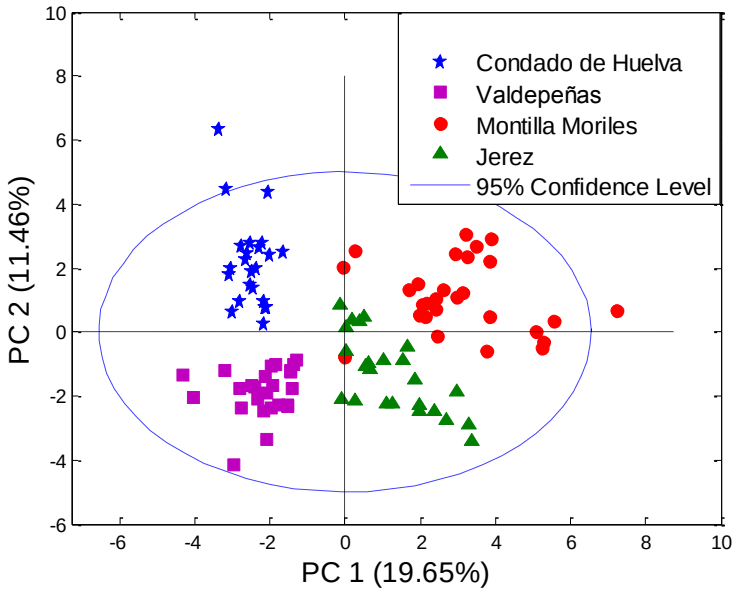


Figure 4.14 PCA model of GC dataset.

Although a good classification of the white wine samples has been achieved with both methods, with the CFS-GPS-UV-IMS method none sample pretreatment is required while in the chromatographic method a prior dilution and addition of the internal standard to the wine sample are necessary before analysis. Others advantages of the IMS technique are the short analysis times and a lower cost of the CFS-GPS-UV-IMS compared with a GC system.

Possible identification of the profile of the wine samples analyzed by IMS

The wine present a huge number of volatile compounds, such as, alcohols, esters, aldehydes or ketones (Holmberg, 2010). Although the objective of the work was not to carry out the identification of the volatiles compounds presents in the wine samples, the profile of the spectra from wine samples obtained by CFS-GPS-UV-IMS was studied to identify some analytes. To achieve this goal, all the compounds identified by GC-FID method were studied to check if these analytes could also be determined by UV-IMS. In Table 4.4, the boiling point, ionization potential and vapor pressure to ambient temperature of these compounds are summarized. Only the compounds with an ionization potential lower that 10.6 eV can be ionized by a UV lamp. Therefore, all the compounds with a ionization potential lower that 10.6 eV could be determined using the proposed IMS method but only three compounds (acetaldehyde, methyl acetate and ethyl acetate) shown signal. Although the other compounds have an ionization potential lower that 10.6 eV their boiling points are above 80ºC. Moreover these compounds have a very low vapor pressure at room temperature for all these reasons, these compounds were not identified in the IMS spectrum. IMS has the disadvantage that it is difficult to identify peaks in their spectra due to ion-molecule reactions and its low resolving power, but by comparing the ion mobility spectrum of a wine with one spiked wine sample shows an increase can be seen in the bands from spectrum between 15 and 27 ms. Therefore, in this preliminary study we could confirm that acetaldehyde,

ethyl acetate and methyl acetate contribute to obtaining the fingerprint of a wine sample analyzed using the proposed method.

Compounds	Ionization potential (eV)	Boiling temperature (°C)	Vapor pressure (mmHg to 20 °C)	GC(1)	IMS(2)
Acetaldehyde	10.23	20.08	756.8	yes	yes
Methyl acetate	10.25	57	165	yes	yes
Ethyl acetate	10.01	77	73	yes	yes
Methanol	10.85	64.7	97.7	yes	no
2-Butanol	9.88	99	12.5	yes	no
1-Propanol	10.22	97.1	14.9	yes	no
Isobutanol	10.02	108	8	yes	no
1-Butanol	9.99	118	5	yes	no
2-Methyl-1-butanol	>10.6	128	3	yes	no
3-Methyl-1-butanol	>10.6	132	2	yes	no
Acetoin	>10.6	148	2.7 (25 °C)	yes	no

Table 4.4 Compounds analyzed by GC-FID(1) and CFS-CPS-UV-IMS(2) (Garrido-Delgado et al., 2011)

In this work, a vanguard analytical system (CFS-GPS-UV-IMS) has been proposed for extraction in-line of volatile compounds present in liquid samples. This method has been applied to the analysis of white wine samples from different origins and different alcohol content. In this way, characteristics profiles from each group of wine samples have been obtained. Later a detailed chemometric signal processing was carried out to classify the different wine samples. A good classification was obtained by firstly reducing the data dimensionality by PCA followed by LDA and finally using a kNN classifier.

On the other hand, these same wine samples have been analyzed using a chromatographic method using GC-FID. Later a chemometric treatment has been carried out too. Using the data of the areas from all the peaks (36 peaks) from the chromatogram and applying a PCA, all the wine samples have been classified correctly. Moreover LDA and later a kNN classifier have been used to get a good classification too.

4.4. MCR and SFFS as classification methodology: Application for detection of SEPSIS in rats.

In the previous section a classification model was done using the whole spectra information. Another option is trying to extract pure compounds with multivariate curve resolution techniques and use the concentration profile for the subsequent classification. This methodology is going to be used in the application for detection of SEPSIS in rats. The main goal is to find a set of compounds responsible for discrimination between healthy rats and rats that were induced SEPSIS(Guaman et al., 2012).

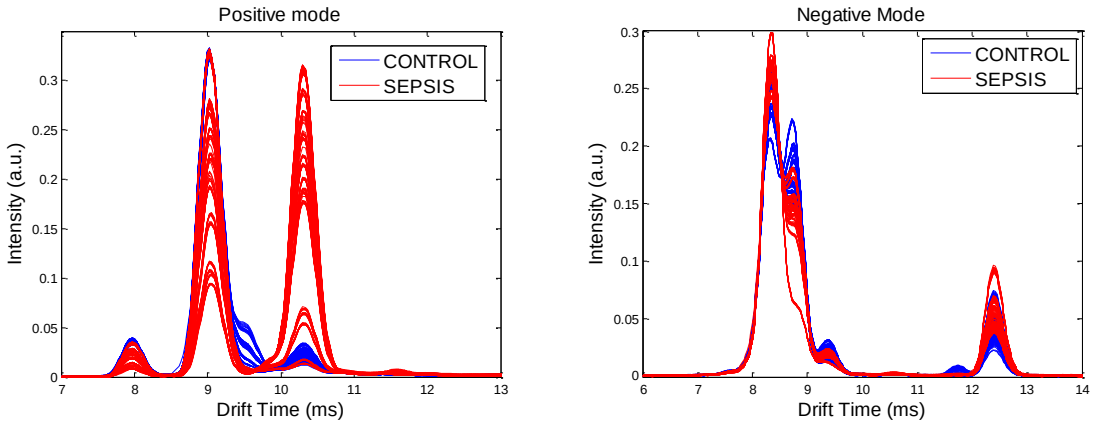
Despite the evolution of intensive care medicine and the broad range of clinical systems nowadays, sepsis is still the first cause of death in non-coronary critical care units. Traditionally, sepsis diagnostics use culturing techniques of blood, urine, cerebrospinal fluid and bronchial fluid, among others. The major drawback of culturing techniques is the time needed to develop the culture, usually between 24 and 48 h. Although other techniques such as ELISA, ProCalcitonin Test (PCT) assays and DNA detection by Polymerase Chain Reaction (PCR) are faster, they need between 2 and 6 h to obtain a response and they are incapable of following the dramatic changes occurring in sepsis. In the face of a lack of a real-time monitoring system for sepsis, breath analysis with IMS must be considered a promising and prospective alternative.

The potential capability of breath tests for the diagnosis of sepsis has been indicated in some works (Miekisch and Schubert, 2006) but, as far as we know, sepsis still remains untested by IMS technology. Other technologies such as GC/MS are also capable of offering a high performance in breath analysis but usually they cannot provide the portability and simplicity of the IMS measurements. IMS is more suited to the clinical trend of developing bedside patient systems but unfortunately it cannot identify easily unknown volatile compounds in a sample, so, in this respect, GC/MS measurements complement this lack of knowledge as a reference technique. This study includes, for the first time, the measurement with IMS technology of rats' breath infused with LPS from *E. coli* as a sepsis animal model. This represents a first step in the potential applicability of IMS for the diagnosis of sepsis in human patients.

A pathophysiological rat status was carried out to each rat. As expected, pulmonary edema was found only in the LPS-treated rat group (SEPSIS) compared to control animals. Moreover, concentrations of circulating inflammatory markers in plasma were significantly increased in LPS-infected mice compared to controls. Whereas in the control animals the concentration of IL1-β and TNF-α were 1.51±1.01 pg/mL and 1.43±0.14 pg/mL, respectively, in the LPS-injected animals these concentrations rose to 313.45±81.80 pg/mL and 5.99±0.30 pg/mL, respectively.

The IMS dataset featured 10 spectra from 40 breath samples (10 healthy rats + 10 LPS treated rats and an additional replicate of each one). Since the breath analysis was done using GDA2 spectrometer, two spectra of each sample is obtained -one positive and one negative IMS mode. Figure 4. 15 (a) and (b) shows the spectra of both classes in positive and negative mode respectively. It can be observed that there are some peaks remarkably different between both classes. In addition, spectra are really reproducible between them, it may be because of rats are genetically identical and its

diet and external condition were precisely controlled. Certainly, in human samples the spectra would be quite different due to the high and uncontrolled variability.



(a) (b)

Figure 4. 15 Spectra from breath analysis in control and SEPSIS rats. (a) Positive mode IMS and (b) Negative mode IMS.

MCR-LASSO(Pomareda et al., 2010) was used to decompose IMS raw spectra into their pure contributions: pure spectra components, S, and their related concentration time evolution, C, were extracted (Figure 4. 16). As a result, fourteen relevant pure components were obtained from negative and positive spectra. Undesirable contributions appeared at a drift time of 9.575ms in positive mode and at a drift time of 8.99ms in negative mode. Anesthesia (drift time=12.48ms in negative mode) as well as pure components related to the RIP peaks in positive mode (drift time=8.06ms and 9.03ms) and negative mode (drift time 8.363ms) were identified but were not considered for further evaluation. At the end of this process, eight pure components had been obtained.

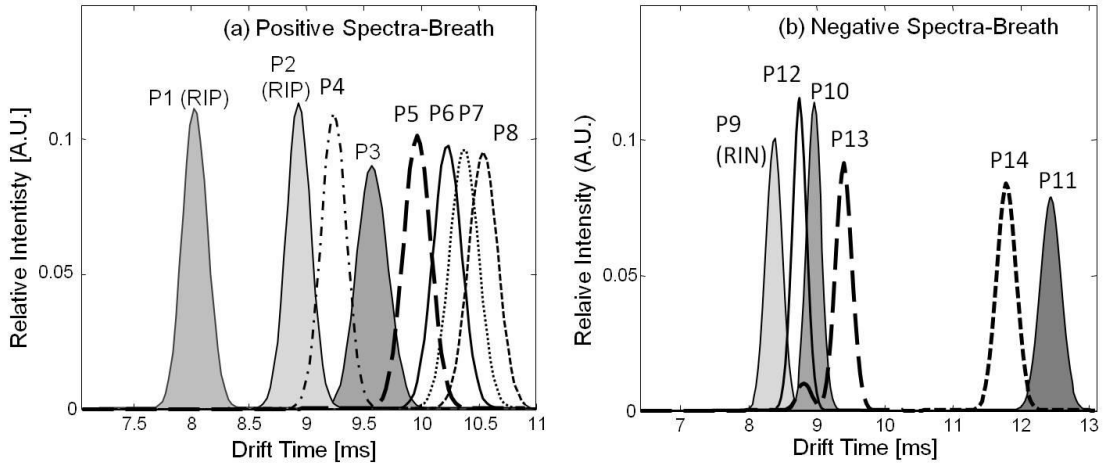
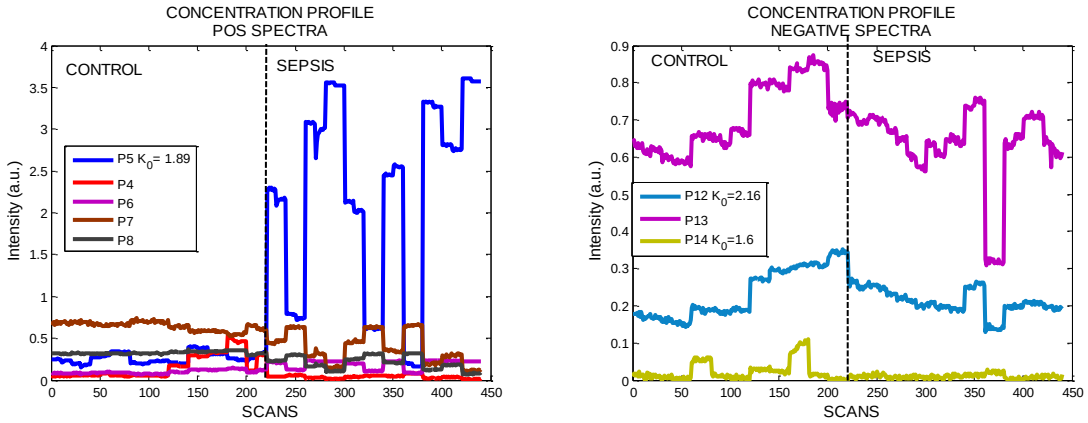


Figure 4. 16 Spectra profile from MCR-LASSO analysis. Pure components peaks (P) from MCR-LASSO results for rat's breath. Every component from P1 to P14 has its Reduced Mobility *K*0 (cm2 V−1 s−1) for positive and negative mode. Filled peaks correspond to anesthesia, air pollution and reactant ion peak from IMS, and the others are related to compounds from breath. Positive Spectra: P1 (*K*0 = 2.35): RIP comes from Nitrogen ion species, P2 (*K*0 = 2.11): RIP comes from water ion species, P3 (*K*0 = 1.97): a component from laboratory room air, P4 (*K*0 = 2.04), P5 (*K*0 = 1.89), P6 (*K*0 = 1.84), P7 (*K*0 = 1.82), P8 (*K*0 = 1.79). Negative Spectra: P9 (*K*0 = 2.25): RIN, P10 (*K*0 = 2.11): a component from laboratory room air, P11 (*K*0 = 1.52): anesthesia, P12 (*K*0 = 2.16), P13 (*K*0 = 2.01), P14 (*K*0 = 1.60).(Guaman et al., 2012)

The concentration profile of the compounds that do not have undesirable contributions for both positive and negative mode is shown in Figure 4.17 (a) and (b) respectively. These concentration profiles were used to build a model for discrimination between control and SEPSIS rats. In addition, a feature selection algorithm was used in order to get a subset of discriminatory compounds correlated with the disease. Actually, some compounds are really different between control and SEPSIS.



(a) (b)

Figure 4.17 Concentration profile of different compounds present in breath samples (a) positive mode and (b) negative mode.

As a result of the SFFS selection, the subset consisting of compounds with reduced mobility of $K0_1$=1.89 $cm^2V^{-1}s^{-1}$ (positive spectra), $K0_2$=2.16$cm^2V^{-1}s^{-1}$ and $K0_3$=1.60 $cm^2V^{-1}s^{-1}$ (negative spectra) were selected. Figure 4.18 shows the distribution of rats in the space of the three selected compounds. For easier interpretation, two plots of $K0_1$ versus $K0_2$ and $K0_1$ versus $K0_3$ have been shown, as opposed to a three-dimensional plot. Bootstrap validation was applied to estimate the discrimination between healthy and LPS-treated rats and the final result was an accuracy of 99.8% (99.7%-99.9%), a specificity of 99.6% (99.5%-99.7%) and a sensitivity of 99.9% (99.8%-100%). The confidence limits were calculated at a 95% confidence level.

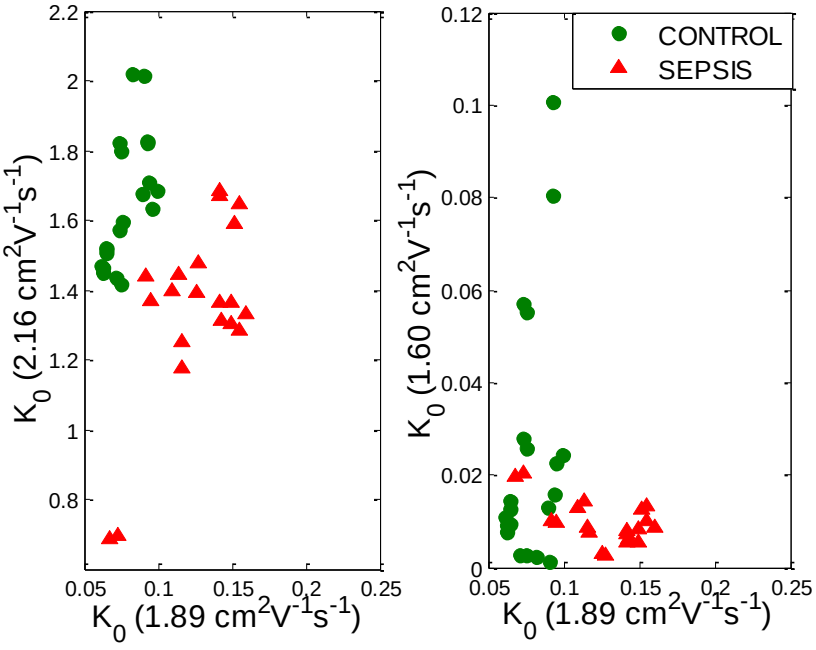


Figure 4.18 Plot of IMS samples using the three compounds that were selected by SFFS algorithm.

As it was seen before the results are really promising due to there are a set of discriminative compounds. Nonetheless, these compounds are unknown; thereby the same breath samples were analyzed with a reference analytical technique as GC/MS. Figure 4.19 shows chromatograms obtained from diseased rats and healthy rats. Note the abundance of peaks and slight differences between both chromatograms. Moreover, how reproducible are the chromatograms between them as it was seen in IMS data.

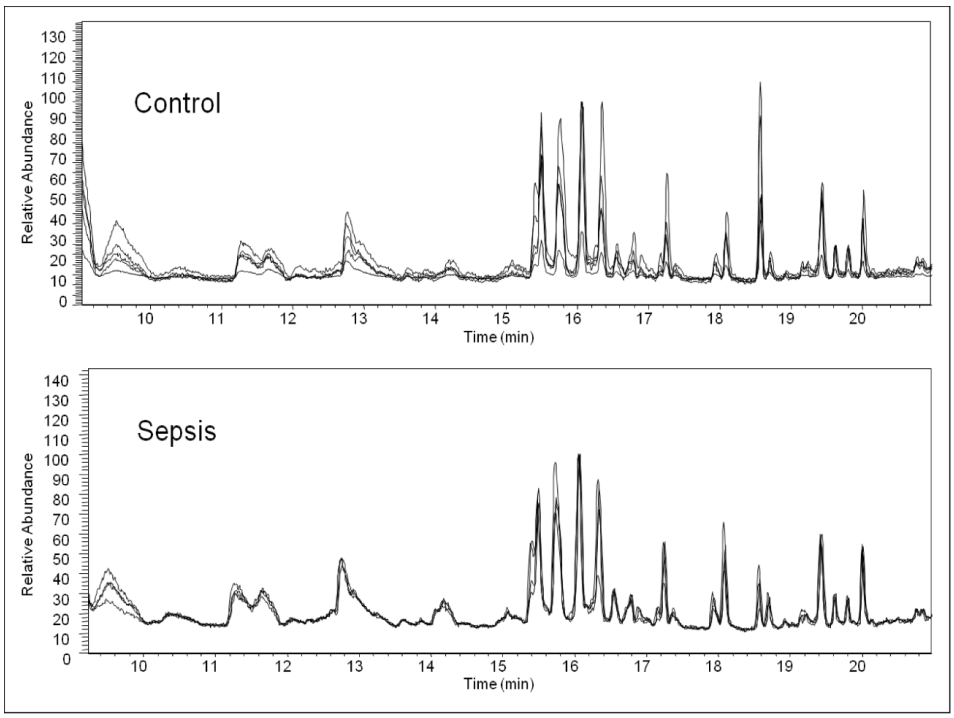


Figure 4.19 Chromatogram of breath samples from control and SEPSIS rats.

Although not all the peaks of the samples can be identified, Table 4.5 lists nineteen compounds found and identified in breath samples from diseased and healthy rats. Three compounds were identified as related to a fiber induced by LPS and one compound was identified as linked to the anesthesia. All of these were discarded for the subsequent data evaluation study. In the end, fifteen compounds were selected as possible compounds associated with sepsis and the area under the peak was calculated for each one using MzMine2(Katajamaa et al., 2006). The results of the application of PCA-LDA with rank products are shown in Table 4.5. Five compounds with a p-value less than 0.001 were chosen by the algorithm as possible compounds related with sepsis.

Figure 4.20 shows the plot resulting from the discriminant model. Bootstrap validation was implemented for a strict validation of the discrimination model. The final results obtained with bootstrap validation have an accuracy of 85% with a confidence interval between 84.6% and 85.9%. The results for sensitivity and specificity are 91% (89.7%-92.2%) and 80% (79.3%-80.7%), respectively. Again, the confidence limits were calculated to a 95%.

GC/MS measurements provided a list of compounds in the rat's breath. After the elimination of the compounds from the SPME-fiber and the anesthesia, fifteen compounds can be potentially used to separate healthy rats from treated rats. To obtain a subset of compounds related to sepsis, PCA-LDA and rank products were used as techniques that allow a maximum discrimination between classes and a ranking of compounds according to their discrimination importance. Moreover, this methodology allows us to obtain a significance level for selected compounds considered as a p-value. Thus, the p-value represents the probability of observing a compound at a certain rank, and compounds with the lowest rank are the most important in the separation. In this study we selected compounds with a p-value lower than 0.001. In the end, the first five compounds listed in Table 4.5 were selected as the most representative compounds in the discrimination between septic and healthy animals, and this could be considered a pattern correlated with sepsis. In this reduced space, a pattern recognition system provides promising rates of bootstrap validation: 85% of accuracy, 91% of specificity and 80% of sensitivity. These percentages must be understood in the light of the bootstrap validation procedure: they mean that, after 500 random selections of different sets of rats, overall 85% of the rats were well classified, and the same interpretation can be made for the specificity and sensitivity figures.

Compounds	Identification	Rank Product (p-value)
1	Cyclohexane, methyl	0.000005
2	Acetone	0.000007
3	CO2	0.00001
4	Pentafluoropropionamide	0.00003
5	Dimethylether	0.0002
6	Retention Time (18.57) Mazas(42,48,56)	0.0010
7	o-Xylene	0.0191
8	Hexane, 2,3,4-trimethyl-	0.2676
9	Octane, 4-methyl-	0.5343
10	Decane	0.6611
11	2-Propanol, 1,3-dichloro-	0.8983
12	Toluene	0.9702
13	Acetic acid	1.6955
14	Propane, 2-ethoxy-2-methyl-	2.3828
15	Benzene	4.1241
FIBER	Silanediol, dimethyl-	
	Cyclotrisiloxane, hexamethyl-	
	Cyclotetrasiloxane, octamethyl-	
ANESTHESIA	Ketanone	

Table 4.5 Identification of compounds from GC dataset

Despite the good figures achieved with GC/MS measurements, the time, cost and infrastructure needed for the sampling and measurement make it impossible to use of these instruments in a bedside setting. The IMS alternative, however, does allow for this possibility because the sampling and measurement time takes only few minutes. With respect to the IMS results, multivariate signal processing was able to detect the spectra of pure breath constituents. After a fine counteraction of external pollutants and anesthesia, and after applying pattern recognition procedures, a pattern of three components was found. Although it is not possible to identify these compounds, they can be separated into two classes, with good levels of accuracy (99.8%), specificity (99.6%) and sensitivity (99.9%) figures under bootstrap validation. It must be stressed that bootstrap validation is designed to avoid over-optimistic results. It is interesting to note that even better results are achieved by processing the full IMS spectra instead of selected molecules. In this respect, we believe that sepsis produces a general alteration in the breath pattern and not just the secretion of a single or few biomarkers.

Lack of knowledge about the metabolic pathway is therefore not a major issue, since the levels of many different VOCs are probably altered. The outstanding results obtained are encouraging and open up the prospect of performing new experiments to validate the model developed for the diagnosis of sepsis and beginning carefully controlled studies with human patients.

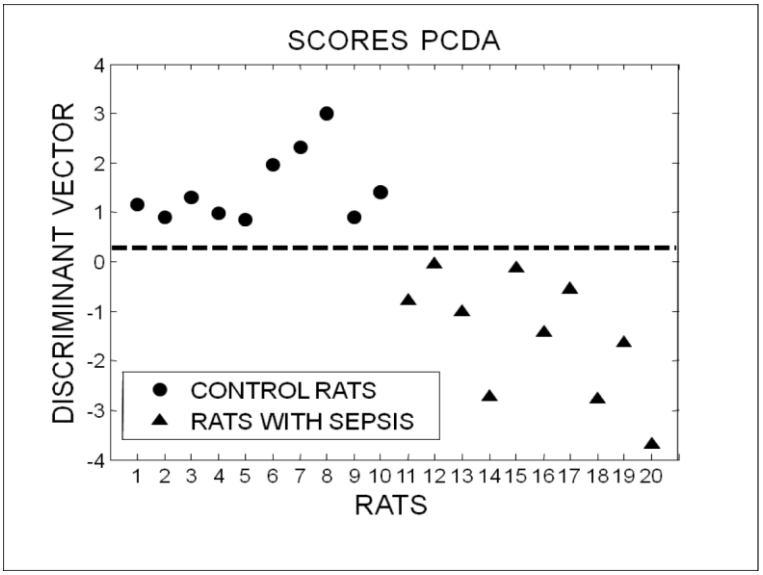


Figure 4.20 Score plot from LDA analysis

In conclusion, breath analysis with IMS has been presented as an alternative for a rapid diagnosis of sepsis. The performance of this methodology in separating a healthy rat group from a diseased rat group is excellent and provides encouraging conceptual evidence at the experimental level. Therefore, the results obtained in the present animal study warrant further clinical studies in septic patients, in order to explore the routine capability of IMS as a non-invasive point-of-care diagnostic tool.

4.5. Summary

This chapter has addressed solving qualitative problems in IMS field. Different analytical techniques have been presented for building qualitative models from real data which were measurement with IMS. Moreover, in this chapter an explanation about how to improve the signal to noise ratio of the spectra IMS has done.

The pre-processing of the IMS spectra has enclosed three main parts: noise reduction, baseline correction and peak alignment. The noise reduction has been covered from easy algorithms to more complex strategies. PCA or ICA was selected as alternative to reduce noise that was coupled to the signal, which was not eliminated using conventional filtering algorithms. The results show good performance, however the methodology is not fully optimized.

On the other hand, two different alternatives have been used for resolve a classification problem. The first approach is to use the whole spectra without taking to account the information of individual compounds. The second approach is to use blind source separation techniques to extract the pure compounds of a sample. The results of both techniques mainly differ in the goal that is going to be achieved. The goal of the first approach is build a model that discriminate wine classes but the main attention is not found which compounds are involved. The second approach apart from build a classification model, it is also interesting to choose the compounds more discriminative between classes.

Reference

Arvanitoyannis, I. S., Katsota, M. N., Psarra, E. P., Soufleros, E. H. & Kallithraka, S. 1999. Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics). *Trends in Food Science & Technology,* 10**,** 321-336.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning,* New York, NY, Springer Science and Buisness Media.

Bisson, L. F., Waterhouse, A. L., Ebeler, S. E., Walker, M. A. & Lapsley, J. T. 2002. The present and future of the international wine industry. *Nature,* 418**,** 696-699.

Cen, H. & He, Y. 2007. Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends in Food Science & Technology,* 18**,** 72-83.

Chen, H.-P., Liao, H.-J., Huang, C.-M., Wang, S.-C. & Yu, S.-N. 2010. Improving liquid chromatography-tandem mass spectrometry determinations by modifying noise frequency spectrum between two consecutive wavelet-based low-pass filtering procedures. *Journal of Chromatography A,* 1217.

Comon, P. 1994. INDEPENDENT COMPONENT ANALYSIS, A NEW CONCEPT. *Signal Processing,* 36**,** 287-314.

de Juan, A., Maeder, M., Martinez, M. & Tauler, R. 2000. Combining hard- and soft-modelling to solve kinetic problems. *Chemometrics and Intelligent Laboratory Systems,* 54**,** 123-141.

de Juan, A. & Tauler, R. 2006. Multivariate curve resolution (MCR) from 2000: Progress in concepts and applications. *Critical Reviews in Analytical Chemistry,* 36**,** 163-176.

Dixon-Woods, M., Fitzpatrick, R. & Roberts, K. 2001. Including qualitative research in systematic reviews: opportunities and problems. *Journal of Evaluation in Clinical Practice,* 7**,** 125-133.

Efron, B. 1979. 1977 RIETZ LECTURE - BOOTSTRAP METHODS - ANOTHER LOOK AT THE JACKKNIFE. *Annals of Statistics,* 7**,** 1-26.

Ewing, R. G., Atkinson, D. A., Eiceman, G. A. & Ewing, G. J. 2001. A critical review of ion mobility spectrometry for the detection of explosives and explosive related compounds. *Talanta,* 54**,** 515-529.

Felsenstein, J. 1985. CONFIDENCE-LIMITS ON PHYLOGENIES - AN APPROACH USING THE BOOTSTRAP. *Evolution,* 39**,** 783-791.

Garrido-Delgado, R., Arce, L., Guaman, A. V., Pardo, A., Marco, S. & Valcarcel, M. 2011. Direct coupling of a gas-liquid separator to an ion mobility spectrometer for the classification of different white wines using chemometrics tools. *Talanta,* 84**,** 471-479.

Guaman, A. V., Carreras, A., Calvo, D., Agudo, I., Navajas, D., Pardo, A., Marco, S. & Farre, R. 2012. Rapid detection of sepsis in rats through volatile organic compounds in breath. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences,* 881-82**,** 76-82.

Holmberg, L. 2010. Wine Fraud. *International Journal of Wine Research,* 2**,** 105–113.

iCEX, V. 2014. *El Vino en Cifras – El Vino en Cifras –Año 2014* [Online]. Available: http://www.winesfromspain.com/icex/cma/contentTypes/common/records/mostrarDocumento/?doc=4779156 [Accessed 2014 2015].

Katajamaa, M., Miettinen, J. & Oresic, M. 2006. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics,* 22**,** 634 - 636.

Krause, P. J. 1998. Learning probabilistic networks. *Knowledge Engineering Review,* 13**,** 321-351.

Miekisch, W. & Schubert, J. K. 2006. From highly sophisticated analytical techniques to life-saving diagnostics: Technical developments in breath analysis. *Trac-Trends in Analytical Chemistry,* 25**,** 665-673.

Patton, M. Q. 1999. Enhancing the quality and credibility of qualitative analysis. *Health Services Research,* 34**,** 1189-1208.

Pomareda, V., Calvo, D., Pardo, A. & Marco, S. 2010. Hard modeling Multivariate Curve Resolution using LASSO: Application to Ion Mobility Spectra. *Chemometrics and Intelligent Laboratory Systems,* 104**,** 318-332.

Razifar, P., Engler, H., Blomquist, G., Ringheim, A., Estrada, S., Langstrom, B. & Bergstrom, M. 2009. Principal component analysis with pre-normalization improves the signal-to-noise ratio and image quality in positron emission tomography studies of amyloid deposits in Alzheimer's disease. *Physics in Medicine and Biology,* 54**,** 3595-3612.

Ren, X., Yan, Z., Wang, Z. & Hu, X. 2006. Noise reduction based on ICA decomposition and wavelet transform for the extraction of motor unit action potentials. *Journal of Neuroscience Methods,* 158**,** 313-322.

Saruwatari, H., Kawamura, T., Nishikawa, T., Lee, A. & Shikano, K. 2006. Blind source separation based on a fast-convergence algorithm combining ICA and beamforming. *Ieee Transactions on Audio Speech and Language Processing,* 14**,** 666-678.

Savitzky, A. & Golay, M. J. E. 1964. SMOOTHING + DIFFERENTIATION OF DATA BY SIMPLIFIED LEAST SQUARES PROCEDURES. *Analytical Chemistry,* 36**,** 1627-&.

Smit, S., van Breemen, M. J., Hoefsloot, H. C. J., Smilde, A. K., Aerts, J. M. F. G. & de Koster, C. G. 2007. Assessing the statistical validity of proteomics based biomarkers. *Analytica Chimica Acta,* 592**,** 210-217.

Statheropoulos, M., Pappa, A., Karamertzanis, P. & Meuzelaar, H. L. C. 1999. Noise reduction of fast, repetitive GC/MS measurements using principal component analysis (PCA). *Analytica Chimica Acta,* 401**,** 35-43.

Stevenson, F. K., Ottensmeier, C. H. & Rice, J. 2010. DNA vaccines against cancer come of age. *Current Opinion in Immunology,* 22**,** 264-270.

Tomasi, G., Savorani, F. & Engelsen, S. B. 2011. icoshift: An effective tool for the alignment of chromatographic data. *Journal of Chromatography A,* 1218.

Tomasi, G., van den Berg, F. & Andersson, C. 2004. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics,* 18**,** 231-241.

Tsoukias, A. 2008. From decision theory to decision aiding methodology. *European Journal of Operational Research,* 187**,** 138-161.

Wang, M., Perera, A. & Gutierrez-Osuna, R. 2004. Principal discriminants analysis for small-sample-size problems: application to chemical sensing. *Proceedings of the IEEE Sensors 2004 (IEEE Cat. No.04CH37603)***,** 591-4 vol.2|3 vol. (xlvii+1596).

Webster, G. & Bertoletti, A. 2001. Quantity and quality of virus-specific CD8 cell response: relevance to the design of a therapeutic vaccine for chronic HBV infection. *Molecular Immunology,* 38**,** 467-473.

Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J. J., van Duijnhoven, J. P. M. & van Dorsten, F. A. 2008. Assessment of PLSDA cross validation. *Metabolomics,* 4.