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 TWO
Quantitative and Qualitative Analysis of Ion Mobility Spectrometry: from univariate to multivariate

2.1. Introduction

Both hardware development and the use of IMS have considerably increased in the last decade, as well as the amount of data to be analyzed. Initially, when the applications were based on on-off detection, the analysis was done by visual inspection or without any sophisticated software. However, novel application requires a deeper exploration and understanding of the spectra, especially when the important information seems to be hidden in complex spectra or the amounts of the compounds are really closed to signal to noise ratio. Therefore, multivariate strategies can bring useful solutions to deal with this kind of signal processing problems as well as pattern recognition and chemometrics tools, which have been used day-to-day in standard analytical techniques.

IMS dataset usually have a high dimensionality due to a single spectrum has hundreds of drift time points, and a single measurement can generate scores of spectra. Thus, a single experiment may lead many samples with a bunch of spectra. That is the reason why is really important to treat and set out rules and validation strategies for avoiding overffiting results.

This chapter stars dealing to pre-processing requirements of IMS. Then, qualitative and quantitative analysis is explained from a general perspective to a specific IMS approach. Cross validation strategies is also pointed out. Finally, limit of detection is also covered giving a perspective of different uses and formulation and how to be used in quantitative analysis of IMS.

2.2. Spectral description

A spectrum of IMS represents the ion current as a function of the drift time. An IMS usually provide several scans of a single measurement in few miliseconds (~5ms to ~20ms). However, this scans are usually very noisy. In order to reduce the noise, IMS instruments perform an average of a subset of consecutive scans completing less noisy spectra. This averaging provides slow instrument responses (around 0.5 to 2 seconds), but improves the signal to noise ratio. So, in order to estimate the dimensionality of data provided by an IMS, we can assume that an IMS instrument uses a recording spectrum speed of around a spectrum (scan) by second. Every spectrum is recorded at a sampling frequency of around 30 KHz and the drift time can be around 30 ms. Thus, the dimensionality of a single measurement that lasts N seconds, can be a matrix of N scans x M drift points, which means a high dimensional data.

Figure 2.1 shows a single measurement of IMS where only a small part of the drift time is studied, the part with useful information. The data matrix has a dimension of around 70 scans x 200 drift time points ≈ 6 ms. The high dimensionality can be easily noticed. On the right of the figure, the evolution of the whole measurement is represented in an image. Three of the 70 scans are depict on the left of figure, which represent different times of the whole measurement showing on the right of the figure. The intensity of the peaks is linked to the intensity of the image on the right of the figure. These scans depict the variety of the information from compounds that arise at different times during the whole measurement. In fact, depending on the application, the information either can be located in a unique spectrum, or a further and thorough analysis of several spectra is required to extract accurate information. For instance, in Figure 2.1 the information in spectrum 20 and spectrum 40 is quite different, thus means that analysis should be done using at least both spectra.

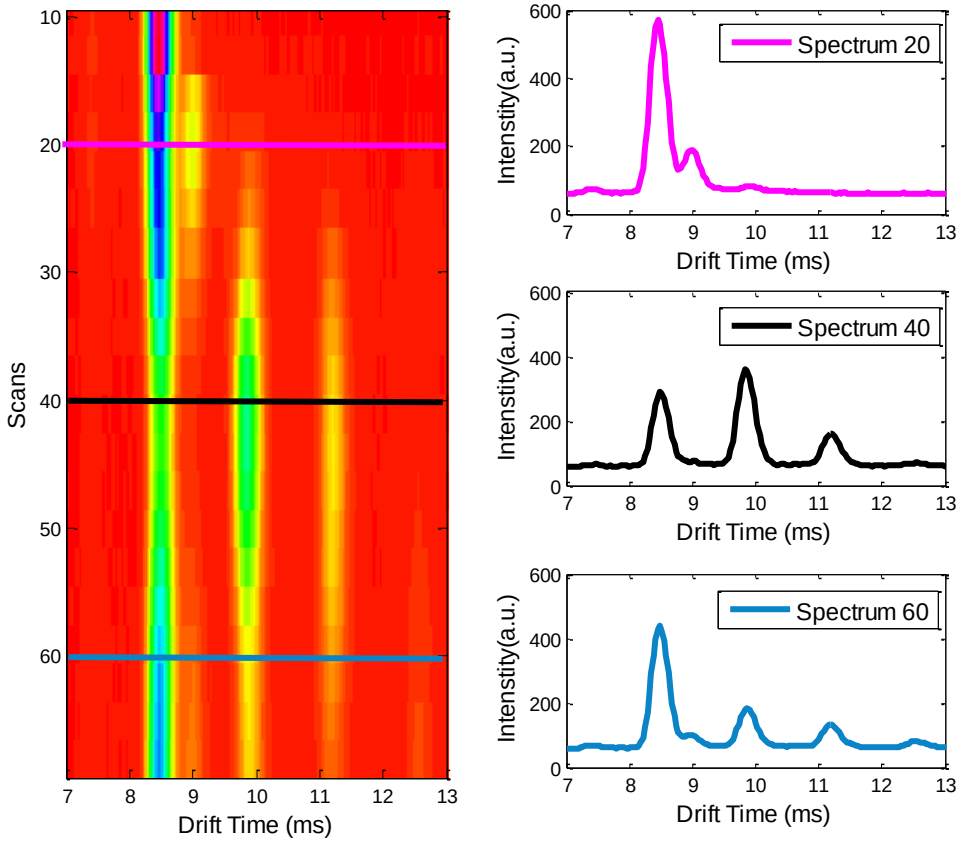

Figure 2.1 An exemplification about the high dimensionality of a single IMS measurement. Left: An image of the scans vs drift time of a single measurement. Right: Different spectra at different measurement time (scans) showing the information variability.

Clearly, the complexity of IMS data cannot be solved with simple methods and this example shows how IMS data processing must be faced with intelligent signal processing approaches. In a typical framework, some general data processing steps are needed. For a good reliability on results, to perform a preconditioning of spectra is mandatory. Then, and depending on the main purpose, qualitative or quantitative strategies must be implemented (Figure 2.2). Specific strategies for preprocessing, quantitative and qualitative IMS signal processing are presented below in sections 2.3, 2.4, and 2.1 respectively

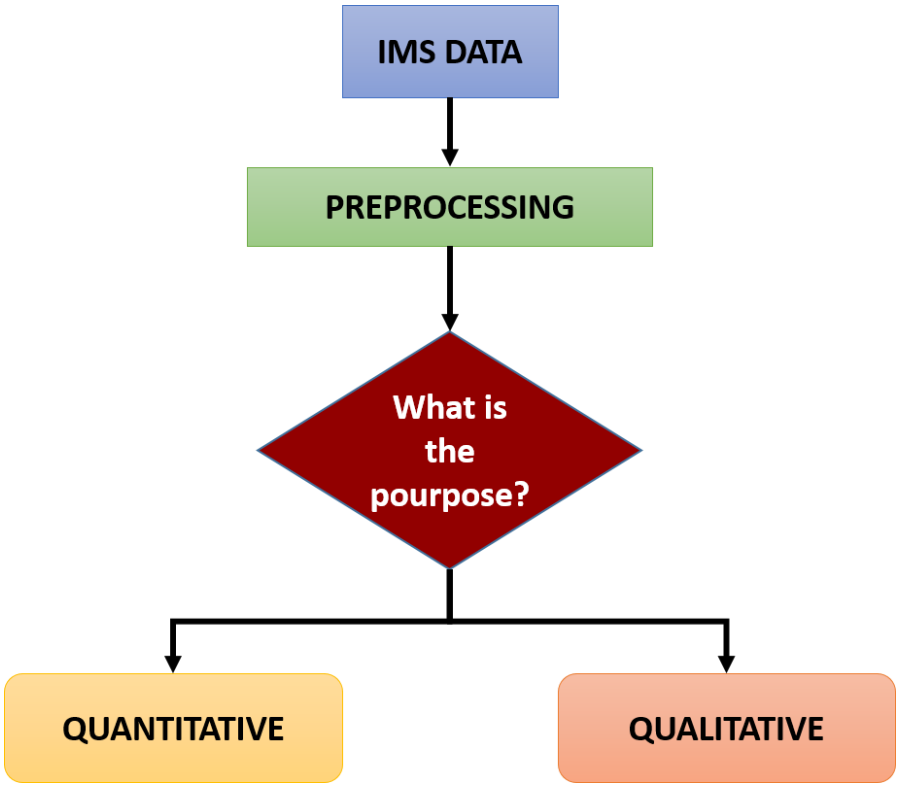

Figure 2.2 Genera block diagram

2.3. Pre-processing

Before addressing the quantitative or qualitative IMS data processing, some initial specific issues with the raw IMS spectra must be faced. These initial specific issues are solved in techniques that are known as pre-processing.

The preprocessing of IMS spectra consists of several consecutive steps for enhancing the signal to noise ratio (SNR). First issue is about noise. IMS spectra are noisy and, depending on the signal to noise ratio, this issue might affect to the detection of peaks at low concentration. As it was explained above, the spectra provided by IMS instruments are an average of several scans. Although this operation improves the signal to noise ratio, additional noise reduction operations are usually needed.

Apart from the noise, spectra present a certain baseline that need to be corrected in order to get a proper peak determination. In a single measurement, the baseline of each spectrum is quite similar between them. Thus, it is not necessary to estimate different baseline for each spectrum. In addition, there is not any difference between samples when the spectra come from the same IMS. Although, baseline correction do not implies hard algorithms strategies, performing a proper baseline correction allows avoiding a miss interpretation of the results. For instance, baseline correction should be needed when the peak area need to be estimated or a spectra analysis need to be done. Otherwise, the results might show difference due to different baseline of spectra, not due to differences between samples.

In addition, a misalignment of peaks is sometimes observed. The main source of misalignments is due to changes in temperature, pressure and humidity conditions. Effects of misalignment are greater, the larger the measurement time. The implementation of these three preprocessing steps to IMS spectra improves accuracy and precision of the final results. Below, a brief explanation of practical implementations of every step is presented. Specific results of the preprocessing used in this dissertation will be presented in chapter four.

2.3.1. Noise reduction

Noise reduction and filtering is a common problem in different fields, and several methods have been developed in order to solve it, among of them, digital filtering, smoothing, mean filtering. Some commercial spectrometers usually implement hardware or software tools to enhance the SNR. This improvement, as it was explained before, is done by averaging several scans of IMS. Nevertheless, a high frequency noise is still present in the spectra that need to be smoothed. For removing high frequency noise components from the sprectra savitzky and golay algorithm (Savitzky and Golay, 1964) is typically used. The savitzky and golay algoritm performs a local polynomial regression to determine the smoothed value for each data point. The polynomial regression (of degree k) is calculated using evenly spaced k+1 data points. The coefficients of the filter are chosen for preserving features of the data such as peak height and width, which are usually dismissed by adjacent averaging. Thus, the number of points for calculating the polynomial regression and the degree of the polynomial needs to be determined.

In practice, despite all the cautions and due to the complexity of the instruments, it is not uncommon to have a periodical low frequency noise coupled to the signal. This undesirable noise may distort peaks, causing loss of information and reduce the performance. Digital filters as savitzky and golay filter cannot deal efficiently with this kind of noise due to them might alter the shape of the peaks. A better alternative is to use techniques that aim to extract pure components from the signal and separate them. An example of this issue is depicted in Figure 2.3 in which 0.5 ppm of trimethylamine(TMA) is measured during 30 seconds while a chemical hood is turning on. It can be seen on the left how a periodical noise is coupled in the measurement. On the right it is shown the sinusoidal noise caused by the chemical hood superimposed on the TMA spectra. In order to uncouple this noise strategies based on component analysis has been used. Some of the algorithms available for this purpose are independent component analysis (Comon, 1994, Saruwatari et al., 2006) and principal component analysis (Statheropoulos et al., 1999) which have been mainly used in image, biomedical and voice preprocessing (Razifar et al., 2009, Ren et al., 2006, Saruwatari et al., 2006, Takahashi et al., 2009). This new approach applied to IMS spectra are going to be explained in chapter four.

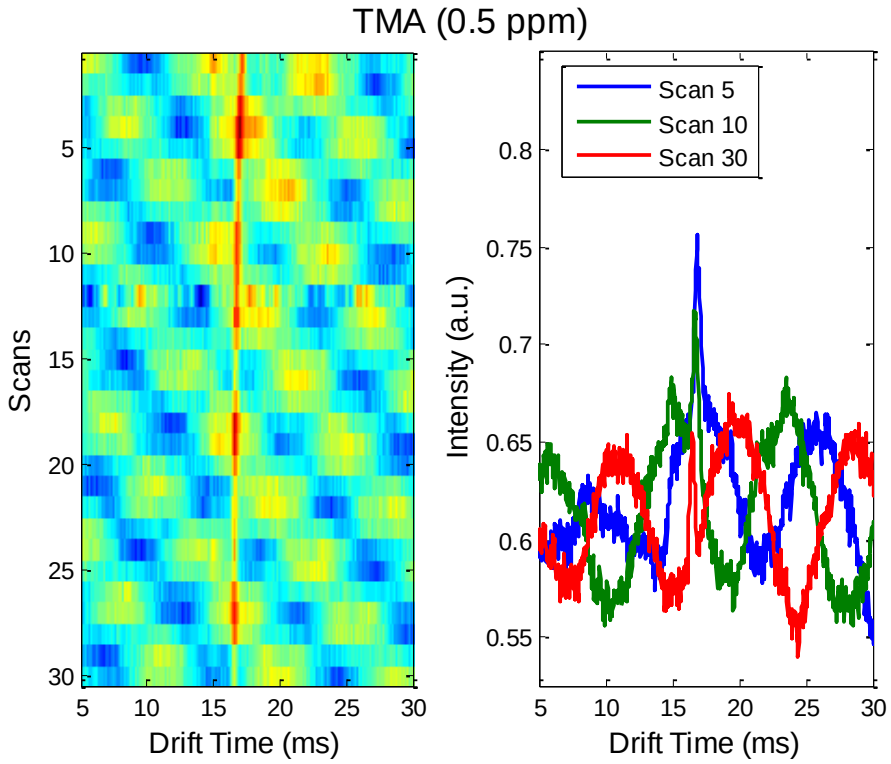

Figure 2.3 Low frequency noise coupled to IMS spectra.

2.3.2. Baseline Removal

The approach for baseline removal cannot be thinking as a general solution, it should be considered one solution for each spectrometer. Note, peak area or height of the peak of interest is regularly used in IMS analysis, thus baseline need to be removed before any further analysis. One simple option is to connect a straight line between the endings of the peak of interest, but is not the optimal choice.

In Figure 2.1, the peaks are really well defined if it is compared to the baseline, thus the baseline estimation might not be difficult. However, set up a correct baseline sometimes is not easy to perform as it can be seen in Figure 2.3. For this purpose, there are some general approximations that can also fit in the IMS field. The easiest one is to select intervals where no information (peaks) appears, which can be at the beginning and the end of each spectrum. Then a polynomial of a certain order can be fitted using these intervals (Pomareda et al., 2010). Certainly, it can occur that the baseline does not fully fit in the real baseline, so a certain amount of error should be expected.

Iterative approaches have been developed for fitting and removing baseline from a signal using some initial specifications. One of them is to set up iteratively automatic thresholds, which could be a polynomial, and cut out the signal above of them. A criterion indicates if the baseline fulfill all the requirements and the final baseline is determined and removed from the signal(Gan et al., 2006). Another approach is based on asymmetric least squares in combination with a smooth factor for modeling the baseline without any prior information of the signal (Peng et al., 2010). There are two parameters as initial information that need to be set up, and are related with the smooth shape of the baseline and asymmetric algorithm behavior of the baseline. Zheng et al (Zhang et al., 2010) proposes an adaptive iteratively penalized least squares (airPLS) method in which the

baseline is gradually approximate changing weights of the penalized least squared algorithm(Good and Gaskins, 1971). The weights are adjusted adaptively by means of sum square errors between the fitted baseline and original signal and the smooth of the baseline by a parameter provided by the user.

2.3.3. Misalignments

The last problem from a signal preprocessing perspective is to solve misalignment. The misalignment can be produced by changes in experimental condition or instrument variability. As it has been explained above, temperature and pressure are completely related with the formation of clusters and, in turn, shift of peaks may come from changes in these parameters. These changes are clearly observed during an experiment when different samples are analyzed. In this case, peaks from sample to sample are misalignment and consequently performing any kind of analysis becomes unfeasible. The alignment has to be done in such a way there is no peak distortion.

A first alignment can be addressed performing a correction of the mobility coefficient, which is inversely proportional to drift time, using Eq. 3.2. This first alignment has the main advantage that does not require any reference vector. However, this first order correction is sometimes not enough, especially when complex mixture are been analyzed and producing different effects such as overlapping peaks.

Additional techniques/algorithms for peak alignment, coming from gas chromatography analysis or NMR analysis (Brown et al., 1996, Trygg et al., 2007), can be used in IMS spectra. Nevertheless, an indispensable requirement of this kind of algorithms is to have a reference peak or pattern. The fact of having a reference can be quite challenging when non-target experiments are analyzed. For those IMS spectrometers that are able to generate a RIP peak (those with a radioactive ionization source, for example) or a dopant peak, these peaks can be used as reference for alignment purposes and a first order alignment can be done in the time basis direction.

There are other solutions when the spectra are more complex than performing a shift in the time basis direction. One solution is the use of warping algorithms. Warping algorithms are a well-established solution which goal is to align a sample data towards a reference pattern by allowing limited changes using constraints in segments lengths on the sample vector (Tomasi et al., 2004). The most popular and may be used are known as correlation optimized warping (COW) and dynamic time warping (DTW) (Tomasi et al., 2004). The computational cost of both algorithms could be quite expensive and the peaks can be slightly altered as consequence of the changes in time domain. Hence, in the last years a faster version called *icoshift (interval correlation optimized shifting algorithm) (Tomasi et al., 2011)* has been developed. Icoshift use an efficient Fast Fourier Transformation for performing an interval shift of the sample that requires less computational time, but using the same concept of warping algorithms. The reference vector can be any spectrum and there is a certain grade of flexibility in the use of intervals. At the end, the most important in the IMS is to not distort the spectrum as consequence of the alignment, in order to not lose their physical meaning.

2.4. Qualitative Analysis

The IMS data is a matrix of huge dimension of *N* scans per *M* drift points. Moreover, the information can be placed in different scans and many times the evolution during the measurement can add information in the analysis. Actually, the information of the IMS data is focused in the peaks present in the spectra. Consequently, the *M* points might be reduced to a subset of *K* peaks. Nonetheless, it can be overlapped peaks that make difficult to extract peak information. One solution can be to analyze the whole spectrum information instead of extract peak information. On the other hand, the *N* scans of a single measurement have to be taken as unique information and not as having *N* measurements. These consideration are really important especially to avoid the called "curse of dimensionality" (Duda et al., 2000, Raudys and Jain, 1991). The curse of dimensionality is well known in pattern recognition field which means get results that are over fitting for the fact of having fewer samples than features or sensors. Certainly, if this consideration is not taking into account in IMS data, the results can suffer from over fitting.

Qualitative analysis provides a certain grade of discrimination between different categorical classes. This qualitative analysis can be done building a model and using a classification algorithm to get a classification rate. It is important to highlight the high dimensionality of a single measurement of IMS as can be seen in Figure 2.1. Consequently, it is really necessary to perform a dimensional reduction to avoid over fitting results as it was explained above. Besides dimensionaly reduction, it is important to do a suitable validation. The validation should also consider that a single sample has scores of scans and not misinterpret scans with samples.

2.4.1. Feature extraction, feature selection and dimensionality reduction

In order to avoid over fitting results, the dimension of the data must be reduced in such a way to ensure to have enough samples for getting realistic results. According to Raudys and Jain there is an exponential relationship between samples and features (Raudys and Jain, 1991). Thus implies that dataset from IMS, which provides hundreds of drift points, will be require thousands of samples for obtaining reliable results.

There are two main approaches for dimensional reduction called feature extraction and feature selection. Feature extraction seeks to create a new subspace from a projection of the original space. Feature selection attempt to select a subset of original features which maximize a figure of merit such as classification rate. In IMS feature selection can be only possible if the features are information of peaks such as height or width of peaks. Other alternative is use feature selection after a pre deconvolution of the spectra; hence pure components can be obtained instead of whole spectra. On the other hand, feature extraction can be used over IMS spectra and get a new projection with fewer dimensions of the original IMS data.

A common search strategy attempt to extract enlightening information based on an expert knowledge such as area or height from the peak or peaks of interest in IMS. In this case the whole spectra get reduced to a subset of peaks, instead of having hundreds of drift points. The effectiveness of getting peaks information will depend on the resolution of the peaks (if they are overlapped or not) and the informatics tools that help in this demanding task. Univariate techniques used as qualitative analysis can be done

using statistical tests or ploting one feature against the others. Therefore, the most significant and representative features can be keep it.

Other alternative is to generate a subspace where the most relevant information should be keep it. Nonetheless, this new subspace tends to be no lineal and is necessary to generate systematic lineal transformations. In order to avoid theses repetitive transformation, linear approximations has been developed, among of them the most popular is principal component analysis (PCA) (Bishop, 2006). PCA is defined as the orthogonal projection of the data onto a lower dimensional linear subspace called principal components (PCs), which are linear independent, such that the variance of the projected data is maximized (Hotelling, 1933)**.** The mathematical formulation is given by Eq. 2.1, where *X* is the dataset of N samples by F features, *T* represents the samples projected onto the new subspace *D* (*D << F)* called scores and the *W* are the loadings which measure the feature importance in this new subspace in accounting for the variability. The number of PCs can be determined either by visual inspection or using cross validation methods.

$$X = TW^T \qquad \text{Eq. 2.1}$$

The PCA is an unsupervised method because the labels of the sample do not take part in the compression, and it is usually to be hoped that the new PCs describe really well the original data without losing important information. In this sense, supervised methods seek to compress the data but considering the labels of the sample. One of the most popular is the linear discriminant analysis or fisher discriminant analysis (LDA) (Bishop, 2006, Eisenbeis and Avery, 1972, Friedman, 1989, Sugiyama, 2007) which compresses the data into a lower dimension equal to the number of classes minus one. LDA seeks to perform dimensionality reduction while preserving as much of the class discriminatory information as possible. It finds a new subspace where examples of the same class are projected very close to each other but, at the same time the projected means of every class are farther apart as possible. Despite of the effect of dimensionality reduction, LDA is sensitive to over fitting especially when the number of samples is smaller than the number of features. A strategy of combining PCA and LDA is commonly used in order to take profit of advantages of both. Normally a first unsupervised PCA step is applied on the dataset in order to reduce dimensionality and prevent the overfiting, and then a second LDA step is applied for improving the performance. (PCDA) (Wang et al., 2004, Garrido-Delgado et al., 2011a).

Other approach for combining dimensional reduction techniques and discriminant analysis is PLS-DA (Westerhuis et al., 2008, Barker and Rayens, 2003). PLS-DA use partial least squares (PLS), explained below in section 2.5.3, using binary labels, therefore PLS-DA seeks to predict the response of the data using the explanatory variables from the PLS. PLS-DA can be used as either classifier or model using the score projection for applying a different classifier. There other combinations using PLS such as orthogonal PLS (OPLS) (Trygg and Wold, 2002) which tries to enhance the discrimination between classes removing systematic variations of the data that are not correlated with the categorical classes, or a recent variation called OPLS-DA (Bylesjo et al., 2006) which attempt to enhance the variance between groups in a single dimension or latent variable, and separate the within group variance into orthogonal latent variables.

Feature selection seeks to select a subset of features that maximize an objective function. One alternative is to use a variable selection approach such as interval PLS (IPLS) (Norgaard et al., 2000) which main idea is to select a subset of features or intervals that maximizes the prediction between classes comparing to the use of the whole data. Another way is to use bilinear models such as multivariate curve resolution (MCR) which is deeply explained bellow in section 2.6. There are also more conventional algorithms such as sequential feature selection (SFS) or sequential backward selection (SFS)(Narendra and Fukunaga, 1977). The objective is to sequentially add or remove features while the objective function is maximized. There are other versions, in which the algorithm goes forward and backward seeking the best option. This algorithms is knowing as floating search algorithms (Pudil et al., 1994). Finally, genetic algorithms (Leardi et al., 1992) is another alternative for feature selection which is biological inspired. In any case, these techniques build models and their predictive power needs to be obtained and it is done using classifiers.

2.4.2. Classifiers

At the end, the goal of a qualitative analysis is to get a quantitative value regarding a separation between substances. The classifiers create decision regions using decision boundaries or surfaces to divide an input space (Bishop, 2006). The input space is made up by the training data or model and giving a test vector x the idea is to assign it to one of the K discrete classes C_k of the training data. The final decision can be either assigns each sample to one and only one class or to provide a probability value for each class. The decision boundaries can be obtained from linear and nonlinear models.

Among of the linear models, LDA is the most known and was explained above, with the variations for dimensional reduction. Another alternative is logistic regression(Bishop, 2006) which is a probabilistic statistical classification model. The boundaries can be nonlinear, hence the classifiers, among of them support vector machine (SVM) (Burges, 1998, Suykens and Vandewalle, 1999, Chang and Lin, 2011) using kernel functions, decision trees such as random forest (Svetnik et al., 2003, Strobl et al., 2007)**,** multilayer perceptrons (Ruck et al., 1990, Huang and Huang, 1991, Pal and Mitra, 1992)**,** and k nearest neighborhood (kNN) (Henley and Hand, 1996) or in the fuzzy version fuzzy-kNN (Kuske et al., 2005) which provide a grade of membership.

In Table 2.1 shows a brief summary about the main requirements of classifiers.

	Overfitting	Computational requirements	Setting up Parameters	Decision boundaries	Membership
LDA	Yes	Low	None	Linear	Yes
Logistic Regression	Yes	Very Low	Regression parameters	Linear	Yes
SVM	No	Medium /High	Kernel	Linear/NonLinear	Yes
kNN	No	Medium	k neighbors	Non Linear	No
Random forest	No	High	Number of variables to start the decision tree	Non-Linear	Yes
Multilayer perceptrons	Yes	High	Number of neurons	Non-Linear	Yes

Table 2.1 Summary about requirements of classifiers.

Actually, choosing a classifier will be depend on how the data is distributed and how many samples has the model, otherwise overfitting results may lead. Surely, the classification rate is very important in order to demonstrate if a model is good or not. Apart from the fact of using a proper classifier, validate the model are essential. Thus, performing a precise validation methodology is required.

2.4.3. Qualitative analysis used in IMS.

In IMS field, qualitative analyses have been carried out from screening strategies to discrimination of classes. Most of them have been based on identification of a peak or set of peaks of interest in target experiments by visual inspection, and then a discussion about the behavior of the compound in the experiment (Bell et al., 1995, Fernandez-Maestre et al., 2010, Verkouteren and Staymates, 2011, Zamora et al., 2011, Dunn et al., 2012). Other groups perform calibration curves using either height or area of the peak or peaks of interest, which sometimes are done using commercial software developed by the manufacturer, and calculate the limit of detection of them, see more detail in section 2.1 and 2.7.

There are just few papers that use the information of the whole spectra for discrimination aims. For instance, Snyder et al (Snyder et al., 1995) use PCA for cluster compounds of different families, Garrido et al (Garrido-Delgado et al., 2011a) use PCDA for discriminate wine of different origin denominations, also in another publications Garrido et al (Garrido-Delgado et al., 2012)use the same technique for differentiate olive oils. The detection of fungal infestations of wood has been studied by Huebert et al (Huebert et al., 2011) in which PCA was used. A different approach was applied in the study of breath analysis in rats where MCR following by sequential feature selection was used for building a model that discriminate between control and inflammatory disease(Guaman et al., 2012). This methodology is relative new in the IMS field, which results are explained in chapter four.

2.5. Quantitative Analysis

Quantitative analysis is referred to build quantitative model in order to be able to predict an amount of a substance or substances of interest. Quantitative analysis is also known as calibration process. Calibration has a widespread use in different science such as chemistry, physics, medicine, engineering , and instrumental measurements (Thomas, 1994). The main idea is to seek a relationship, which could be a linear relationship, between a signal response and a target variable; thereby calibration relates, correlates or model a measured response based on any physical or chemical properties of the sample (Brown et al., 1996, Kalivas, 2005, Olivieri et al., 2006, Danzer et al., 1998). Actually, calibration or quantitative prediction models have become a main objective of signal processing for chemical sensing in the last years (Marco and Gutierrez-Galvez, 2012). In the same context both instrumental and computational advances have allowed to develop numerous calibration methods which are able to also manage new challenges in emergent applications like food chemistry, biomedical applications and environmental industry.

2.5.1. Univariate Calibration Model

Univariate calibration is conceived as the analysis of a single feature. Usually, peak height or area of the analyte of interest is extracted and used for the quantitative analysis. For instance, in the case of having a linear regression, a calibration model can be mathematical modeled as Eq. 2.2 which express a relationship between a target value X (*I*(calibration samples) *x 1* analyte of interest) (e.g., concentration, level of absorbance of a dilute solution, etc) and the response of the instrument Y (*Ix1*)- i.e height of a peak of interest. Thus, the model is formed by β_0 which represents the intercept, and β which is the slope of calibration curve. It is noteworthy that linear regression is the simple case, but it is possible to use curve of polynomials of higher orders when the assumption of linearity is no longer suitable.

$$Y = \beta_0 + \beta X \qquad \text{Eq. 2.2}$$

The calibration model is obtained using a set of measurements of the analyte of interest, which are also known as training, within a desired region. The following step is known as prediction in which new set of measurements is projected into the calibration model in order to predict its associated concentration level. Indeed, the main idea is to determine β which is usually obtained using least squares and the mathematical expression is given by Eq. 2.3. The new prediction (*x)* of a new set of measurement *(y)* is given by Eq. 2.4 (Hastie et al., 2003).

$$\beta = (X^T X)^{-1} X^T Y \qquad \text{Eq. 2.3}$$

$$x = y/\beta \qquad \text{Eq. 2.4}$$

In a simple scenario, when data can be obtained by a single selective sensor, univariate methods are fully feasible and therefore Eq. 2.3, and Eq. 2.4 can be directly applied. In a real scenario, such as biological samples, to achieve these requests are quite difficult because interferences cannot be fully eliminated, and real samples are composed by hundreds or thousands compounds. Thus means that is necessary to acquire pure standards for all responding species and to do this is not completely manageable. In addition to that, it is well studied that univariate calibration in presence of interference

give severely biased predictions (Olivieri et al., 2006). Despite of the univariate calibration issues, it is still accepted and deeply used in different fields among of them pharmaceutical(Sasic et al., 2015) ,biology(Dimitrov et al., 2015) , meteorology(Feldmann et al., 2015, Kiesel et al., 2015).

2.5.2. Univariate Calibration applied to IMS

Univariate calibration has been deeply accepted and used in the field of IMS owing to its simplicity, but its use might be questionable owing to IMS intrinsic complexity. The idea under univaritae paradigma is to find where the peak of interest is located and calculate either peak area or peak height whose final value is used to build a calibration curve.

Note that univariate calibration can be really dificult in real samples, since it is expected that a sample include several compounds and a more complex spectra for extracting the area or height of the peaks.There are several publicashion where univariate calibration is used, i.e. (Armenta and Blanco, 2012b, Garrido-Delgado et al., 2011b, Marcus et al., 2012, Jafari et al., 2007, Satoh et al., 2015, Atmanene et al., 2012, Morsa et al., 2011). In those cases, a univariate measurement were calculated in order to determine a quantitative information, even though the measurements were not done with pure analytes. Even though the results shown are quite acceptable, the main problem is an expert needed to certificate that one peak belonged to a particular analyte. In some cases, as it was mention before, when the experiment was performed in a control situation, it could be posible assure previous statement. However, it is difficult to guaranty that samples from biological sources in non target studies do not have interfernts or a high backround. Thus means that there will be other interferents in the sample with the same reduced mobility or a background with more intensities values that hide the compounds of interest. Therefore, results from univarite calibration must be used when there is a guaranteee of do not having any other interfent in the sample.

Figures of Merit for Univariate Calibration

The univariate calibration use can be linked to its easy implementation. Thus, organizations such as IUPAC have developed a compendium of accepted rules in order to be stricter when univariate calibration is used. (Danzer et al., 1998). These rules are, also known as figures of merit, oriented towards to ensure a proper use and validation of calibration model. In addition, the figures of merit provide a good way to compare results and models.

In general, the figures of merit most used are sensitivity, selectivity, signal to noise ratio and root mean square error of prediction and cross validation.

- Sensitivity (R^2), in the case of univariate calibration of a given analyte is defined as the slope of the calibration curve.
- Selectivity is a ratio between sensitivity of the analyte (s_a) and sensitivity of a particular interference (s_i) and it is given by $\xi_{i,a} = s_a / s_i$.
- Signal to noise ratio (SNR) is the ratio of the useful analytical signal to the background noise.
- Root mean square error of prediction (RMSEP) of M measurements is given by Eq. 2.5.

$$\mathrm{RMSEP} = \sqrt{\frac{1}{\mathrm{M}}\sum_{\mathrm{i}}^{\mathrm{M}}(\mathrm{X_i} - \mathrm{x_i})^2} \qquad \text{Eq. 2.5}$$

This figures of merit are going to be used for compare the results between different models, especially between multivariate calibration models.

2.5.3. Multivariate Calibration Model

In the last decade the use of multivariate calibration model has increased and become popular as an analytical tool. There are several reasons why multivarite calibration has been so well accepted, among of them the fact of being able to include multiple features and several measurements into the model –i.e. sensors response, spectral information and not single response as univariate case. Thus allow to improve the precision and applicability of quantitative analysis (Forina et al., 1998, Thomas and Haaland, 1990). Indeed, multivariate calibration allows the study of analyte concentration in mixtures becoming the most growing area in chemometrics (Brown et al., 1996). Additionally, analyzing multivariate signal enable to compensate contributions of interferences in the predictions samples. The use of multivariate calibration have been succesfully and extended used in a variety among of them, spectroscopic, biomedical, food chemistry, industrial and clinical chemistry (Escandar et al., 2006, Forina et al., 1998, Forina et al., 2007, Thomas, 1994).

A model is built using the dimension of the original data. In our case, the model is going to be applied directed to IMS dataset. Thus, lest have a matrix, for instance IMS matrix *Y (IxJ)*, which contain the spectra of *I* calibration samples at *J* measurements at different drift time, and a vector of concentrations *X (IxN)* for each *I* calibration sample from *N* analytes. Thus, a calibration model can be described by Eq. 2.6 if the model fullfill the superposition propierties. In this case (Eq. 2.6), *B (JxN)* are the sensitivities or regression vector for each analyte at the *J* measurements of drift time and *E (IxJ)* represent the error. Therefore, the *N* concentrations are obtained by fitting the regressors matrix *B* to the spectrum of the prediction $\hat{y}$ whose fitting is usually done using ordinary least-squares (OLS) Eq. 2.3. This approach is also known as classical least-squares (CLS) and it shown in Eq. 2.7. However, it is important to take into account its limitations such as CLS requires the spectra information of all contributing analytes to be measured or estimated from mixture spectra (Olivieri et al., 2006) thereby all interferences must be known.

$$Y = x_1 b_1 + \cdots + x_N b_N + E = XB + E \qquad \text{Eq. 2.6}$$

$$\hat{x} = B^{+}\hat{y} \qquad \text{Eq. 2.7}$$

Since, knowing all interferences in advance is impractical or even unreal, a different approach called inverse model have been developed to solve these problems by treating analyte concentrations as a function of spectral values (Kalivas, 2005). Thus, the calibration model is given by Eq. 2.8 and Eq. 2.9. As long as the number of samples *I* is larger than the number of variables *J* in the matrix *Y*, OLS can be used to get a regressor matrix *B*. Nevertheless, this rarely happens so that alternative methods called inverse least squares (ILS) have been used to reduce dimensionality to a lower space *F<I*. For

instance, artificial neural networks (Swierenga et al., 2000) or genetic algorithms (Leardi, 2001) have been used to select a best subset of variables for implementing ILS. Hence, *B* will be determined by

$$X = YB + E \quad \text{Eq. 2.8}$$

$$B = XY^{+} \quad \text{Eq. 2.9}$$

Another approach tries to perform linear combinations of the original variables instead of selecting only a set of them. This linear combinations are also called scores in which the idea is to move from *J>I* to a new reduced space *F<I.* The leading approaches of the so called "full-spectrum" method are principal component regression (PCR) (Massy, 1965) and partial least squares (PLS) (Geladi and Kowalski, 1986, Haaland and Thomas, 1988, Helland, 1990). Both PCR and PLS addresses to replace the original variables of Y^{+} by the called scores, but both of them differ on how to estimate these scores.

On one hand, PCR takes advantage of dimensional reduction using principal component analysis (PCA) thereby $Y = TP^{T} + E$ in which *T* are the scores of *Y* and *P* are their respectively loadings. *T* and *P* both collect only a limited percentage of the total variance of *Y*. The number of principal components (PCs) is selected using a rank-determination method such as cross validation. At the end, PCR estimates Y^{+} in the base of *T* and *P*, and regressor matrix *B* will be determine as Eq. 2.10.

$$B = Y^{+}X = [P(T^{T}T)^{-1}T^{T}]X \quad \text{Eq. 2.10}$$

On the other hand, PLS seeks to maximize the covariance between *Y* and *X*, therefore the new factors called latent variables (LV) capture variance in *Y* but also achieve correlation with *X*. PLS decomposes *Y* and *X* matrices into the form Eq. 2.11 and Eq. 2.12, where *T* and U are the score matrices (dimensions *I x F*) of the *F* extracted latent variables. The matrix *P (JxF)* and matrix *Q (NxF)* are the loadings and *E* and *F* represent the residuals.

$$Y = TP^{T} + E \quad \text{Eq. 2.11}$$

$$X = UQ^{T} + F \quad \text{Eq. 2.12}$$

According to non-iterative partial least squares of NIPALS method (Geladi and Kowalski, 1986) PLS provides two types of loadings matrices *W* (weight loadings), which are needed to keep the scores *T* orthogonal, and *P* that explains the maximum covariance between instrumental signals and information of constituents - i.e. concentrations. The scores *T* are given by Eq. 2.13 and the final matrix *B* as Eq. 2.14

The LV in PLS should be selected using methods of cross-validation to get an objective determination of the importance of prediction errors in the final calibration model.

$$T = YW(P^{T}W)^{-1} \quad \text{Eq. 2.13}$$

$$B = [W(P^{T}W)^{-1}(T^{T}T)^{-1}T]X = Y^{+}X \quad \text{Eq. 2.14}$$

There are slightly variations of PLS when the calibration problem is no longer linear such as poly-PLS (Wold et al., 1989, Wold, 1992). Poly-PLS attempts to establish a non-linear relationship between score *U* and *T* whose fit is a polynomial of a certain order (nth) and is given by Eq. 2.15.

$$u_i = b_{0i} + b_{1i}t_i + \cdots + b_{ni}t_i^n \qquad \text{Eq. 2.15}$$

Besides the typical techniques of PLS and PCR, there are less used approaches based on machine learning strategies which results are positively accurate such as support vector regression (SVR) (Desai et al., 2006, Smola and Scholkopf, 2004, Cherkassky and Ma, 2004), or random forests applied to regression (Svetnik et al., 2003).
In our thesis PLS are explored as multivariate calibration technique applied to IMS dataset. The multivariate techniques are compared with the use of univariate calibration models which results are deeply explained in chapter five.

2.5.4. Multivariate Calibration applied to IMS

As it has been discussed in the chapter one, the ionization process in IMS, and hence the response of the instrument to the analytes, is affected by the nature of ion species. The ion competitiveness, the collisions driven atmospheric pressure and the clustering of ions can provide an unwanted IMS response due to variation on temperature or humidity, and matrix effects during ionization(Eiceman and Karpas, 2005). Univariate techniques are not able to handle with the complexity of the process, explained before, and they have poor performance in complex scenarios. On the other hand, multivariate techniques are able to use the entire information of the whole spectra featuring better the IMS response in real scenarios.

Actually, it was not long ago that the IMS use was focused on detecting the presence or absence of an analyte. Thus, the effort was concentrated on the study of dopants for achieving an appreciable grade of selectivity of the studied analyte/es. This usually happen because of the complexity in the dynamics of ion formation of the IMS. Even though the use of the dopant is still in use and its usefulness is guaranteed, new applications have been emerged where the use of dopant is not an option. This new applications usually claims to find new compounds from complex samples, such as biological samples, in order to use IMS as detection or monitoring instrument from a research perspective. These new application have brought new problems among of them, complex spectra. Therefore, the use of multivariate calibration techniques has been tabled as a perfectly good solution giving reliable results. In this sense, (Zheng et al., 1996, Fraga et al., 2009, Zamora and Blanco, 2012) are the only ones who made a contribution in the field of IMS using multivariate techniques.

Zheng (Zheng et al., 1996) proposed the use of cascade correlation networks (CCN) (Fahlman and Lebiere, 1991) to perform both qualitative and quantitative analysis of a set of 15 volatile organic compounds that were measured independently. In addition, PLS was implemented to evaluate the performance of neural networks. They found a good fit when CNN is used even in cases in which peaks are overlapped with the RIP of the instrument and better results comparing to PLS results. However, number of latent variables was fixed and not selected using cross validation; thereby the PLS results could be too pessimistic.
Fraga (Fraga et al., 2009) compared the common univariate technique versus multivariate techniques such as partial least squares and principal component regression. Explosives compounds 2,4,6-trinitramine (RDX), 2,4,6-trinitrotoluene(TNT) and 1,3,5,7-tetrazocie (HMX) were analyzed by temperature step desorption (TSD) coupled to IMS. TSD was used to partially resolve mixture components before ion

mobility spectrometry. The measurements consisted of mixtures at different concentrations of the three explosive compounds. In the univariate case the area of the peak of each compound were calculated and in the multivariate case the whole spectra were used to perform PLS and PCR. The authors concluded that the results are remarkably better with PLS or PCR both than univariate case. They also said that in terms of accuracy the variability in peak area calculation does not allow to have sufficiently precise quantitative results even the substances were well resolved. Moreover, multivariate calibration is able to capture the effects like competition among the reagents among the different IMS signals while peak-area calibration cannot do it.

(Zamora and Blanco, 2012) used IMS to determine active principal ingredients (APIs) present at low concentrations in pharmaceuticals. The main objective was to test multivariate techniques to solve overlapping problems present in APIs. They measured two concecutives peaks which were strongly overlapped and through multivariate curve resolution and a second derivative algorithm both APIs were extracted given qualitative results. To get quantitative results PLS were performed to the whole plasmagrams (spectra) using mixtures and pure APIs. They conclude that APIs are able to be quantitative determined using PLS and also overcomes the problems of their mutual interferences, primarily present in mixtures.

In this thesis, an deeply exploration of multivariate technique is done using novel biological application, where the main difficulties are pointed out togheter with its respective methodological solutions. Also, new altarnetive of calibration is presented where a deconvolution technique is used. All of these results are presented in chapter five.

Figures of Merit for Multivariate Calibration

The figures of merit for multivariate calibration are extensively explained in (Olivieri et al., 2006). Nevertheless, an overview about inverse model will be described.

- Sensitivity is determined by $1/\|B\|$. The sensitivity is also known as slope of the calibration curve and is denoted by R^2. A calibration model works better as the sensitivity is close to 1 or 100%.
- Root mean square error based on cross-validation (RMSCV) provides a measure of how well the prediction is estimated on the basis of the average analyte level (Thomas, 1994). RMSECV is a robust prediction measurement, due to the error is calculated over a several partitions of the original data. A calibration model is enough accurate and reliable when RMSEC is minimum or closer to an ideal zero.

$$RMSCV = \sqrt{\frac{1}{n}\sum_{i=1}^{m}\left(\hat{X} - X\right)^2} \qquad \text{Eq. 2.16}$$

, where $\hat{X}$ is the prediction concentration using a multivariate model and *m* represents the number of predictions samples.

2.6. Self-modeling mixture analysis techniques

Real samples are composed by both pure standards and complex mixtures. In addition, the content of real samples are usually formed by thousands of compounds, but there are just few of them that are fully correlated with an specific application. The compounds that are not linked with the desired application can be considered as background, and some of them migth be pollutatants of the sample. Therefore, the fact of extracting just the informative compounds from the whole sample is not a simple procedure.

One of the goals of multivariate techniques is to extract a pattern from the whole compounds which are totally realated with the application. In this sense, there are some techniques that are able to extract the basic information from the sample, among of them the most popular are PCA and PLS. PCA and PLS are techniques that perform mathematical transformations building a new space where the samples are projected. Even though, the projection of the data in this new space are able to separate or discriminate different clases, identifying the features (compounds) which are responsible for this separation results to be complex. The features are projected in this new sapace in a matrix called loadings. These loadings do not provide a physical interpretation of the compounds, but the featuares are wheighted for each principal component. For these reasons, PLS and PCA are deeply used as exploratory techniques, but has to be carefully taken as ultimate model without any validation.

Moreover, depending upon the instrumental technique, the final data matrix could be a multi-component system of 2 dimensions that is also known as two-way data or bilinear data. For example, a single measurement using ion mobility spectrometry generates a data matrix in which each row is a multivariate spectrum and each column represents a drift time of the ionized molecules. This kind of data should follow Beer-Lambert law that means i.e. there are a linear dependency between pure compound and its concentration (Stlouis and Hill, 1990). Thus means that changes in concentration will lead changes of the pure compound peaks. Therefore, self-modeling techniques attempt to perform a bilinear decomposition or a mathematical deconvolution to extract pure compounds present in the sample without any a priori information about them (Jiang and Ozaki, 2002, Windig and Guilment, 1991, de Juan and Tauler, 2006). This kind of techniques are also known as Blind Source Separation (BSS) (Cichocki et al., 2008) in the field of signal processing and their usefulness has been tested in different scenarios both within and outside the chemical scope (Comon, 1994, Esteban et al., 2000, Berry et al., 2007, Zhang et al., 2013, Kim et al., 2007, Osten and Kowalski, 1984, de Juan et al., 2000).

Bilinear models can be applied provided exist a linear relationship between the samples that are analyzed by the spectrometer and the underlying contributions of the pure components. It means that each concentration belongs to a specific compound. In this sense, the mathematical model is represented as Eq. 2.17

$$D = c_a s_a^T + c_b s_b^T = CS^T + E \qquad \text{Eq. 2.17}$$

In the specific case of IMS data, matrix *D* is formed by *I* spectrum x *J* drift time points. The resulting concentration profile *C* will have a dimension of *I x N* pure compounds and *S* represents the spectra profile for each pure variable. In this case c_a, s_a, c_b,and s_b are the pure compounds present in the sample.There are different approaches for resolving bilinear model (Eq. 2.17) but it could be divided in two groups: non-iterative approaches (unique resolution techniques) and iterative approaches (rational resolution methodologies) (de Juan and Tauler, 2006, Jiang and Ozaki, 2002).

Non-iterative procedure explores a subspace of dataset both selective and zero-concentration regions which are usually built using local rank analysis. It is based on performing repetitive PCA analysis within small regions of the data matrix which can be done using fixed or variable windows. This final process gets as results information from the pure compounds that evolve along the dataset. According to (Manne, 1995), the best scenarios to use these methods are when the overlapping is in the same direction of concentration. However, the fact of selecting a unique region makes this kind of methodologies dependent on an expert, who needs to have a priori knowledge of the regions for assuring accurate estimations. Amongst the most popular algorithms can be found Evolving Factor Analysis (EFA) (Maeder, 1987) usually used as local rank analysis technique, Window Factor Analysis (WFA) (Malinowski, 1992)(Malinowski, 1992)(Malinowski, 1992)(Malinowski, 1992), Heuristic Evolving Latent Projections (HELP) (Kvalheim and Liang, 1992, Liang and Kvalheim, 1993). WFA and HELP use a preselected region of raw data at zero-concentration in order to find main variations between them, and recover the pure variable that is uncorrelated with information at zero-concentration. Nowadays, Fixed Size Moving Window Evolving Factor Analysis (Keller and Massart, 1991) is the most popular algorithm in this category because it uses different size of windows to detect the presence of minor species in data matrix.

Moreover, there are other techniques that aims to evaluate the purity of the most representative compounds and concentrations of the spectral variables. These methods assume that the pure variables must be selective and vary during the whole experiment, and can be used for either determining the number of pure compounds present in a sample or extracting the concentration and spectra profile. The pioneer and the most used method is called simple-to-use interactive self-modeling mixture analysis (SIMPLISMA) (Windig and Guilment, 1991). Windig define a pure variable in such a way it has only contribution from one of the components in the sample. The basic idea is to estimate C of Eq. 2.17 with pure variables which is defined by

$$p_{k,j} = w_{k,j} \times \frac{\sigma_j}{\mu_j + \alpha} \qquad \text{Eq. 2.18}$$

for which $p_{k,j}$ is the purity spectrum *k* of variable *j*. The $w_{k,j}$ is a determinant based weight function which measure the linear independence of the j_{th} candidate variable with respect to the previous pure variable candidate, and α is a constant value that prevent dividing by zero in case of no signal content and bias the purity slightly towards variables with higher intensity (Windig et al., 2005). Then, the spectra profile for each k pure variable can be calculated using Eq. 2.18 as Eq. 2.19, and the predicted concentration profile can be determined using Eq. 2.20

$$s^*_{k,j} = w_{k,j}\sigma_j \qquad \text{Eq. 2.19}$$

$$C^* = DS^*\left(S^{*T}S^*\right)^{-1} \qquad \text{Eq. 2.20}$$

In the last years, some modifications have been proposed by Windig such as the use of the second derivative data for a better selection of the pure compounds especially in cases when the data present baseline problems and there is higher overlapping between the components (Windig et al., 2002). SMAC (stepwise maximum angle calculations) (Windig et al., 2005) is an angle measurement of the pure variables previously selected with each other, i.e. using SIMPLISMA, in order to reject pure variables that represent several components -in other words variables that are not completely pure. In addition, SMAC is a helpful tool in cases when a dataset have both narrow and broad spectral features and SIMPLISMA are not able to accurately extract the correct pure variables. Iterative approaches yields feasible solutions in complex chemical process or data without any a priori knowledge, thus adding knowledge to the models is known as adding constraints for refining the pure variables.

Constraints, which can be either mathematical or physical information, are an active field of study, and new kinds of constraints are being developed nowadays (deJuan et al., 1997, de Juan et al., 2000). These methods requires initial estimations of *C* or S^T, which can be obtained by applying methods such as SIMPLISMA or PCA. Among of the set of constraints, non-negative, unimodality, closure are the most common, specially applied above all in the spectrometry field. Another approach is to fit hard/rigid models that define our data in a mathematical expression, in general the shape of the concentration or/and spectra profile. Figure 2.4 illustrates the most common constraints applied in multivariate curve resolution approaches.

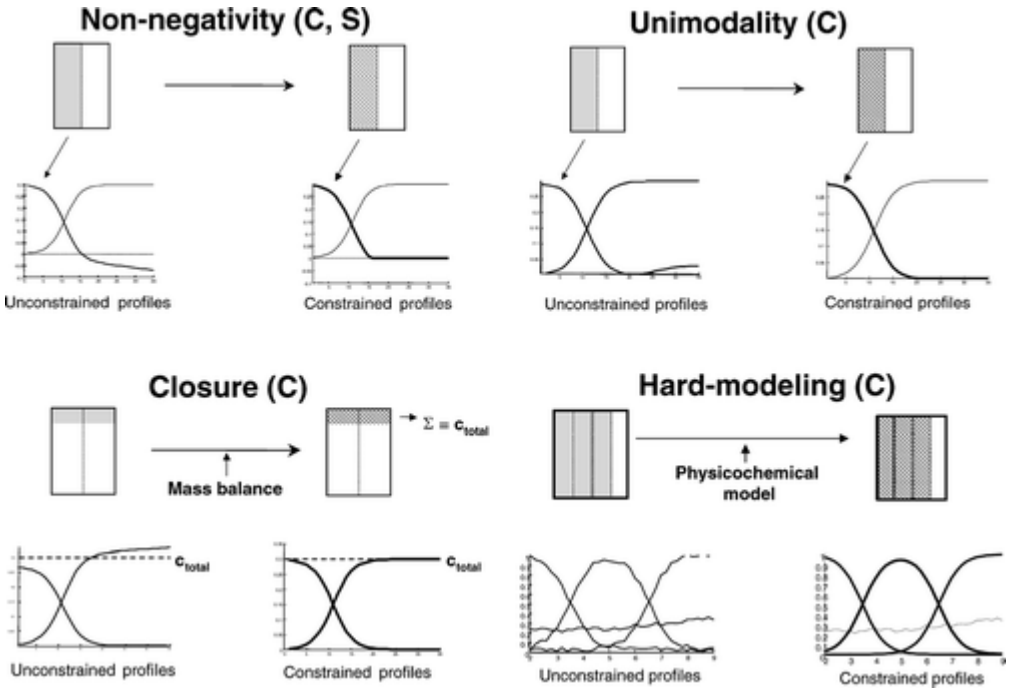

Figure 2.4 Common constraints used in iterative MCR approaches (de Juan and Tauler, 2006)

One of the main drawbacks of these algorithms is that can produce a set of different feasible solutions, which is also called rational and/or intensity ambiguities, and this uncertainties have been the center of attention of many researches (Gemperline, 1999, Vosough et al., 2006, Leger and Wentzell, 2002, Abdollahi and Sajjadi, 2010, Golshan et al., 2012). Nevertheless, one possible solution is to set minimum and maximum boundaries to each pure compound, one at a time, for the feasible solutions bands of resolved profiles (Tauler, 2001, Rajko and Istvan, 2005, Abdollahi and Tauler, 2011, Abdollahi and Sajjadi, 2010, Sawall et al., 2012, Beyramysoltan et al., 2013).

The most popular algorithm within iterative methods is Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) (Tauler et al., 1993). In principal, ALS attempts to minimize the error in Eq. 2.17 over *C* for fixed *S* and over *S* for fixed *C*. An initialize profiles either *C* or *S* is needed before to start with ALS and it usually is done either using PCA or SIMPLISMA. For example, if the initial estimation is the spectra profile *S*, the least square equation to estimate *C* is:

$$C = DS(S^T S)^{-1} = D(S^T)^+ \qquad \text{Eq. 2.21}$$

$$S = D(C^T C)^{-1} C^T = DC^+ \qquad \text{Eq. 2.22}$$

where $(S^T)^+$ is the pseudoinverse matrix of S^T when it is full rank. To estimate the spectra profile the equation is Eq. 2.21. Both Eq. 2.21 and Eq. 2.22 are iteratively estimated until an optimal solution is obtained and/or a convergence criterion is fulfilled. As previously stated, constraints need to be applied in the iterative process of ALS, so that ambiguities can be avoided. A diagram block is shown in Figure 2.5 about MCR-ALS procedure. Apart from MCR-ALS, there are other algorithms related to iterative methodologies such as Iterative Target Transformation Factor Analysis (Gemperline, 1984, Vandeginste et al., 1985).

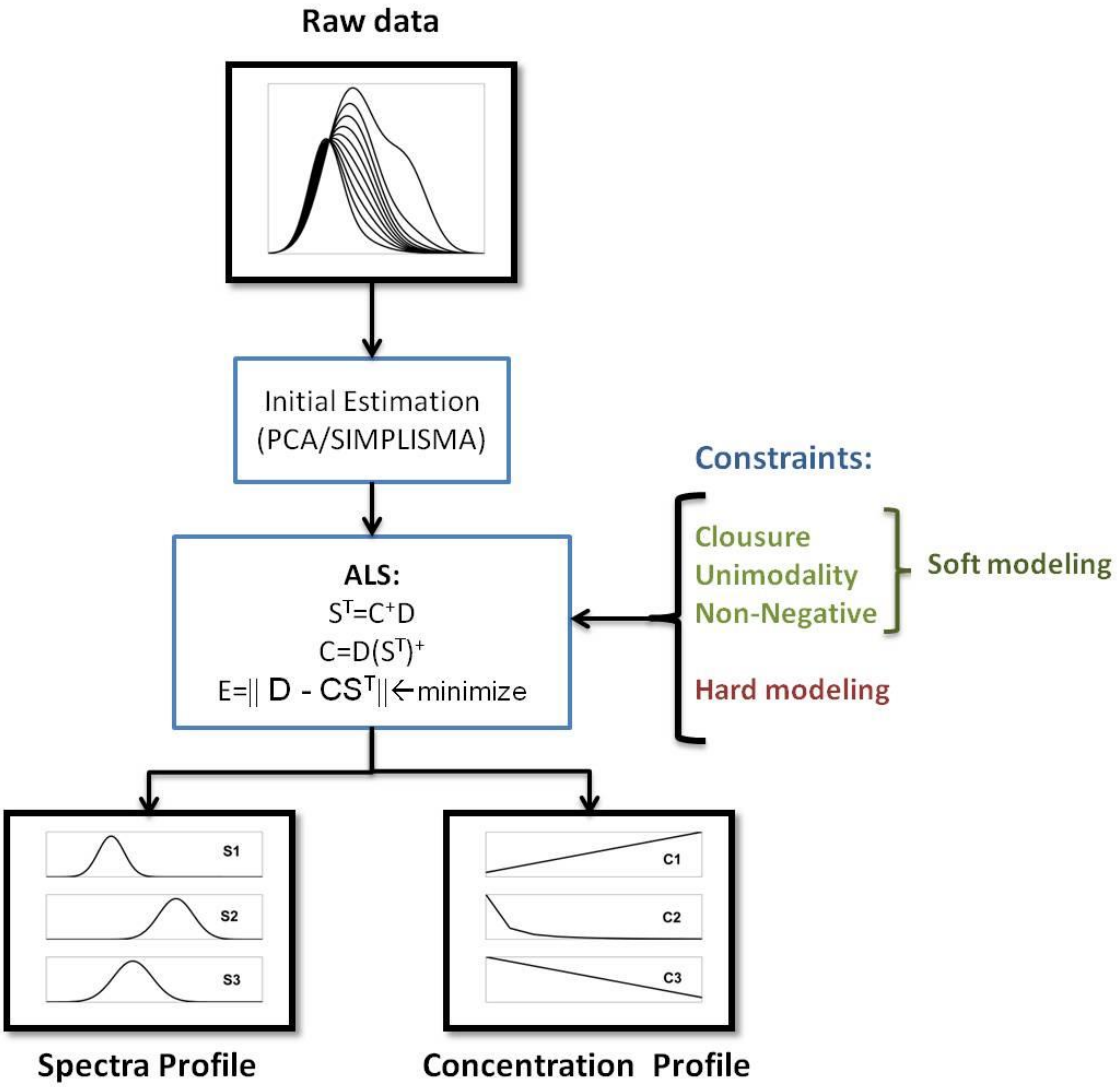

Figure 2.5 Block diagram of MCR-ALS approach

2.6.1. Self-modeling mixture analysis applied to IMS

One of the first studies aiming to extract pure compounds from IMS spectra was developed by (Bell et al., 1995). In this work Bell et. al. use deconvolution techniques to extract the compound of interest from a noisy background. They propose to use fitting curves as models of peaks in IMS like Gaussian, Lorentzian, and Error functions. In addition, they want to know the effects in deconvolution under a certain experimental conditions such as drift tube temperature, concentration of the analyte, moisture and component mixtures. Besides concluding that error function was not as good approximation for deconvolution as the others, they found that concentration and the physic-chemical properties of the compounds are totally correlated with peak shape, which is a predominant factor for deconvolution. Special care must be taken in cases of mixture of substances, in which proton affinity and interaction among them play an important role. Another important factor was the amount of moisture to guarantee the success of deconvolution. Nevertheless, they admitted that some deconvolution parameters must be optimized for every application, making less attractive than other approaches.

(Buxton and Harrington, 2001) was one of the pioneer of introducing SIMPLISMA as techniques for deconvoluting peaks in IMS. In this case, they used compounds related to explosives as pure standards and interferents at different temperatures. The results showed that pure compounds were correctly extracted, even though some of them were missed by visual inspection. Furthermore, they said the successful of SIMPLISMA was due to the changes both concentration and temperature occurred at the measuring time. Pure compounds in presence of interferents were also extracted using SIMPLISMA, but they noted that concentration level of pure standard was significantly higher than interferents which make a less complicate scenario. Later on (Chen and Harrington, 2003, Chen and Harrington, 2001, Cao et al., 2005) introduced a compression method based on wavelet transformation in order to make feasible real time processing using SIMPLISMA. Data used to test the algorithms were warfare agent stimulants (Cao et al., 2005) and narcotics (Chen and Harrington, 2003) both at high concentration. In both cases, they used wavelet compression in both retention time and measuring time direction, and then either SIMPLISMA or ALS was used to extract pure compounds. They claimed the fact of using wavelet was not worse if it is contrasted against original algorithm with the advantage of having less dimensions. Then (Harrington and Chen, 2004) developed a new constraint to be applied into ALS called equilibrium of compounds in IMS which uses a formulation of theoretical evolution of monomer and dimer when concentration increases. They concluded that this method could be a good model to estimate vapor concentrations in non-multivariate scenarios.

(Lu et al., 2009) presented an alternative approach closer to the biological scenario. They applied ALS to the data of samples that were measured with solid phase extraction (SPE)-ion mobility spectrometry. In this case, they tried to get cocaine metabolites in urine samples using SPE to increase instrumental sensitivity and selectivity, and subsequently they applied ALS to obtain spectra and concentration profile from the metabolites of interest. They finally concluded that coupling SPE-IMS is a good alternative for screening drugs in urine, and the use of ALS offers good performance compared with traditional screening methods.

In the last years, just few contributions have emerged linked with this topic. Pomareda et al. (Pomareda et al., 2010) proposed to impose a hard modeling into MCR algorithm. The objective was to create a very dense Gaussian model (it could be other peak model) and then fitting it using a least absolute shrinkage and selection operator, thus the algorithm is called MCR-LASSO and its diagram block is shown in Figure 2.6. This method assumes that the peak shape can be modeled by a Gaussian and the weight of the Gaussian is calculated using the resolution of the instrument.. Under this approach, a regularization parameter (λ) should be adjusted by cross-validation. The algorithm was tested using synthetic and real data and its performance were compared to SIMPLISMA . The results showed a better fit when MCR-LASSO was applied, specially when high level of noise was introduced in the system. The algorithm was able to model wider asymmetric peaks and provide a better resolution in real experiments. The final pure spectra and concentration profiles had less noise becoming more interpretable than SIMPLISMA.

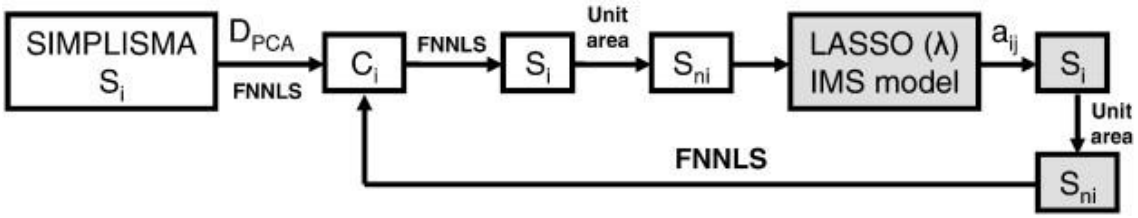

Figure 2.6 MCR-LASSO algorithm's block diagram.(Pomareda et al., 2010)

(Khayamian et al., 2012) aims to build a cubic data matrix (*IxJxK*) using IMS data in which *I* is a whole measurement of IMS of *J* spectra *x K* drift time points, and use tucker 3 model to decompose the pure compounds of the sample into three modes and three-way core array. Each mode represents the evolution in time of the experiments, the concentration behavior and the spectra information, respectively and the three-way core array represent the capture variance of the model. Although, the main idea is similar to the methodologies explained above, the parameters setting up and the subsequent interpretation it is not as simple as other methodologies. The author's also propose to use the final concentration profile to build a calibration curve as quantitative procedure.

(Zamora and Blanco, 2012) proposes to use MCR-ALS and a second derivative deconvolution technique to deconvolute peaks that have a huge overlap between them. According to the authors both techniques had a good performance and allowed the correct identification of the pure compounds. Besides identification techniques, they used PLS as quantitative procedure using the whole information of IMS. In turn, (Armenta and Blanco, 2012a) aims to find a set of pharmaceutical substances that might be present in the air of a pharmaceutical workplace using an IMS which is coupled to a thermal desorption unit. Thus, the final data matrix represents a spectrum of IMS for each retention time. Furthermore, they identified a set of compounds that has a closer reduced mobility between them using MCR-ALS. The first estimation was performed using evolving factor analysis to enhance the selectivity and sensitivity and then it is refined using MCR-ALS. The results were satisfactory and the concentration profile gave a quantitative value about the concentration of the compounds of interest. Finally, both instrumental technique and the deconvolution methodology were tested in simulated workplace exposure and in a real situation, and the final results were contrasted against a reference method given comparable results. Thus, confirms the potential of IMS as monitoring techniques in an occupational pharmaceutical assessments.

In this thesis, MCR-ALS and MCR-Lasso is combined with other multivariate techniques such as PLS and sequential floating feature selection. In the first case, the concentration matrix is going to be used for transforming the semiquantitative response into a purely quantitative response using PLS. This is going to be applied in non-linear datasets that comes from samples of wine, and rat's breath. The second case, the concentration profile from the compounds are the input of a feature selection algorithm. Thus, a subset of the compounds will be selected by the algorithm which maximize the separation between the classes. Both strategies are going to be used to solve quantitative and qualitative problems.

2.7. LIMIT OF DETECTION

Limit of detection (LOD) is one of the most used figure of merit in analytical chemistry and many other fields. LOD has the capability for quantify a trace element or molecule in chemical and biological matrices, and measure the power of an analytical procedure to deal with traces amount of analyte. However, there is a lack of agreement about the best definition and interpretation of limit of detection, and it could imply an order of magnitude in its calculation.

$$L_D = \mu_0 + k_D\sigma_0 \qquad \text{Eq. 2.23}$$

The basic formulation of LOD (Eq. 2.23) can be expressed in terms of population mean (μ_0) and the population standard deviation of the blank of the signal (σ_0). The standard deviation is multiplied by a constant value k_D which represents a probability that the blank does not exceed the LOD value (α or type I error), where k_D has been typically received a value of 3.

A wide study about hypothesis based on detection limit theory was presented by Currie in 1968 (Currie, 1968), and it has rapidly become one of the most popular formulation of LOD. This study was based on define LOD in basis of the null hypothesis (H_0), which tests if the analyte is not present. Currie establishes two concepts: a critical level (Lc), which is the maximum acceptable value for avoid detecting a substance when it is not; and detection limit (Ld) which is the smallest true signal that will be reliable detected against to make a false conclusion that a blank observation is a real signal.

The mathematical formulation of critical level and detection limit are given by Eq. 2.24 and Eq. 2. 25

$$L_C = k_\alpha\sigma_0 \qquad \text{Eq. 2.24}$$

$$L_D = L_C + k_\beta\sigma_D \qquad \text{Eq. 2. 25}$$

where k_α and k_β correspond to probabilities of commit error type I and type II according to central *t* distribution. Error type II (β) is committed when a compound is not detected when it should be detected. σ_D is the standard deviation which stands for measurements that contains the analyte at the level of limit of detection.

In 1995, Currie (Currie, 1995) addressed how to calculate k_α and k_β coefficients, and confidence limits of LOD. Later on Hubaux and Vos(Hubaux and Vos, 1970) introduced a new method in calculation of LOD based on confidence limits for linear calibration curves. This approach attempts to calculate upper and lower confidence limit in a regression curve which represents a confidence band where predicted values must fall inside of the region.

Moreover, the width of this confidence band will depend on knowledge of data and uncertainty about the true position of the calibration line. In order to estimate LOD the authors define a value y_c that represent the lowest measurable signal, which is similar to critical level on Currie estimation (Currie, 1968). The lowest concentration to be predicted is represented by x_D and it is the projection of y_c on the lower confidence limit. Therefore, x_D represent the detection limit (L_D) that Currie(Currie, 1968) had previously estimated.

Figure 2.7 depicts the concept under LOD calculation, and it is important remark that x_D and y_c are related to α and β probabilities (type I and II errors). This approach assumes data are homoscedastic that means equal variance through the whole experiment or constant variance. Thus, if data are not homoscedastic, the confidence band will be narrower or wider around calibration curve, and LOD calculation will be affected by these data distribution as it can be noticed in Figure 2.7.

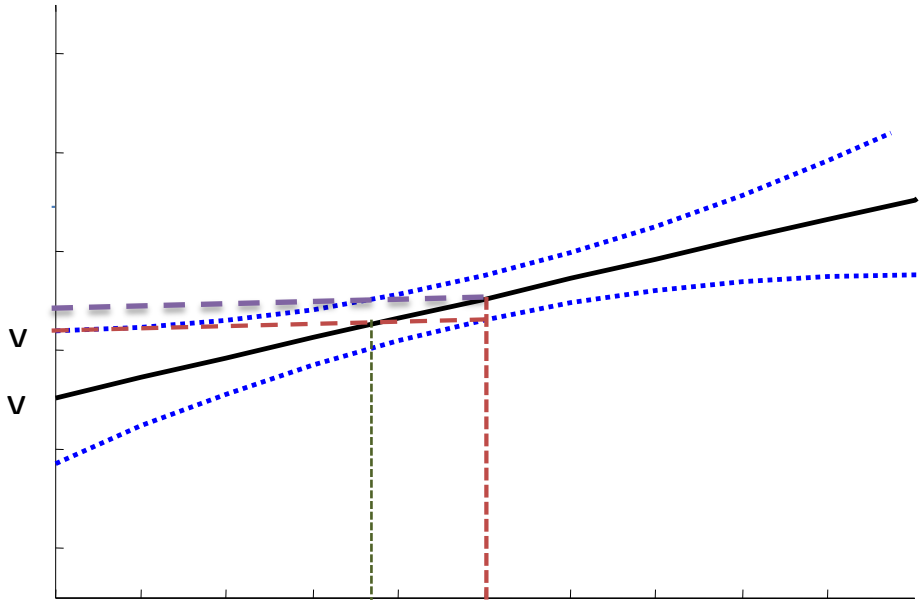

Figure 2.7 Calibration curve with upper and lower confidence limits. y_C decision limit, x_C critical level and x_D the detection limit

The mathematical development goes through determine y_c (Eq. 2.26)

$$y_C = Y_0 + s \cdot A \qquad \text{Eq. 2.26}$$

where Y_0 represents the intercept of the curve, s is given by equation Eq. 2.27 and is the residual variance of the theoretical and real signals, and A it is a factor related to confidence interval of regression curve equation Eq. 2.28.

$$s^2 = \frac{\sum(y_i - Y_i)^2}{N-2} \qquad \text{Eq. 2.27}$$

$$A = (t_{1-\alpha} + t_{1-\beta})\sqrt{1 + \frac{1}{N} + \frac{\hat{x}^2}{\sum(x_i - \hat{x})^2}} \qquad \text{Eq. 2.28}$$

N is the number of samples in the calibration curve, y_i is the measured signal at concentration x_i and Y_i is the theoretical signal based on regression line.

Despite of the fact that the theoretical approximations of Currie(Currie, 1968) and Hubaux and Vos(Hubaux and Vos, 1970) are quite similar, in practice there are one order of magnitude different in LOD calculation. Actually, Voigtman (Voigtman, 2008a) performed a set of simulation to test both approaches and concluded that Hubaux and Vos approximation is overly pessimistic resulting in false negatives rates (failing to decide that an analyte is present when it is) and the final LOD is twice of Currie(Currie,

1968) estimation. In addition, he claims that the fundamental problem with Hubaux and Vos method (Hubaux and Vos, 1970) is that the confidence prediction levels do not fit in the detection process because the authors are considering limit of detection on the content (x_D) and not in the response of the instrument (y_D). Hence, the LOD should have been lower than they report. On the other hand, Long and Winefordner (Long and Winefordner, 1983) were settled that the curve approximation is only valid when the major source of variation is present in the blank, otherwise the LOD could give artificially low values. Moreover, Analytical Methods Committee of Royal Society of Chemistry (1987) claims that Hubaux and Vos (Hubaux and Vos, 1970) approach have to be only used when the calibration curve is linear near the detection limit region.

In 1997 International Union of Pure and Applied Chemistry (IUPAC) published an overview in order to standardize the procedures for determining the limits of detection and quantification (Mocak et al., 1997). The LOD was defined as "limit of detection must reflect the value of the true signal (related to some non-zero analyte concentration) which is significantly different from the blank signal value". They emphasized about having a finite number of measurement both in blank measurements and calibration points in real experiments and how the original Eq. 2.24 formulation have to change taking to account a finite number of measurements. The first consideration is how to estimate k_D. Thus, the new formulation that they proposed is an upgrading of Eq. 2.24, and it is shown in Eq. 2.29

$$L_D = \bar{y}_0 + t(v_0, \alpha)(1 + 1/n_0)^{1/2} s_0 \qquad \text{Eq. 2.29}$$

where, $\bar{y}_0$ and s_0 are the sample characteristics of both mean and standard deviation of blank samples, $t(v_0, \alpha)$ critical value of t-distribution with v_0 degrees of freedom which is calculated as number of blank samples (n_0) minus one., and the term $(1 + 1/n_0)^{1/2}$ is a correction of the uncertainties of the determination of $\bar{y}_0$ and s_0. Mocak (Mocak et al., 1997) also alluded to the consideration of estimate LOD using information of calibration curve Eq. 2.26. It was called as Upper limit approach to the calibration curve and is given as,

$$L_D = t(v, \alpha)\frac{s^2}{b}A \qquad \text{Eq. 2.30}$$

where, s^2 and A is given by respectively Eq. 2.27 and Eq. 2.28, b is the slope of the calibration curve and $t(v, \alpha)$ is a value of t distribution with v degrees of freedom calculated as number of samples plus one minus number of the regression parameter for instance 2 when is a line.

Obviously, Mocak (Mocak et al., 1997) and (Currie, 1968) determinations differs from each other and the major discrepancy corresponds to whether include or not error of the second kind which at the end means doing error type I and II comparable. Furthermore, Mocak claimed that the differences goes beyond to be a theoretical disagreement but it is a philosophical understanding of what it is detection limit; and he considered that LOD determination is considered by to be different from a blank signal such as critical level Eq. 2.24 of Currie(Currie, 1968). Indeed, he mentioned if there are several measurements, the error of the second kind shifts to the error of the first kind resulting in consider only α probabilities.

Mocak (Mocak et al., 1997)also gives some considerations when LOD is calculated such as:

- Blank samples need to be included in the regression procedure.
- It is recommendable to have the same number of replications for each concentration in the calibration curve.
- The range of concentrations to calculate LOD has to be as close to the expected value to avoid over-optimistic value.

Later on in 2007 Voigtman performed an extensive monte carlo studies in order to test four different scenerios like (i) to compare Currie (Currie, 1968) and Hubaux and Vos (Hubaux and Vos, 1970) methodology (Voigtman, 2008a)(ii) having non-central distributions (Voigtman, 2008b)(iii) what happen when heteroscedastic noise is present in samples (Voigtman, 2008c), and (iv) exploring error of the second kind in LOD(Voigtman, 2008d).

The first part (Voigtman, 2008a) was discussing above, but in summary he found that Hubaux and Vos method is negative biased so the LOD is too low. The second part (Voigtman, 2008b) concluded there is not any relevant effect when data are not central distributed, therefore there is not any need to apply critical values of the noncentrality parameter of the noncentral t distribution. In the third part, Voigtman explored the value of use weighted least squares when heteroscedastic noise is present in a sample. In order to explore this scenario, he assumed that the noise precision model is known. Nevertheless, in reality it is far from be true. He also tested a Currie (Currie, 1997) formulation dealing with heteroscedastic noise. After 10000 simulations, Voigtman found that Currie formulation can accurately calculate when heterostedastic noise is present as long as weights of WLS have been accurately estimated. Finally, in part 4, he studied the effect of whether include or not Type 2 errors. The main advantage is the simplicity of the formulation when 50% Type 2 error rates are assumed. Since, having this percentage of error sometimes has not a big impact; the author suggested the use of Currie schema Eq. 2.22 and Eq. 2.23. He also mentioned the effect of having a fixed k_D as it is mentioned by Mocak (Mocak et al., 1997) and he could observe that when k_D varies, LOD changes.

The main difference between the approach by Mocak and Currie is to add Type 2 error into the formulation that results in a philosophical understanding of what it is limit of detection. On the other hand, the other method, where the parameters of the calibration curve are used for estimate LOD, ((Hubaux and Vos, 1970), Mocak(Mocak et al., 1997) has brought some controversy about its use. Nevertheless, this method brings a solution when predicted calibration points give as result negative values whose response do not have any physical meaning. Perhaps this approach is not useful from univariate standpoint, but it could be valuable when there are not enough blanks in the experiments and the calibration curve can add useful information in the LOD calculation.

The need of multivariate detection limit (MDL) is given by the fact of solving the presence of other analytes apart from the analyte of interest in the sample to be analyzed. Therefore, LOD does not only depend on the mathematical model, but depends on influence of the other analyte over sample. The pioneer of introducing MDL was Lorber (Lorber, 1986) starting from the analysis of net analyte signal. However, it had not been

used in the field of chemistry until Bauer (Bauer et al., 1991) improve Lorber method (Lorber, 1986) and used in error propagation theories. Later on, in 1997 Faber and Kowalski (Faber and Kowalski, 1997) made a good approximation of Currie univariate method (Currie, 1968) estimating L_C Eq. 2.24 and L_D Eq. 2. 25 from calculating error propagation(Bauer et al., 1991) of predicted concentration of ordinary least square solution Eq. 2.6, and Eq. 2.7

Another approach of MDL was developed by Boqué (Boque and Rius, 1996) who used criteria discussed by Hubaux and Vos(Hubaux and Vos, 1970) in univariate calibration and turned into multivariate calibration point of view. In this approach MDL is calculated taking to account the variance of the predicted concentration of the analyte of interest testing the null hypothesis, and the MDL will be given by intersection of the lower confidence interval of the multivariate model with a straight line for an analyte concentration at zero level.

Besides LOD and MLD techniques, another approach is related to use the information of receiver operation curve (ROC) for determining LOD (Brown and Davis, 2006). This approach is a univariate technique that is used as an alternative when data has a non-parametric distribution whose noise is heteroscedastic.

2.7.1. Limit of detection applied to IMS

Limit of detection in ion mobility spectrometry is commonly determined from predictions of blanks that are obtained from a calibration curve which is built using the information of either peak area or height intensity of the peak of the analyte of interest. The common way of estimate LOD is determined by three times standard deviation of blanks divided by slope of calibration curve (Armenta and Blanco, 2012b), (Marquez-Sillero et al., 2012),(Moran et al., 2012),(Zhou et al., 2012) and (Garrido-Delgado et al., 2011b). This approximation is similar to Eq. 2.23 when k_D has a value of 3 which is equal to have 99.85% probability that the blank signal not exceed the LOD. Actually, neither the number of samples to build a calibration curve nor influences of other analytes present within sample are considered in this approximation. As seen in Eq. 2.29 (Mocak et al., 1997) a correction of k_D has to be introduced in the formulation of LOD in order to correct the uncertainties of having a limited number of points in calibration curve. Furthermore, it is important to pay attention in the fact that the usefulness either peak area or height intensity of a peak of interest can bring misunderstandings specially when there are present other substances, or even worst when the peak of interest is overlapped with another substance. Thus, multivariate techniques are needed to get a better profit of the IMS response.

As far as I know, in the literature there is only one publication about LOD applied to IMS techniques. Fraga in 2007 (Fraga et al., 2007) introduced the approach about ROC-curve (Brown and Davis, 2006) in order to be adopted in IMS field. In this case, IMS is used to detect the presence or absence of a proprietary chemical marker placed into diesel fuel in such a way to observe if there is any interaction and suppression of this chemical marker due to diesel. Several experiments were performed at different concentration of the marker. From these experiments different ROC curves were built, thus the best AUC of the ROC curve was obtained when the presence of the marked is considerable high compared with the background of the diesel. This strategy seems to be suitable for

achieving the LOD when there is a complex chemical matrix. However, this approach does not provide a specific LOD, but a concentration range where LOD is located.

2.8. Cross validation methodologies

As it was mention above, over fitting results are obtained when the number of features are much bigger than the number of samples, or when the parameters of the model was only set up using a subset of samples. The fact of having over fitting results implies that the results seem better than they really are. Therefore, validation is always a crucial step for avoiding over fitting results. Validation helps to confirm results of a model and to set up parameters of a model. The validation can be divided into external and internal validation. Internal validation is used for setting up the parameters of the model and external validation is used for measuring the robustness of the model.

The simplest method is to keep apart a subset of samples as test data and the rest of the samples are used for building the model. This method is called hold-out validation and the test data is used for giving the final accuracy of the model. Indeed, the results from this partition out of a finite number of samples cannot be generalized and the result is not enough accurate. In contrast, the use of cross validation techniques gives a better description about the data and the models.

Cross validation is an iterative approach that split the data in training and validation data, thus different models are build and validated, so the final error is the mean error between all models. There are three methods commonly used as cross validation technique which are k-fold, leave one out and random sub-sampling cross validation (Filzmoser et al., 2009). Another approach is bootstrap validation (Efron, 1979, Felsenstein, 1985). Bootstrap validation is the most powerful methodology for validation which is also considered as a random sub-sampling validation with replacement. In each iteration samples for training set are chosen randomly, and the samples were not selected in training set are part of validation set. The “replacement” means that training set can have repeated samples. It is considered that in each iteration 30% of the data is left out for validation purposes.

The use of cross validation is really important for setting up parameters of models and classifiers. Thus, some authors propose the use of a double cross-validation with inner loops for setting up parameters as it is shown in Figure 2.8. The data is initially split in validation and training set, then the training set is used for model optimization purposes in which a cross validation strategy can be used. Once the model parameters are optimized, a final model is built with the training data set and validated with the validation set to get the final error results. A cross validation outer loop can be used either to test the model with the parameters previously selected or to generate different models and combine them using other strategies.

IMS dataset is made up of several spectra of different measurements, and each measurement is composed by a number of spectra per hundred of drift time points. Thus, in a validation process is important to taking into account a specific subset of scans or spectra that belongs to a sample, and not only individual spectrum. Therefore, when a sample is left out, all spectra of the sample have to be kept apart and not only one spectrum. At the end, the error should be calculated as mean or majority vote from all spectra of a sample.

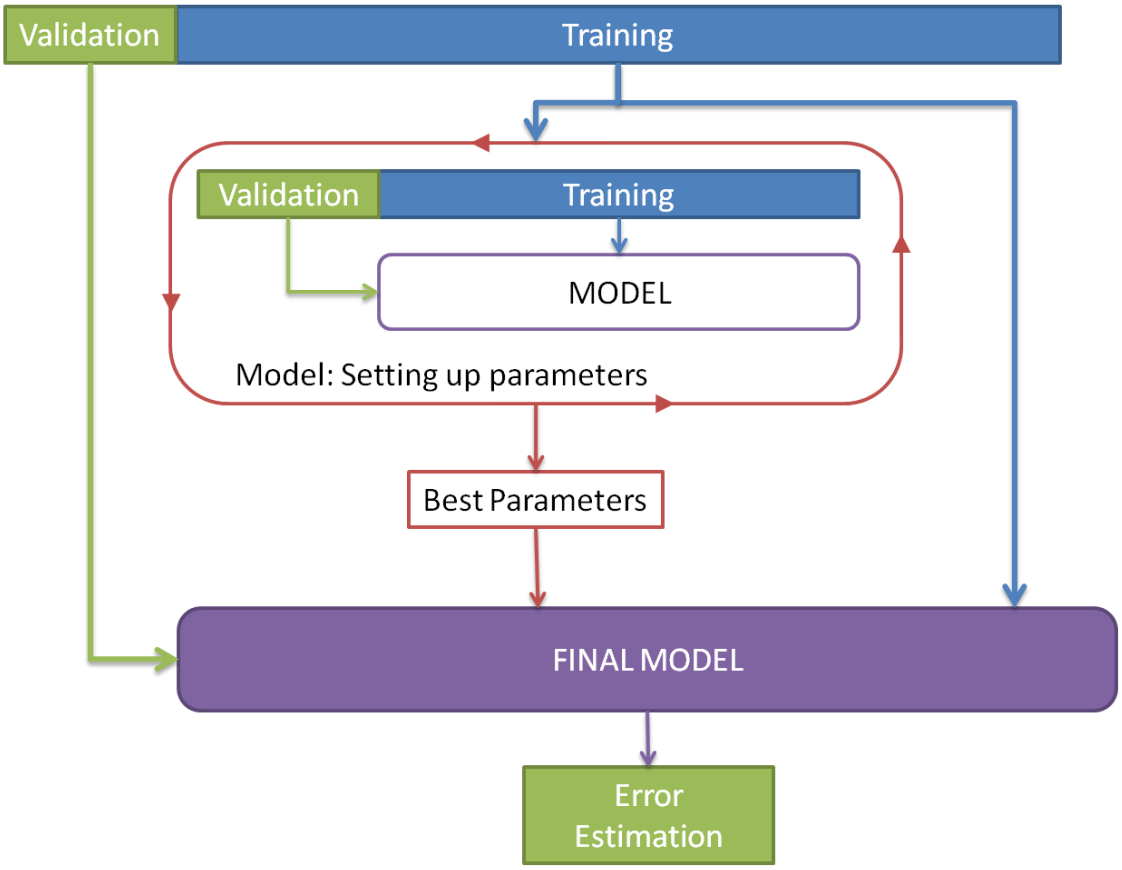

Figure 2.8 Cross validation for setting up model parameters.

Both quantitative and qualitative model require a thorough validation process. The best method for validation is boostrap validation, but it is necessary to have considerable amount of samples in order to get feasible results. However, many times is not possible to have too much samples, especially in clinical applications where it is common to have just few of them.

In addition, it is necessary to add external and internal validation in the methodology. Thus means, the model will be always tested with data that is not part of the model. This implies that the model will be more robust and ensure that any sample can be properly classified. This methodology is really strict and requires to add a double loop of validation. The first one for setting up the internal model parameters and the second one for testing the external dataset. This procedure is surely time consuming, and requires to have enough samples for this iteration process.

The validation process is not usually deeply studied in IMS field. It might be due to the fact that usual scenarios where IMS has been used do not require an exhaustive

validation process, but the biological/biomedical scenario need an adequate validation for ensuring reliable results. In this thesis, different biological applications have been studied and the validation process has been carefully design in order to avoid over fitting results.

2.9. Summary

Nowadays, the measurement of Volatile Organic compounds (VOCs) is an active field for research and development. Certainty, there are reference techniques for VOC measurement that are extendedly used, but besides their high economic cost, their high cost in time for each measure is a drawback in applications when it is necessary to perform on-line measurements. In this sense, IMS can be a complement or an alternative to reference techniques, but there has been a lack about how to perform a proper signal processing analysis. Actually, when new applications have emerged in IMS field, the need of developing new signal processing strategies were evident for obtaining reliable and accurate results.

It is important to remark that IMS has a moderate selectivity and non-linear behave, and changes in temperature and humidity might affect the analysis of the information of the IMS. The primarily use of IMS was in explosive and illicit chemicals detection in which the detection was based on a binary decision using one or few known compounds. In virtue of novel bio related applications has appeared, the interest of using multivariate techniques to address complex data from IMS have been raised too.

The samples, which are obtained from this kind of applications, contain thousands of compounds that make unfeasible the use of univariate strategies. Moreover, the compounds behave into IMS brings additional drawbacks, such as overlapping of peaks from unknown compounds, that need to be solved. Therefore, multivariate strategies are proposed to handle with these problems and extract informative information.

This chapter has covered from a general view a comparison of multivariate and univariate analysis. In addition, how multivariate and univariate techniques has been applied into the IMS field. The main proposal of this thesis is tackled IMS problems such as non linearities and mixtures through multivariate techniques. These problems occur due to chemical response of the spectrometer or complexity of the application. Next chapters will cover the following issues. Chapter three explains the experiments that were carried out and the spectrometers that were used. Chapter four explores the qualitative applications. In chapter five, IMS data is explored from a quantitative point of view, where calibration models are built and limit of detection are evaluated.

Reference

1987. RECOMMENDATIONS FOR THE DEFINITION, ESTIMATION AND USE OF THE DETECTION LIMIT. *Analyst,* 112**,** 199-204.

Abdollahi, H. & Sajjadi, S. M. 2010. On rotational ambiguity in parallel factor analysis. *Chemometrics and Intelligent Laboratory Systems,* 103**,** 144-151.

Abdollahi, H. & Tauler, R. 2011. Uniqueness and rotation ambiguities in Multivariate Curve Resolution methods. *Chemometrics and Intelligent Laboratory Systems,* 108**,** 100-111.

Armenta, S. & Blanco, M. 2012a. Ion mobility spectrometry as a high-throughput analytical tool in occupational pyrethroid exposure. *Analytical and Bioanalytical Chemistry,* 404**,** 635-648.

Armenta, S. & Blanco, M. 2012b. Ion mobility spectrometry for monitoring diamine oxidase activity. *Analyst,* 137**,** 5891-5897.

Atmanene, C., Petiot-Becard, S., Zeyer, D., Van Dorsselaer, A., Hannah, V. V. & Sanglier-Cianferani, S. 2012. Exploring Key Parameters to Detect Subtle Ligand-Induced Protein Conformational Changes Using Traveling Wave Ion Mobility Mass Spectrometry. *Analytical Chemistry,* 84**,** 4703-4710.

Barker, M. & Rayens, W. 2003. Partial least squares for discrimination. *Journal of Chemometrics,* 17.

Bauer, G., Wegscheider, W. & Ortner, H. M. 1991. SELECTIVITY AND ERROR-ESTIMATES IN MULTIVARIATE CALIBRATION - APPLICATION TO SEQUENTIAL ICP-OES. *Spectrochimica Acta Part B-Atomic Spectroscopy,* 46**,** 1185-1196.

Bell, S. E., Wang, Y. F., Walsh, M. K., Du, Q., Ewing, R. G. & Eiceman, G. A. 1995. QUALITATIVE AND QUANTITATIVE-EVALUATION OF DECONVOLUTION FOR ION MOBILITY SPECTROMETRY. *Analytica Chimica Acta,* 303**,** 163-174.

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P. & Plemmons, R. J. 2007. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis,* 52**,** 155-173.

Beyramysoltan, S., Rajko, R. & Abdollahi, H. 2013. Investigation of the equality constraint effect on the reduction of the rotational ambiguity in three-component system using a novel grid search method. *Analytica Chimica Acta,* 791**,** 25-35.

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

Boque, R., Faber, N. K. M. & Rius, F. X. 2000. Detection limits in classical multivariate calibration models. *Analytica Chimica Acta,* 423**,** 41-49.

Boque, R. & Rius, F. X. 1996. Multivariate detection limits estimators. *Chemometrics and Intelligent Laboratory Systems,* 32**,** 11-23.

Brown, C. D. & Davis, H. T. 2006. Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems,* 80**,** 24-38.

Brown, S. D., Sum, S. T., Despagne, F. & Lavine, B. K. 1996. Chemometrics. *Analytical Chemistry,* 68**,** R21-R61.

Burges, C. J. C. 1998. A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery,* 2.

Buxton, T. L. & Harrington, P. D. 2001. Rapid multivariate curve resolution applied to identification of explosives by ion mobility spectrometry. *Analytica Chimica Acta,* 434**,** 269-282.

Bylesjo, M., Rantalainen, M., Cloarec, O., Nicholson, J. K., Holmes, E. & Trygg, J. 2006. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics,* 20**,** 341-351.

Cao, L. B., Harrington, P. D. & Liu, J. D. 2005. SIMPLISMA and ALS applied to two-way nonlinear wavelet compressed ion mobility spectra of chemical warfare agent simulants. *Analytical Chemistry,* 77**,** 2575-2586.

Chang, C.-C. & Lin, C.-J. 2011. LIBSVM: A Library for Support Vector Machines. *Acm Transactions on Intelligent Systems and Technology,* 2.

Chen, G. & Harrington, P. B. 2001. Real-time interactive self-modeling mixture analysis. *Applied Spectroscopy,* 55**,** 621-629.

Chen, G. X. & Harrington, P. D. 2003. SIMPLISMA applied to two-dimensional wavelet compressed ion mobility spectrometry data. *Analytica Chimica Acta,* 484**,** 75-91.

Cherkassky, V. & Ma, Y. Q. 2004. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks,* 17**,** 113-126.
Cichocki, A., Zdunek, R. & Amari, S.-I. 2008. Nonnegative matrix and tensor factorization. *Ieee Signal Processing Magazine,* 25**,** 142-145.
Clayton, C. A., Hines, J. W. & Elkins, P. D. 1987. DETECTION LIMITS WITH SPECIFIED ASSURANCE PROBABILITIES. *Analytical Chemistry,* 59**,** 2506-2514.
Comon, P. 1994. INDEPENDENT COMPONENT ANALYSIS, A NEW CONCEPT. *Signal Processing,* 36**,** 287-314.
Currie, L. A. 1968. LIMITS FOR QUALITATIVE DETECTION AND QUANTITATIVE DETERMINATION - APPLICATION TO RADIOCHEMISTRY. *Analytical Chemistry,* 40**,** 586-&.
Currie, L. A. 1995. NOMENCLATURE IN EVALUATION OF ANALYTICAL METHODS INCLUDING DETECTION AND QUANTIFICATION CAPABILITIES (IUPAC RECOMMENDATIONS 1995). *Pure and Applied Chemistry,* 67**,** 1699-1723.
Currie, L. A. 1997. Detection: International update, and some emerging di-lemmas involving calibration, the blank, and multiple detection decisions. *Chemometrics and Intelligent Laboratory Systems,* 37**,** 151-181.
Danzer, K., Currie, L. A. & Commission Gen Aspects Analyt, C. 1998. Guidelines for calibration in analytical chemistry - Part 1. Fundamentals and single component calibration (IUPAC recommendations 1998). *Pure and Applied Chemistry,* 70**,** 993-1014.
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.
deJuan, A., VanderHeyden, Y., Tauler, R. & Massart, D. L. 1997. Assessment of new constraints applied to the alternating least squares method. *Analytica Chimica Acta,* 346**,** 307-318.
Desai, K., Badhe, Y., Tambe, S. S. & Kulkarni, B. D. 2006. Soft-sensor development for fed-batch bioreactors using support vector regression. *Biochemical Engineering Journal,* 27**,** 225-239.
Dimitrov, B. D., Motterlini, N. & Fahey, T. 2015. A simplified approach to the pooled analysis of calibration of clinical prediction rules for systematic reviews of validation studies. *Clinical epidemiology,* 7**,** 267-80.
Duda, R. O., Hart, P. E. & Stork, D. G. 2000. *Pattern Classification*.
Dunn, J. D., Gryniewicz-Ruzicka, C. M., Mans, D. J., Mecker-Pogue, L. C., Kauffman, J. F., Westenberger, B. J. & Buhse, L. F. 2012. Qualitative screening for adulterants in weight-loss supplements by ion mobility spectrometry. *Journal of Pharmaceutical and Biomedical Analysis,* 71**,** 18-26.
Efron, B. 1979. 1977 RIETZ LECTURE - BOOTSTRAP METHODS - ANOTHER LOOK AT THE JACKKNIFE. *Annals of Statistics,* 7**,** 1-26.
Eiceman, G. A. & Karpas, Z. 2005. *Ion Mobility Spectrometry,* Florida, Taylor & Francis Group.
Eisenbeis, R. A. & Avery, R. B. 1972. *Discriminant Analysis and CLassification Procedures: Therory and Applications,* Lexington, D. C. Healthy and Company.
Escandar, G. M., Damiani, P. C., Goicoechea, H. C. & Olivieri, A. C. 2006. A review of multivariate calibration methods applied to biomedical analysis. *Microchemical Journal,* 82**,** 29-42.
Esteban, M., Arino, C., Diaz-Cruz, J. M., Diaz-Cruz, M. S. & Tauler, R. 2000. Multivariate curve resolution with alternating least squares optimisation: a soft-modelling approach to metal complexation studies by voltammetric techniques. *Trac-Trends in Analytical Chemistry,* 19**,** 49-61.
Faber, K. & Kowalski, B. R. 1997. Improved estimation of the limit of detection in multivariate calibration. *Fresenius Journal of Analytical Chemistry,* 357**,** 789-795.
Fahlman, F. E. & Lebiere, C. 1991. The cascade-correlation architecture. Carnegie Mellon University Report, CMU-CS-90-100.
Feldmann, K., Scheuerer, M. & Thorarinsdottir, T. L. 2015. Spatial Postprocessing of Ensemble Forecasts for Temperature Using Nonhomogeneous Gaussian Regression. *Monthly Weather Review,* 143**,** 955-971.
Felsenstein, J. 1985. CONFIDENCE-LIMITS ON PHYLOGENIES - AN APPROACH USING THE BOOTSTRAP. *Evolution,* 39**,** 783-791.

Fernandez-Maestre, R., Harden, C. S., Ewing, R. G., Crawford, C. L. & Hill, H. H. 2010. Chemical standards in ion mobility spectrometry. *Analyst,* 135**,** 1433-1442.

Filzmoser, P., Liebmann, B. & Varmuza, K. 2009. Repeated double cross validation. *Journal of Chemometrics,* 23**,** 160-171.

Forina, M., Casolino, M. C. & Martinez, C. D. P. 1998. Multivariate calibration: applications to pharmaceutical analysis. *Journal of Pharmaceutical and Biomedical Analysis,* 18**,** 21-33.

Forina, M., Lanteri, S. & Casale, A. 2007. Multivariate calibration. *Journal of Chromatography A,* 1158**,** 61-93.

Fraga, C. G., Kerr, D. R. & Atkinson, D. A. 2009. Improved quantitative analysis of ion mobility spectrometry by chemometric multivariate calibration. *Analyst,* 134**,** 2329-2337.

Fraga, C. G., Melville, A. M. & Wright, B. W. 2007. ROC-curve approach for determining the detection limit of a field chemical sensor. *Analyst,* 132**,** 230-236.

Friedman, J. H. 1989. REGULARIZED DISCRIMINANT-ANALYSIS. *Journal of the American Statistical Association,* 84.

Gan, F., Ruan, G. H. & Mo, J. Y. 2006. Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems,* 82**,** 59-65.

Garrido-Delgado, R., Arce, L., Guaman, A. V., Pardo, A., Marco, S. & Valcarcel, M. 2011a. 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.

Garrido-Delgado, R., Arce, L. & Valcarcel, M. 2012. Multi-capillary column-ion mobility spectrometry: a potential screening system to differentiate virgin olive oils. *Analytical and Bioanalytical Chemistry,* 402**,** 489-498.

Garrido-Delgado, R., Mercader-Trejo, F., Arce, L. & Valcarcel, M. 2011b. Enhancing sensitivity and selectivity in the determination of aldehydes in olive oil by use of a Tenax TA trap coupled to a UV-ion mobility spectrometer. *Journal of Chromatography A,* 1218**,** 7543-7549.

Geladi, P. & Kowalski, B. R. 1986. PARTIAL LEAST-SQUARES REGRESSION - A TUTORIAL. *Analytica Chimica Acta,* 185.

Gemperline, P. J. 1984. A PRIORI ESTIMATES OF THE ELUTION PROFILES OF THE PURE COMPONENTS IN OVERLAPPED LIQUID-CHROMATOGRAPHY PEAKS USING TARGET FACTOR-ANALYSIS. *Journal of Chemical Information and Computer Sciences,* 24**,** 206-212.

Gemperline, P. J. 1999. Computation of the range of feasible solutions in self-modeling curve resolution algorithms. *Analytical Chemistry,* 71**,** 5398-5404.

Golshan, A., Abdollahi, H. & Maeder, M. 2012. The reduction of rotational ambiguity in soft-modeling by introducing hard models. *Analytica Chimica Acta,* 709**,** 32-40.

Good, I. J. & Gaskins, R. A. 1971. NONPARAMETRIC ROUGHNESS PENALTIES FOR PROBABILITY DENSITIES. *Biometrika,* 58**,** 255-&.

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.

Haaland, D. M. & Thomas, E. V. 1988. PARTIAL LEAST-SQUARES METHODS FOR SPECTRAL ANALYSES .1. RELATION TO OTHER QUANTITATIVE CALIBRATION METHODS AND THE EXTRACTION OF QUALITATIVE INFORMATION. *Analytical Chemistry,* 60.

Harrington, P. B. & Chen, P. Equilibrium modeling of Ion Mobility Spectrometry. The Conference Proceedings of the Thirteenth International Workshop on IMS, 2004. 160-181.

Hastie, T., Tibshirani, R. & Friedman, J. 2003. *The Elements of Statistical Learning. Data Mining, Inference and Prediction. ,* New York, Springer Verlag.

Helland, I. S. 1990. PARTIAL LEAST-SQUARES REGRESSION AND STATISTICAL-MODELS. *Scandinavian Journal of Statistics,* 17**,** 97-114.

Henley, W. E. & Hand, D. J. 1996. A k-nearest-neighbour classifier for assessing consumer credit risk. *Statistician,* 45**,** 77-95.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology,* 24**,** 417-441.

Huang, S. C. & Huang, Y. F. 1991. BOUNDS ON THE NUMBER OF HIDDEN NEURONS IN MULTILAYER PERCEPTRONS. *Ieee Transactions on Neural Networks,* 2.
Hubaux, A. & Vos, G. 1970. DECISION AND DETECTION LIMITS FOR LINEAR CALIBRATION CURVES. *Analytical Chemistry,* 42**,** 849-&.
Huebert, T., Tiebe, C. & Stephan, I. 2011. Detection of fungal infestations of wood by ion mobility spectrometry. *International Biodeterioration & Biodegradation,* 65**,** 675-681.
Jafari, M. T., Khayamian, T., Shaer, V. & Zarei, N. 2007. Determination of veterinary drug residues in chicken meat using corona discharge ion mobility spectrometry. *Analytica Chimica Acta,* 581**,** 147-153.
Jiang, J. H. & Ozaki, Y. 2002. Self-modeling curve resolution (SMCR): Principles, techniques, and applications. *Applied Spectroscopy Reviews,* 37**,** 321-345.
Kalivas, J. H. 2005. Multivariate calibration, an overview. *Analytical Letters,* 38**,** 2259-2279.
Keller, H. R. & Massart, D. L. 1991. PEAK PURITY CONTROL IN LIQUID-CHROMATOGRAPHY WITH PHOTODIODE-ARRAY DETECTION BY A FIXED SIZE MOVING WINDOW EVOLVING FACTOR-ANALYSIS. *Analytica Chimica Acta,* 246**,** 379-390.
Khayamian, T., Sajjadi, S. M., Mirmahdieh, S., Mardihallaj, A. & Hashemian, Z. 2012. Simultaneous analysis of bifenthrin and tetramethrin using corona discharge ion mobility spectrometry and Tucker 3 model. *Chemometrics and Intelligent Laboratory Systems,* 118**,** 88-96.
Kiesel, J., Schroeder, M., Hering, D., Schmalz, B., Hoermann, G., Jaehnig, S. C. & Fohrer, N. 2015. A new model linking macroinvertebrate assemblages to habitat composition in rivers: development, sensitivity and univariate application. *Fundamental and Applied Limnology,* 186**,** 117-+.
Kim, T., Attias, H. T., Lee, S.-Y. & Lee, T.-W. 2007. Blind source separation exploiting higher-order frequency dependencies. *Ieee Transactions on Audio Speech and Language Processing,* 15**,** 70-79.
Kuske, M., Rubio, R., Romain, A. C., Nicolas, J. & Marco, S. 2005. Fuzzy k-NN applied to moulds detection. *Sensors and Actuators B-Chemical,* 106**,** 52-60.
Kvalheim, O. M. & Liang, Y. Z. 1992. HEURISTIC EVOLVING LATENT PROJECTIONS - RESOLVING 2-WAY MULTICOMPONENT DATA .1. SELECTIVITY, LATENT-PROJECTIVE GRAPH, DATASCOPE, LOCAL RANK, AND UNIQUE RESOLUTION. *Analytical Chemistry,* 64**,** 936-946.
Leardi, R. 2001. Genetic algorithms in chemometrics and chemistry: a review. *Journal of Chemometrics,* 15**,** 559-569.
Leardi, R., Boggia, R. & Terrile, M. 1992. GENETIC ALGORITHMS AS A STRATEGY FOR FEATURE-SELECTION. *Journal of Chemometrics,* 6.
Leger, M. N. & Wentzell, P. D. 2002. Dynamic Monte Carlo self-modeling curve resolution method for multicomponent mixtures. *Chemometrics and Intelligent Laboratory Systems,* 62**,** 171-188.
Liang, Y. Z. & Kvalheim, O. M. 1993. HEURISTIC EVOLVING LATENT PROJECTIONS - RESOLVING HYPHENATED CHROMATOGRAPHIC PROFILES BY COMPONENT STRIPPING. *Chemometrics and Intelligent Laboratory Systems,* 20**,** 115-125.
Long, G. L. & Winefordner, J. D. 1983. LIMIT OF DETECTION. *Analytical Chemistry,* 55**,** A712-&.
Lorber, A. 1986. ERROR PROPAGATION AND FIGURES OF MERIT FOR QUANTIFICATION BY SOLVING MATRIX EQUATIONS. *Analytical Chemistry,* 58**,** 1167-1172.
Lu, Y., O'Donnell, R. M. & Harrington, P. B. 2009. Detection of cocaine and its metabolites in urine using solid phase extraction-ion mobility spectrometry with alternating least squares. *Forensic Science International,* 189**,** 54-59.
Maeder, M. 1987. EVOLVING FACTOR-ANALYSIS FOR THE RESOLUTION OF OVERLAPPING CHROMATOGRAPHIC PEAKS. *Analytical Chemistry,* 59**,** 527-530.
Malinowski, E. R. 1992. WINDOW FACTOR-ANALYSIS - THEORETICAL DERIVATION AND APPLICATION TO FLOW-INJECTION ANALYSIS DATA. *Journal of Chemometrics,* 6**,** 29-40.
Manne, R. 1995. ON THE RESOLUTION PROBLEM IN HYPHENATED CHROMATOGRAPHY. *Chemometrics and Intelligent Laboratory Systems,* 27**,** 89-94.
Marco, S. & Gutierrez-Galvez, A. 2012. Signal and Data Processing for Machine Olfaction and Chemical Sensing: A Review. *Ieee Sensors Journal,* 12**,** 3189-3214.

Marcus, S., Menda, A., Shore, L., Cohen, G., Atweh, E., Friedman, N. & Karpas, Z. 2012. A novel method for the diagnosis of bacterial contamination in the anterior vagina of sows based on measurement of biogenic amines by ion mobility spectrometry: A field trial. *Theriogenology,* 78**,** 753-758.

Marquez-Sillero, I., Cardenas, S. & Valcarcel, M. 2012. Headspace-multicapillary column-ion mobility spectrometry for the direct analysis of 2,4,6-trichloroanisole in wine and cork samples. *Journal of Chromatography A,* 1265**,** 149-154.

Massy, W. F. 1965. PRINCIPAL COMPONENTS REGRESSION IN EXPLORATORY STATISTICAL RESEARCH. *Journal of the American Statistical Association,* 60.

Mocak, J., Bond, A. M., Mitchell, S. & Scollary, G. 1997. A statistical overview of standard (IUPAC and ACS) and new procedures for determining the limits of detection and quantification: Application to voltammetric and stripping techniques (technical report). *Pure and Applied Chemistry,* 69**,** 297-328.

Moran, J., McCall, H., Yeager, B. & Bell, S. 2012. Characterization and validation of ion mobility spectrometry in methamphetamine clandestine laboratory remediation. *Talanta,* 100**,** 196-206.

Morsa, D., Gabelica, V. & De Pauw, E. 2011. Effective Temperature of Ions in Traveling Wave Ion Mobility Spectrometry. *Analytical Chemistry,* 83**,** 5775-5782.

Narendra, P. & Fukunaga, K. 1977. BRANCH AND BOUND ALGORITHM FOR FEATURE SUBSET SELECTION. *Ieee Transactions on Computers,* 26.

Norgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L. & Engelsen, S. B. 2000. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy,* 54**,** 413-419.

Olivieri, A. C., Faber, N. K. M., Ferre, J., Boque, R., Kalivas, J. H. & Mark, H. 2006. Uncertainty estimation and figures of merit for multivariate calibration. *Pure and Applied Chemistry,* 78**,** 633-661.

Osten, D. W. & Kowalski, B. R. 1984. MULTIVARIATE CURVE RESOLUTION IN LIQUID-CHROMATOGRAPHY. *Analytical Chemistry,* 56**,** 991-995.

Pal, S. K. & Mitra, S. 1992. MULTILAYER PERCEPTRON, FUZZY-SETS, AND CLASSIFICATION. *Ieee Transactions on Neural Networks,* 3.

Peng, J., Peng, S., Jiang, A., Wei, J., Li, C. & Tan, J. 2010. Asymmetric least squares for multiple spectra baseline correction. *Analytica Chimica Acta,* 683**,** 63-68.

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.

Pudil, P., Novovicová, J. & Kittler, J. 1994. Floating search methods in feature selection. *Pattern Recognition Letters,* 15**,** 1119-1125.

Rajko, R. & Istvan, K. 2005. Analytical solution for determining feasible regions of self-modeling curve resolution (SMCR) method based on computational geometry. *Journal of Chemometrics,* 19**,** 448-463.

Raudys, S. J. & Jain, A. K. 1991. SMALL SAMPLE-SIZE EFFECTS IN STATISTICAL PATTERN-RECOGNITION - RECOMMENDATIONS FOR PRACTITIONERS. *Ieee Transactions on Pattern Analysis and Machine Intelligence,* 13.

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.

Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E. & Suter, B. W. 1990. The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Transactions on Neural Networks,* 1.

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.

Sasic, S., Blackwood, D., Liu, A., Ward, H. W. & Clarke, H. 2015. Detailed analysis of the online near-infrared spectra of pharmaceutical blend in a rotary tablet press feed frame. *Journal of Pharmaceutical and Biomedical Analysis,* 103**,** 73-79.

Satoh, T., Kishi, S., Nagashima, H., Tachikawa, M., Kanamori-Kataoka, M., Nakagawa, T., Kitagawa, N., Tokita, K., Yamamoto, S. & Seto, Y. 2015. Ion mobility spectrometric analysis of vaporous chemical warfare agents by the instrument with corona discharge ionization ammonia dopant ambient temperature operation. *Analytica Chimica Acta,* 865**,** 39-52.

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

Sawall, M., Fischer, C., Heller, D. & Neymeyr, K. 2012. Reduction of the rotational ambiguity of curve resolution techniques under partial knowledge of the factors. Complementarity and coupling theorems. *Journal of Chemometrics,* 26**,** 526-537.

Smola, A. J. & Scholkopf, B. 2004. A tutorial on support vector regression. *Statistics and Computing,* 14**,** 199-222.

Snyder, A. P., Maswadeh, W. M., Eiceman, G. A., Wang, Y. F. & Bell, S. E. 1995. MULTIVARIATE STATISTICAL-ANALYSIS CHARACTERIZATION OF APPLICATION-BASED ION MOBILITY SPECTRA. *Analytica Chimica Acta,* 316**,** 1-14.

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.

Stlouis, R. H. & Hill, H. H. 1990. ION MOBILITY SPECTROMETRY IN ANALYTICAL-CHEMISTRY. *Critical Reviews in Analytical Chemistry,* 21**,** 321-355.

Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *Bmc Bioinformatics,* 8.

Sugiyama, M. 2007. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research,* 8**,** 1027-1061.

Suykens, J. A. K. & Vandewalle, J. 1999. Least squares support vector machine classifiers. *Neural Processing Letters,* 9**,** 293-300.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P. & Feuston, B. P. 2003. Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences,* 43**,** 1947-1958.

Swierenga, H., Wulfert, F., de Noord, O. E., de Weijer, A. P., Smilde, A. K. & Buydens, L. M. C. 2000. Development of robust calibration models in near infra-red spectrometric applications. *Analytica Chimica Acta,* 411**,** 121-135.

Takahashi, Y., Takatani, T., Osako, K., Saruwatari, H. & Shikano, K. 2009. Blind Spatial Subtraction Array for Speech Enhancement in Noisy Environment. *Ieee Transactions on Audio Speech and Language Processing,* 17**,** 650-664.

Tauler, R. 2001. Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution. *Journal of Chemometrics,* 15**,** 627-646.

Tauler, R., Kowalski, B. & Fleming, S. 1993. MULTIVARIATE CURVE RESOLUTION APPLIED TO SPECTRAL DATA FROM MULTIPLE RUNS OF AN INDUSTRIAL-PROCESS. *Analytical Chemistry,* 65**,** 2040-2047.

Thomas, E. V. 1994. A PRIMER ON MULTIVARIATE CALIBRATION. *Analytical Chemistry,* 66**,** A795-A804.

Thomas, E. V. & Haaland, D. M. 1990. COMPARISON OF MULTIVARIATE CALIBRATION METHODS FOR QUANTITATIVE SPECTRAL-ANALYSIS. *Analytical Chemistry,* 62**,** 1091-1099.

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.

Trygg, J., Holmes, E. & Lundstedt, T. 2007. Chemometrics in metabonomics. *Journal of Proteome Research,* 6.

Trygg, J. & Wold, S. 2002. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics,* 16.

Vandeginste, B. G. M., Derks, W. & Kateman, G. 1985. MULTICOMPONENT SELF-MODELING CURVE RESOLUTION IN HIGH-PERFORMANCE LIQUID-CHROMATOGRAPHY BY ITERATIVE TARGET TRANSFORMATION ANALYSIS. *Analytica Chimica Acta,* 173**,** 253-264.

Verkouteren, J. R. & Staymates, J. L. 2011. Reliability of ion mobility spectrometry for qualitative analysis of complex, multicomponent illicit drug samples. *Forensic Science International,* 206**,** 190-196.

Voigtman, E. 2008a. Limits of detection and decision. Part 1. *Spectrochimica Acta Part B-Atomic Spectroscopy,* 63**,** 115-128.

Voigtman, E. 2008b. Limits of detection and decision. Part 2. *Spectrochimica Acta Part B-Atomic Spectroscopy,* 63**,** 129-141.

Voigtman, E. 2008c. Limits of detection and decision. Part 3. *Spectrochimica Acta Part B-Atomic Spectroscopy,* 63**,** 142-153.

Voigtman, E. 2008d. Limits of detection and decision. Part 4. *Spectrochimica Acta Part B-Atomic Spectroscopy,* 63**,** 154-165.

Vosough, M., Mason, C., Tauler, R., Jalali-Heravi, M. & Maeder, M. 2006. On rotational ambiguity in model-free analyses of multivariate data. *Journal of Chemometrics,* 20**,** 302-310.

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).

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.

Windig, W., Antalek, B., Lippert, J. L., Batonneau, Y. & Bremard, C. 2002. Combined use of conventional and second-derivative data in the SIMPLISMA self-modeling mixture analysis approach. *Analytical Chemistry,* 74**,** 1371-1379.

Windig, W., Gallagher, N. B., Shaver, J. M. & Wise, B. M. 2005. A new approach for interactive self-modeling mixture analysis. *Chemometrics and Intelligent Laboratory Systems,* 77**,** 85-96.

Windig, W. & Guilment, J. 1991. INTERACTIVE SELF-MODELING MIXTURE ANALYSIS. *Analytical Chemistry,* 63**,** 1425-1432.

Wold, S. 1992. NONLINEAR PARTIAL LEAST-SQUARES MODELING .2. SPLINE INNER RELATION. *Chemometrics and Intelligent Laboratory Systems,* 14.

Wold, S., Kettanehwold, N. & Skagerberg, B. 1989. NONLINEAR PLS MODELING. *Chemometrics and Intelligent Laboratory Systems,* 7.

Zamora, D., Alcala, M. & Blanco, M. 2011. Determination of trace impurities in cosmetic intermediates by ion mobility spectrometry. *Analytica Chimica Acta,* 708**,** 69-74.

Zamora, D. & Blanco, M. 2012. Improving the efficiency of ion mobility spectrometry analyses by using multivariate calibration. *Analytica Chimica Acta,* 726**,** 50-56.

Zhang, J., Zhang, L., Yang, G., Wu, D., Jiang, L., Huang, L., Wen, Z. & Li, M. 2013. Nonnegative matrix factorization for the improvement in sensitivity of discovering potentially disease-related genes. *Chemometrics and Intelligent Laboratory Systems,* 126**,** 100-107.

Zhang, Z.-M., Chen, S. & Liang, Y.-Z. 2010. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst,* 135.

Zheng, P., Harrington, P. D. & Davis, D. M. 1996. Quantitative analysis of volatile organic compounds using ion mobility spectrometry and cascade correlation neural networks. *Chemometrics and Intelligent Laboratory Systems,* 33**,** 121-132.

Zhou, Q., Wang, W., Cang, H., Du, Y., Han, F., Chen, C., Cheng, S., Li, J. & Li, H. 2012. On-line measurement of propofol using membrane inlet ion mobility spectrometer. *Talanta,* 98**,** 241-246.