Talanta 88 (2012) 95–103

Contents lists available at SciVerse ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Feature selection versus feature compression in the building of calibration models from FTIR-spectrophotometry datasets

Alexander Vergara[a,*], Eduard Llobet[b]

[a] BioCircuits Institute, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0402, USA
[b] MINOS-EMaS, Department of Electronic Engineering, University Rovira i Virgili, Avda. Països Catalans 26, 43007 Tarragona, Spain

ARTICLE INFO

Article history:
Received 23 June 2011
Received in revised form 5 October 2011
Accepted 13 October 2011
Available online 20 October 2011

Keywords:
Minimum Redundancy-Maximum Relevance
Self organizing map
Feature selection
Feature compression
Regression models
FTIR-spectrophotometry

ABSTRACT

Undoubtedly, FTIR-spectrophotometry has become a standard in chemical industry for monitoring, on-the-fly, the different concentrations of reagents and by-products. However, representing chemical samples by FTIR spectra, which spectra are characterized by hundreds if not thousands of variables, conveys their own set of particular challenges because they necessitate to be analyzed in a high-dimensional feature space, where many of these features are likely to be highly correlated and many others surely affected by noise. Therefore, identifying a subset of features that preserves the classifier/regressor performance seems imperative prior any attempt to build an appropriate pattern recognition method. In this context, we investigate the benefit of utilizing two different dimensionality reduction methods, namely the minimum Redundancy-Maximum Relevance (mRMR) feature selection scheme and a new self-organized map (SOM) based feature compression, coupled to regression methods to quantitatively analyze two-component liquid samples utilizing FTIR spectrophotometry. Since these methods give us the possibility of selecting a small subset of relevant features from FTIR spectra preserving the statistical characteristics of the target variable being analyzed, we claim that expressing the FTIR spectra by these dimensionality-reduced set of features may be beneficial. We demonstrate the utility of these novel feature selection schemes in quantifying the distinct analytes within their binary mixtures utilizing a FTIR-spectrophotometer.

1. Introduction

Recently, there has been a growing demand for and rapid increase in developing new and sophisticated analytical methods for monitoring and identifying chemical compounds in a variety of chemical industry applications in general, and, in particular, for monitoring on-the-fly the concentrations of reagents and by-products [1]. Among the widely used tools to achieve this goal, infrared spectroscopy has been a workhorse technique for materials' analysis in the laboratory for over seventy years [2–5]. An infrared spectrum represents a fingerprint of a sample with absorption peaks that correspond to the frequencies of vibrations between the bonds of the atoms making up the material [6]. Since each different material has a unique molecular structured combination of atoms, in principle no two compounds produce the exact same infrared spectrum [7]. Therefore, the resulting infrared spectrum signature defines the chemical compound identity, whereas the size of the peaks in this annotated spectrum is a direct indicator of the amount of material present. With the advent of Fourier transform infrared spectrophotometry (FTIR-spectrophotometry), infrared analysis has become a breakthrough in performing tasks connected with the machine olfaction applications of identifying, classifying, and, when coupled with proper computational regression methods, quantifying chemical analytes of an unknown mixture [7]. The FTIR-spectrophotometry, which name refers to the manner that the data is collected and converted from an interference pattern to an actual spectrum, is preferred over its former dispersive-type-of-instrument modality for a variety of reasons: first, it is a non-destructive technique; second, it provides a precise measurement methodology with no external calibration needed; third, it has a greater optical throughput; and fourth, it only has one mechanic moving part, which in essence means that the speed and sensitivity of the FTIR-spectrophotometer to quantitate the multiple concentrations of materials within multi-component mixtures augments. And yet, the FTIR spectra are characterized by hundreds if not thousands of variables (transmittances), which present its own set of challenges since the analysis has to be conducted in a high-dimensional feature space, where many of these features are likely to be highly correlated and many others surely affected by noise. Therefore, a feature step leading to a reduction in dimensionality seems imperative prior any attempt to build an appropriate pattern recognition method.

* Corresponding author. Tel.: +1 8585346758; fax: +1 8585347664.
E-mail address: vergara@ucsd.edu (A. Vergara).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.10.014

By far, the most common choice to perform the multivariate modeling of the FTIR spectra, and hence reducing its dimensionality, is utilizing calibration models that will facilitate, yet perform in a more efficient way, the interpretation of the information entangled in the mentioned spectra. The two main approaches consist of either selecting an optimal subset of factors, e.g. the latent variables (LVs) or principal components (PCs) of the partial least square (PLS) or principal component regression (PCR) models, respectively [8–11], or selecting the wavelengths to be used to build a multivariate model (e.g., utilizing a multi-linear regression (MLR) model) [12–16]. The main advantage of PLS and PCR is their ability to compress the relevant information into a few orthogonal LVs or PCs, in which their orthogonality allow irrelevant PCs and LVs to be removed that in essence leads to more efficient models. However, factor selection uses the full spectrum—including noisy wavelengths—to compute the factors before selecting from among them, which means that the variables chosen to perform regression may not have a direct physical meaning with respect to the chemical sample being analyzed. Moreover, the selection of an optimal subset of factors may not be necessarily straightforward because the magnitude of an eigenvalue is not always a measure of its significance for the calibration [17]. On the other hand, methods based on selected wavelengths (e.g., MLR) are often preferred to their alternative factor selection methods because they use original variables (i.e., absorbances that are directly related to the chemical information) to create the regression model; becoming thereby, easier to interpret. Therefore, these methods are expected to be robust toward the experimental conditions of each specific application. However, selecting from the full spectrum of wavelengths is challenging because there is considerable overlapping among the spectra and the distinctive features are almost imperceptible, not to mention the effect of noise to those spectra. Consequently, MLR is only attractive if applied to a few correctly selected spectral variables, since the collinearity within the variables may jeopardize the stability of the method [12].

In the last few years considerable attention has been given to strategies for feature selection (or variable selection, among many other names given) in spectroscopic analysis [18–23]. The task of feature selection is to reduce the number of variables used in training a pattern recognition algorithm (i.e., a classifier or regression tool) [24]. Three main benefits can be drawn from successful feature selection: first, a substantial gain in computational efficiency (especially important for any application that requires classifier execution in real time); second, scientific discovery by determining which features are most correlated with the class labels (which may in turn reveal unknown relationships among features); and, third, reduction of the risk of overfitting if too few training instances are available (a serious problem particularly in situations with high dimensionalities relative to training set sizes). Many methods have been suggested to solve the variable selection problem within the field of machine olfaction and chemometrics. In the most generic way, they can be categorized into three groups [25]. Filter methods that perform feature selection that is independent of the classifier [26–28]. Wrapper methods, which use search techniques to select candidate subsets of variables and evaluate their fitness based on classification accuracy [29–31]. Finally, embedded methods that incorporate feature selection in the classifier objective function or algorithm [32,33]. However, despite all the success stories for feature selection methods to effectively select wavelengths in many different spectroscopic analysis based application scenarios, the solution found should be investigated carefully because many of these algorithms (e.g., genetic algorithms (GA)), may not prevent from meaningless variables (i.e., random non-relevant variables) to be selected [23,34], originating thus prohibitive expensive computational costs.

With this motivation, our work in this paper focuses on three issues that have not been considered in previous attempts. In the first approach, we introduce to the chemo-sensing community a novel formulation, namely minimum Redundancy-Maximum Relevance (mRMR) filter, to select a subset of relevant features from FTIR spectra. Thus, the first goal of this paper is to evaluate the usability of the mRMR feature selection algorithm to select a feature subset that characterizes best the statistical properties of a target quantification variable, subject to the constraint that these features are mutually as dissimilar to each other and relevant as possible. A filter rather than a wrapper approach is considered here in an attempt to keep computational costs low as well as generalization of the selected features on alternative classifiers or regressors. In the second approach, we utilize a self organizing map (SOM) to perform a feature compression step to investigate its capability to reduce the number of features input to the different calibration models. And third, through the comprehensive experiment considered in this work, we compare the mRMR feature selection, the SOM feature compression, and a two stage feature selection + feature compression algorithm, prior the use of any standard calibration models (e.g., MLR, PCR or PLS) for predicting the concentration of a given set of species within a ternary mixture. Going forward, by attaining "*double citizenship*" in feature selection and feature compression we believe to be uniquely positioned in building more robust, more accurate, and more stable calibration quantification models for the specific application at hand. In the remainder of this manuscript, we will first describe the dataset considered in this work to perform the dimensionality reduction studies and the experimental setup utilized to gather it (Section 2). We will then briefly describe the mRMR, the SOM feature selection schemes, and the combined two-stage feature reduction scheme, followed by presenting the quantification results of the proposed dimension reduction algorithms in predicting the concentrations of the distinct analytes of the unknown mixture (Section 3). And finally, we will present some concluding comments drawn from the results presented in this work (Section 4).

2. Experiments

We apply our proposed dimensionality reduction coupled to regression methods' scheme on an extensive dataset recorded by a FTIR-spectrophotometer. In what follows, we will describe the dataset and then the measurement protocol and software utilized to gather it.

2.1. Apparatus

In gathering our dataset, we utilized a VERTEX 70 series FTIR-spectrophotometer, developed by Bruker Optics, Inc. [35]. The considered system has a wide spectral range, between 9800 and 370 cm^{-1}, standard resolution of 0.5 cm^{-1}, and includes a measurement cell (OMNI-CELL, SPECAC, Inc.), in which the samples to be analyzed are injected using high-precision liquid chromatography (HPLC) syringes. These syringes are accurate within ±1% of nominal volume and with precision within ±1%, measured at 80% of total scale volume. The cell windows are in NaCl, and the optical path is set to 0.5 mm by using Teflon spacers [36]. Thus, this system allows versatility in conveying the chemical information of the sample being detected at the desired concentrations to the sensing cell with high accuracy and in a highly reproducible way.

2.2. Software

The spectrophotometer uses specific software from Bruker Optics Inc. for managing spectra (OPUS Spectroscopy Software) [37]. The calibration models were built and validated using

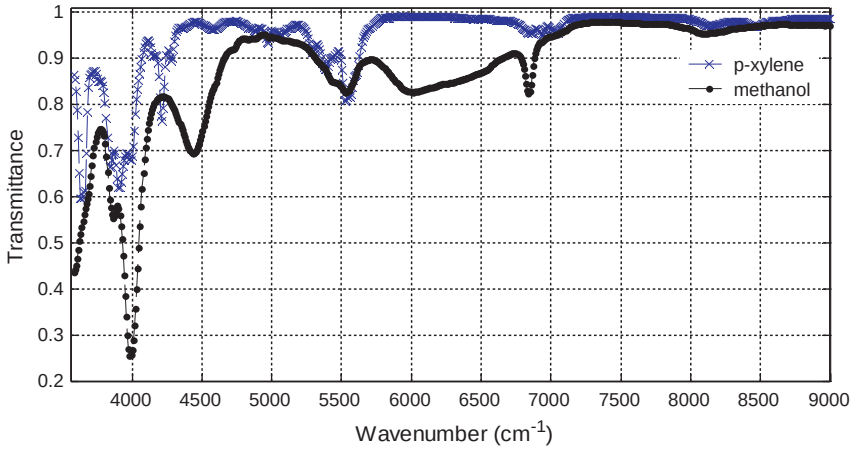

Fig. 1. Typical spectra obtained with the FTIR spectrophotometer for liquid samples containing methanol (10%) or *p*-xylene (10%) in Trichlorotrifluoroethane.

Table 1
Analytes' mixtures and concentrations (in %) covered in the dataset.

Methanol	*p*-Xylene	No. of samples
0	0.1	19
0	0.5	20
0	1	22
0.1	0	20
0.1	0.1	38
0.1	0.5	20
0.1	1	19
0.1	10	20
0.5	0	19
0.5	0.1	38
0.5	0.5	20
0.5	1	19
0.5	10	20
1	0	25
1	0.1	19
1	0.5	19
1	10	22
10	0.1	19
10	0.5	19
10	1	22

standard routines from the PLS-Toolbox® [38] developed by Eigenvector Research Inc. The self organizing feature maps and the mRMR subsets of features were built and validated using standard MATLAB® [39] routines.

2.3. Datasets and procedures

The samples to be analyzed consisted of highly imbalanced binary mixtures of methanol and *p*-xylene diluted in Trichlorotrifluoroethane. The analytes and the solvent were purchased from Sigma–Aldrich and were HPLC grade (their purity was better than 99.9%). Trichlorotrifluoroethane was chosen as solvent because of its inertness and lack of absorption in the NIR region. Table 1 specifies the concentration of each species in the multi-component mixtures.

The dataset[1] used in the work reported here comprises 439 disjoint spectra recorded at different times during five consecutive days, in which each of the samples to be analyzed was prepared independently. In other words, replicate measurements of a given mixture were not performed on aliquot samples but prepared each time mixing its constituents in appropriate quantities. Table 1 specifies the exact concentration and number of replicate measurements taken for every mixture. The raw spectra consisted of the transmittance for wavenumbers between 9800 and 3580 cm^{-1} (recorded at a resolution of 8 cm^{-1}, which gave 777 transmittances per spectra). Fig. 1 shows some typical FTIR spectra for the mentioned analytes' mixtures. The 439 spectra were used to investigate the effects of feature reduction utilizing mRMR feature selection and self organizing maps based feature compression methods on the building and validation of MLR, PCR and PLS models for estimating analyte concentration within the mentioned mixtures. Finally, the choice of these analytes to perform the analytes' mixtures considered in this work was not motivated by any particular reason. The sole peculiarity of the problem addressed in this dataset is its inducement of a non-trivial quantification instance within the selected analytes' mixtures. This dataset is particularly challenging due to the highly imbalanced nature of the concentration values that form every mixture and the overall concentration of the mixture.

3. Dimensionality reduction and experimental results

3.1. The mRMR feature selection

Our goal is to develop a dimensionality reduction method capable of succeeding with our very large dataset with high dimensionality recorded from a FTIR-spectrophotometer.[2] To achieve this goal, our first contribution is a novel formulation, namely the minimum Redundancy-Maximum Relevance (mRMR) principle, a methodology that has previously been successful for a broad range of other quite different applications [40]. The central idea of the mRMR criterion is to find a subset of features, in which the selected promising features jointly have the minimum redundancy (or similarity) on each other and, simultaneously, the largest relevance (or dependency) on the targeted class. Accordingly, assuming an input data D tabled as N samples and M features (i.e., our entire multi-analyte multidimensional dataset that comprises 439 disjoint raw spectra, each of which with 777 transmittances), and the target quantification variable c (i.e., our targeted analyte being quantitated from the mixture), we revisit here the feature selection problem utilizing the mRMR approach on a odor quantification instance as a case of study. The feature selection problem is to find from the M-dimensional observation space, R^M, a subspace of m

[1] The dataset considered in this work is obtainable from the corresponding author upon request.

[2] The loadings of a principal component analysis performed on the data (not shown) confirmed that data were multidimensional.

features, R^m, that characterizes "best" c, in which the best characterization condition means the minimal quantification error rate yielded by the regressor.

The mRMR problem is to maximize a combined single criterion function subject to linear constraints as follows:

$$\max \Theta\ (D, R), \quad \Theta = D - R. \tag{1}$$

Above, the term D is the Maximum Relevance criterion, which seeks to find the feature subset S with m features $\{x_i\}$ that jointly have the maximum dependency on the target class c, thus,

$$\max D\ (S, c), \quad D = I(\{x_i, \quad i = 1, \ldots, m\}; c). \tag{2}$$

In other words, it scores the level of discriminant power of transmittances when they are differentially expressed for different targeted classes. Thus, $I(x_i, c)$ quantifies the relevance x_i for the quantification task. The term R, on the other hand, is the minimum redundancy condition that aims at selecting only the subset of features S that are mutually exclusive, thus,

$$\min R(S), \quad R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j), \tag{3}$$

which in essence $I(x_i, x_j)$ measures the level of similarity between transmittances.

In this paper, we test the feasibility of the mRMR criterion in reducing the dimensionality problem of our dataset, treating both of the above mentioned conditions equally important. Once the combined single criterion from Eq. (1) has been optimized, the components of x represent the weight of each feature. Features with higher weights are better variables to use for the subsequent regression training. Accordingly, an incremental search method can in practice be utilized to find the near-optimal features defined by $\Theta(\cdot)$.

One advantage of the problem formulation above is that it is sufficiently general to permit any symmetric similarity measure to be used. A common choice to measure similarity is the Pearson correlation coefficient ρ. However, it only measures the linear relationship between two random variables, which may not be suitable for some classification or quantification problems. The Mutual Information (MI), in contrast, captures the non-linear dependencies between variables, which is particularly common in the odor quantification task addressed here. Accordingly, in this paper we utilize MI as a symmetric similarity measure. For discrete/categorical variables, the MI between two random variables, say x and y, can formally be defined in terms of their joint probability distribution $p(x, y)$ and the respective marginal probabilities $p(x)$ and $p(y)$ as follows:

$$I(x, y) = \sum_{i,j} p(x_i, y_i) \log \frac{p(x_i, y_i)}{p(x_i)p(y_i)}, \tag{4}$$

where the $I(\cdot)$, as an index measure, provides a ranking of features that takes into account the mutual information between all pairs of features and the relevance of each feature to the class label, simultaneously.

Fig. 2 shows the typical FTIR transmittance spectra for two of the multiple binary mixtures tested: 0.1 and 0.5% of methanol in presence of p-xylene dosed at different concentrations (panel (a)) and 0.1 and 1% of p-xylene in presence of methanol dosed at different concentrations (panel (b)). This figure shows how our proposed mRMR based feature selection scheme captures the transmittances that define the entire analyte-spectrum signature corresponding to each concentration of methanol and p-xylene (panel (a) vs. (b)) irrespective of the concentration of their interferents (i.e., the second constituent analyte of each mixture), thereby promoting the quantification capability of our calibration models. Notice that the region of bands between 6800 and 6900 cm^{-1} represent the change in concentration of methanol regardless the concentration of p-xylene (panel (a)), whereas the region of bands between 5550 and 5700 cm^{-1} are specific to the difference in concentration of p-xylene independently of methanol (panel (b)).

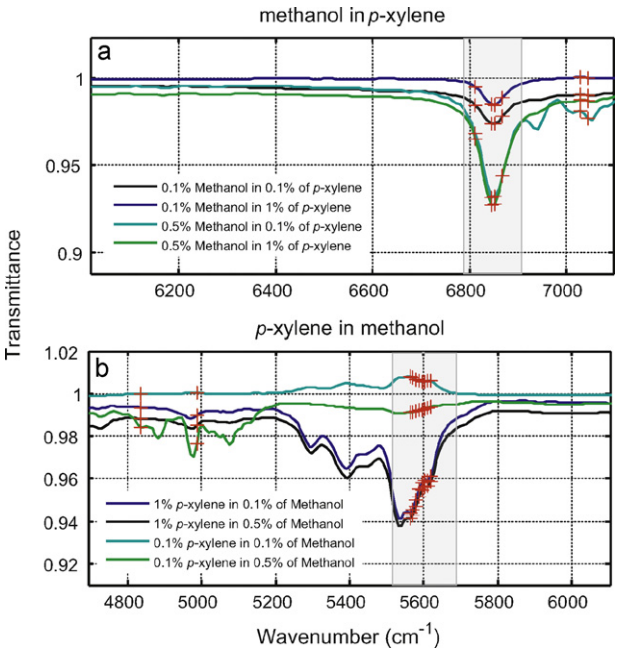

Fig. 2. (a) Typical spectra obtained with the FTIR spectrophotometer for liquid samples containing methanol (0.1% and 0.5%) mixed with different concentration of p-xylene. From panel (a) we can note that region of bands between 6800 and 6900 cm^{-1} (highlighted in gray) represent the change in concentration of methanol regardless the dose of p-xylene. (b) Typical spectra obtained with the FTIR spectrophotometer for liquid samples containing p-xylene (0.1 and 1%) mixed with different concentration of methanol. From panel (b) we can note that the region of bands between 5550 and 5700 cm^{-1} (highlighted in gray) are specific to the difference in concentration of p-xylene independently of the concentrations dosed of the second mixture constituent, i.e., methanol (b).

3.2. The SOM feature compression

Going forward to attaining our dimensionality reduction plan, our second contribution is the self organizing map (SOM), an artificial neural network paradigm that has been widely applied to classify data from multi-sensory systems [41] and to counteract sensor drift [42]. The self organizing map belongs to the category of competitive learning methods with unsupervised training. It performs a topology preserving projection of the data space onto a regular two-dimensional space where similar samples are located together. Accordingly, we utilize the SOM as a feature compressor, in which the full spectra (i.e. 777 features) are input to the network, and the outputs of this network (i.e. transformed features) are fed into the different calibration models. In other words, the MLR, PCR or PLS regression models are not built using the 777 original features of each transmittance spectrum but employing the outputs (i.e. transformed features) of the SOM. Accordingly, since the number of neurons within the two dimensional grid of the SOM is significantly lower than the number of original features in each spectrum, the network can thus be thought of as a feature compressor [43].

3.3. Quantification accuracy results

To evaluate the usability and robustness of the dimensionality reduction methods presented here, we address a chemical analyte mixture quantification problem instance induced by our dataset

described above (see Table 1). Our goal is to assess how much each dimensionality-reduced dataset contribute in the prediction of the analytes' concentrations (i.e., analyte quantification) of each of the constituent species of the highly imbalanced binary mixtures. We are particularly interested in this dataset for their inducement of a non-trivial quantification instance within the highly imbalanced concentration values of the mixtures and the small overall concentration of the mixture. The concentration is a continuous variable, hence the prediction should be made by a regression tool, which takes the features from each created dimensionality-reduced model, and outputs a real number. Our mRMR feature selection method and SOM feature compressor scheme do not convolve with specific regressors. Therefore, we expect the features selected by these schemes have good performance on various types of regression tools. Accordingly, we conducted our validation process by measuring the quantification performance yielded by three, "*gold standard*" in chemometrics, regression tools, namely the principal component regression (PCR) [8], the partial least square (PLS) [9], and the multi-linear regression (MLR) models [12].

We quantified the performance of each created model in the regressor by applying the following training/validation procedure. First, for each individual analyte of the binary mixture, we randomly selected 70% of the resulting 439 spectra recorded data for training the different regression models and kept the remaining 30% for validating them. We then present this randomized split, a batch of data containing 307 training spectra, each of which consisted of 777 transmittances at wavenumbers between 9800 and 3580 cm^{-1}, to each of the above described dimensionality reduction algorithms to compute the feature selection and feature compression processes, individually, creating thereby the different calibration models to be further analyzed (see Fig. 2 for reference on the typical transmittance spectra for methanol and *p*-xylene). For implementing purposes of the mRMR criterion, each variable was discretized in three segments at the positions: $\{(-\infty, \mu, -\sigma), (\mu-\sigma, \mu+\sigma), (\mu+\sigma, \infty)\}$ where it takes -1 if it is less than $(\mu-\sigma)$, 1 if it is larger than $(\mu+\sigma)$, and 0 if otherwise, being μ and σ the mean and standard deviation of training data, respectively. For implementing the SOM network, on the other hand, a set of SOMs, which had 20 through 200 neurons, were trained using the 307 spectra from the training dataset in 200 epochs[3] and employing the default values set in the neural network toolbox of Matlab (i.e., decreasing learning rate and high neighborhood distance during the initial training phase and small learning rate and low neighborhood distance during the fine tuning phase). Once the dimensionality reduction processes were implemented, we utilize their responses to the training spectra to build the MLR, PCR and PLS models, varying the number of features or neurons considered. The resulting response matrices have 307 rows corresponding to the number of spectra considered and a number of columns that equals the number of features selected from the mRMR (from 5 up to 200 features) or the number of neurons employed by each network (from 20 up to 200 neurons). After that, we trained and cross-validated (via leave-one-out cross validation (LOO-CV) process) each calibration model using the training examples to determine the optimal number of factors, e.g., principal components (PC) or latent variables (LV) to be considered in the PCR or PLS regression models, respectively, and then applied the validation batch of recordings to verify the correct concentration prediction rate (i.e., validate our results). The concentration prediction was made by computing the mean squared error of cross validation (MSECV) versus the number of PC and LV used to design the quantification models, respectively. The number

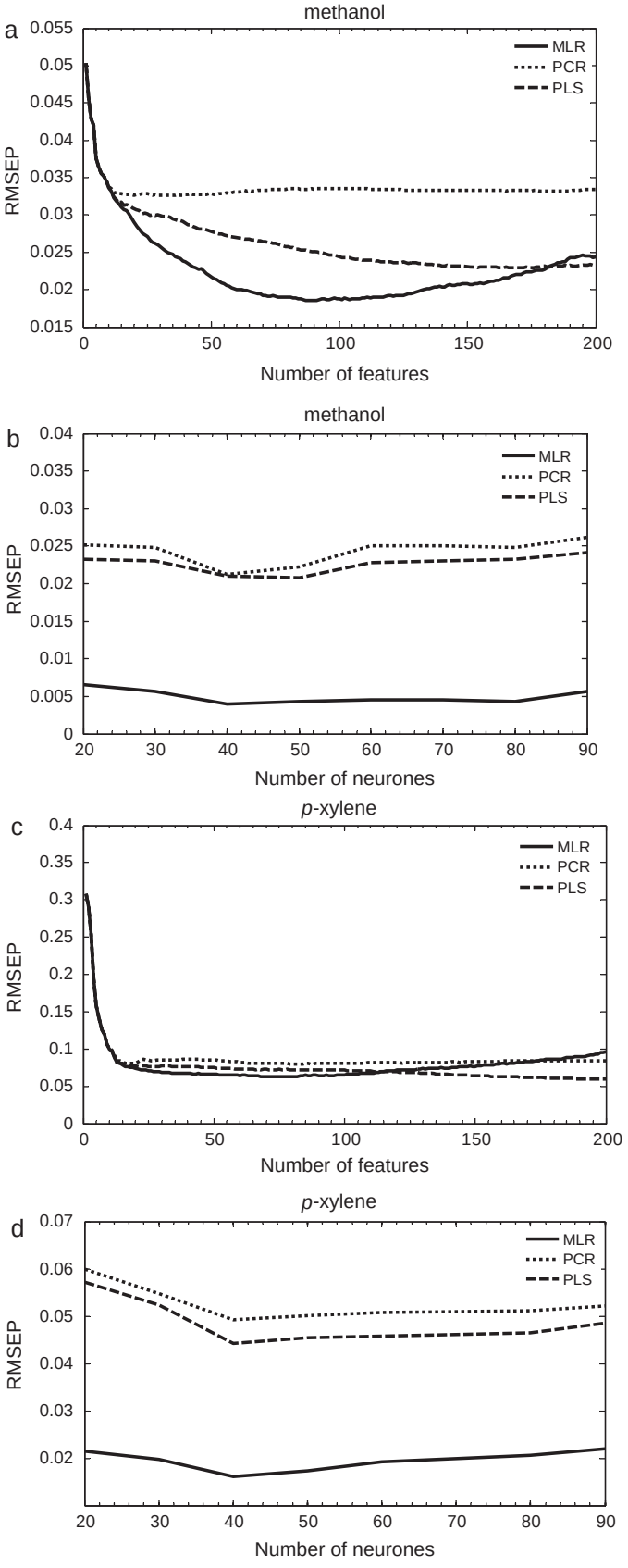

Fig. 3. Evolution of prediction error for the different calibration models as a function of the number of features selected by mRMR versus prediction error for the different calibration models as the number of neurons within the SOM feature compressor. Panel (a) versus panel (b) shows the comparison for methanol. Panel (c) versus panel (d) shows the evolution for *p*-xylene.

[3] The number of training iterations (i.e., 200 epochs) was selected as the minimum number of iterations that led to the lower number of misquantified samples over the training set.

of PC and LV selected was the value after the first sharp decrease in MSECV. This strategy for determining the optimal number of factors is often employed to fight the risk of over fitting [9]. Once the number of PC and LV was determined, we then build the quantification models (one for each analyte), and validated the calibration models, using the 132 spectra data that had been held out (i.e. those that belonged to the validation datasets), by calculating the mean squared error of prediction (MSEPRED) as a function of the number of features or number of neurons[4] considered for each calibration model employed. Finally, we repeat this randomized training/validation episode 50 times with different random splits of the labeled data and estimate the average correct quantification rate. The sole purpose of these 50 runs was to consistently evaluate the robustness of our dimensionality reduction methods.

Fig. 3 shows the evolution of the prediction error (for validation measurements only) as a function of the number of features used versus the evolution of the prediction error as a function of the number of competitive neurons within the SOM for every chemical analyte being analyzed within the mixture with respect to each regression tool (methanol: panel (a) vs. panel (b); *p*-xylene: panel (c) vs. panel (d)). Note that each profile corresponding to every regressor in Fig. 3 (a) and (c) follows a similar pattern along the different number of features considered when the mRMR feature selector is implemented, in the sense that their extreme points (i.e., the lowest prediction errors) occur at the same values. The ordering of these points in magnitude gradient is also preserved to a large extent across the different features considered, being the MLR always the winner in yielding the best quantification performance among the three calibration regression models. Later on this paper, by employing a different validation strategy, it will be shown that the better performance of MLR is only apparent, since its generalization ability is poorer than that of PCR or PLS. Fig. 3 panels (b) and (d), respectively, shows, on the other hand, the evolution of prediction errors for SOMs employing between 20 and 90 neurons. As observed in these figures, the values of prediction errors remain quite stable for all the different number of neurons considered up to 80 neurons and then start to rise as the number of neurons increases. When a 40-neuron SOM is used, the mean squared errors of prediction reach their minimum values (i.e., 4×10^{-3} for MLR, 2.12×10^{-2} for PCR and 2.10×10^{-2} for PLS). The average prediction rates over the 50 training/validation split trials are listed in Tables 2 and 3 for each comprised analyte and stand-alone dimensionality reduction method used and compared against the performance attained when no-reduction method was used (Table 4). More specifically, each Table illustrates the mean values of the MSEPRED, the slopes (m), intercepts (i) and the correlation coefficients (R) of the linear regressions between the actual and predicted concentrations, where the closer to one are the slopes and correlation coefficients and the closer to zero is the MSEPRED and intercepts, the better the quantification calibration models are. As the results indicate, both pre-processing methods boost up the prediction ability of the different models, being the SOM feature compression method the one that performs better by a slight margin not only in terms of performance but also avoiding the burden of envisaging a time-consuming variable selection procedure (see Table 2 vs. Table 3, respectively). Moreover, for PCR and PLS, prediction errors are about 30% lower than those obtained when no feature reduction pre-processing is used to build the models (i.e. errors shown in Table 4).[5] These results clearly indicate that our reduction methods improve the estimation of the analytes' concentrations, and even more importantly, it prevents form collinearity problems presented by the MLR models when being built.

Fig. 4 shows the evolution of the number of miscategorized (i.e., misquantified) samples for SOM networks employing between 20 and 90 neurons. Confusions occur when a given neuron is the winner for spectra that belong to different methanol or *p*-xylene concentrations. Increasing the number of neurons reduces the number of misquantified spectra but at the cost of using a higher number of uncommitted neurons (i.e. neurons that never win during the training process). The weights of uncommitted neurons at the end of the training phase strongly depend on the values initially assigned to them. In other words, since these neurons do not adapt their weights to code a given category, weight values remain close to their initial values, which are selected at random. This is detrimental for the correct estimation of analyte concentrations by the different calibration models.

In evaluating the ability of each pre-processing feature reduction method, a second and more challenging training-test/validation procedure is also envisaged in this work. In this second approach, for each analyte, we partitioned the above described dataset in two parts according to their concentration values. This is, the first part selected a batch of data containing measurements of methanol or *p*-xylene dosed at concentration values different from 1% for training, and held out then the remaining part of the recordings, i.e., those measurements that contain the concentration values dosed at 1%, to quantify the ability of the method to interpolate the unknown concentrations (i.e., validate); see Table 1 for reference on the different concentration values dosed. This validation strategy is more challenging, since the quantitative models built are asked to predict concentrations that do not occur in the first training subset, thereby enabling us to carefully verify the generalization ability of the dimensionality-reduced models considered here.

Thus, for the first sub-dataset part, once again we randomly split it into 70% training and 30% test subsets by a procedure equivalent to the one described above, where this new training subset part was utilized to implement the dimensionality reduction methods and form the calibration models. Then, via the LOO-CV procedure, we trained and cross-validated each calibration model using the dimensionality-reduced training examples, determining thus the optimum number of factors or latent variables to build each calibration model. As in the previous case, the concentration prediction was made by computing the mean squared error of cross validation (MSECV) versus the optimum number of factors or latent variables used to design the quantification models, in which the number of factors selected was the value after the first sharp decrease in MSECV. After that, we applied the part of recordings that had been held out for test to verify the correct concentration prediction rate, and calculated the test mean squared error of prediction (MSEPREDtst). Finally, using the second batch of recordings, the final validation was performed i.e., the interpolation of the unknown concentration values. Thus, utilizing the calibration models calculated above, we compute the validation mean squared error of prediction (MSEPRED) and predict the concentration values of the recordings destined for validation (i.e., second part of the sub-dataset). We repeated this randomized training-test/validation split episode 50 times, which enable us to carefully check the ability of the optimized calibration models to generalize.

[4] The validation of the SOM network is carried out by projecting the validation spectra onto the space of the compressed features, multiplying each spectrum by the neurons' weights. The resulting projected data is then multiplied by the regression vectors of each model to obtain the estimation of the analytes' concentrations of the 132 spectra being validated.

[5] The modeling employing MLR failed due to the high degree of collinearity existing among transmittances at close wavenumbers; therefore, only PCR and PLS models were actually built when full spectra were used.

Table 2
Quantification performance (132-spectra validation dataset) of the MLR, PCR and PLS models employing the first 40 features selected by mRMR. The number of PCs or LVs employed by PCR or PLS models, respectively is indicated.

Product	Model	Factors	Corrcoef.	Slope	Intercept	MSEPRED
Methanol	MLR	–	0.993	1	−0.13	0.0223
	PCR	3	0.989	1	−0.91	0.0325
	PLS	3	0.991	1	−0.56	0.0272
p-Xylene	MLR	–	0.991	1	−0.45	0.0261
	PCR	3	0.991	1	−0.45	0.0272
	PLS	3	0.991	1	−0.38	0.0268

Table 3
Quantification performance (132-spectra validation dataset) of the MLR, PCR and PLS models employing a 40-neuron SOM for feature compression. The number of PCs or LVs employed by PCR or PLS models, respectively is indicated.

Product	Model	Factors	Corrcoef.	Slope	Intercept	MSEPRED
Methanol	MLR	–	0.999	1	−0.017	0.0040
	PCR	3	0.997	1	−0.10	0.0212
	PLS	3	0.997	1	−0.12	0.0210
p-Xylene	MLR	–	0.995	1	−0.025	0.0061
	PCR	3	0.992	1	−0.23	0.0242
	PLS	3	0.992	1	−0.83	0.0242

Table 4
Quantification performance (132-spectra validation dataset) of the PCR and PLS models without feature compression or feature selection. The number of PCs or LVs employed by PCR or PLS models, respectively is indicated.

Product	Model	Factors	Corrcoef.	Slope	Intercept	MSEPRED
Methanol	PCR	7	0.982	1.00	−1.9	0.0301
	PLS	6	0.985	1.00	−1.9	0.0289
p-Xylene	PCR	7	0.980	0.99	−0.82	0.0322
	PLS	6	0.981	0.99	−0.83	0.0319

Table 5
Mean squared error of prediction of the MLR, PCR and PLS models for validation samples in the second[a] training/validation approach employing the first 40 features selected by mRMR. The number of PCs or LVs employed by PCR or PLS models, respectively is indicated.

Product	Model	Factors	MSEPRED
Methanol	MLR	–	0.1034
	PCR	3	0.0800
	PLS	3	0.0786
p-Xylene	MLR	–	0.1298
	PCR	3	0.0834
	PLS	3	0.0821

[a] In this validation approach the model is validated with a concentration that was not found in the training set (more challenging since the model is asked to interpolate).

Table 6
Mean squared error of prediction of the MLR, PCR and PLS models for validation samples in the second[a] training/validation approach. A 40-neuron SOM is used for feature compression. The number of PCs or LVs employed by PCR or PLS models, respectively is indicated.

Product	Model	Factors	MSEPRED
Methanol	MLR	–	0.0970
	PCR	3	0.0634
	PLS	3	0.0632
p-Xylene	MLR	–	0.1031
	PCR	3	0.0647
	PLS	3	0.0645

[a] In this validation approach the model is validated with measurements of a single concentration (i.e., 1%) that was not found in the training set (more challenging since the model is asked to interpolate).

The average prediction rates (MSEPRED) for the validation samples over these trials are presented in Tables 5 and 6. From Table 5, it can be derived that, irrespective of the regression tool used (i.e., PCR, PLS, or MLR), it is still manageable to perform accurate predictions of a concentration that was not in the training phase. It is important to notice, however, that PLS and PCR clearly outperform MLR in this second validation approach. This clearly indicates that MLR tends to over fit training data and does not generalize well. Roughly, an 8% error in the prediction of methanol or *p*-xylene concentration is reached when utilizing the mRMR approach. Table 6, on the other hand, shows a 6% error in the prediction of methanol or *p*-xylene concentration reached when the SOM feature compressor method is utilized. As the results indicate, with this more challenging validation procedure the error in the estimation of methanol or *p*-xylene concentration increased with respect to the previous validation scheme, but it still compares favorably to the 10% error obtained when no feature reduction step was present (see Table 7 for comparison). All these errors, though, are relative to the actual species concentration. These average obtained results shown in these tables indicate, nonetheless, that an accurate estimation of gas concentration is still possible even when the

Table 7
Mean squared error of prediction of the PCR and PLS models without feature compression for validation samples in the second training[a]/validation approach. The number of PCs or LVs employed by PCR or PLS models, respectively is indicated.

Product	Model	Factors	MSEPRED
Methanol	PCR	7	0.0907
	PLS	6	0.0926
p-Xylene	PCR	7	0.0947
	PLS	6	0.0934

[a] In this validation approach the model is validated with measurements of a single concentration (i.e., 1%) that was not found in the training set (more challenging since the model is asked to interpolate).

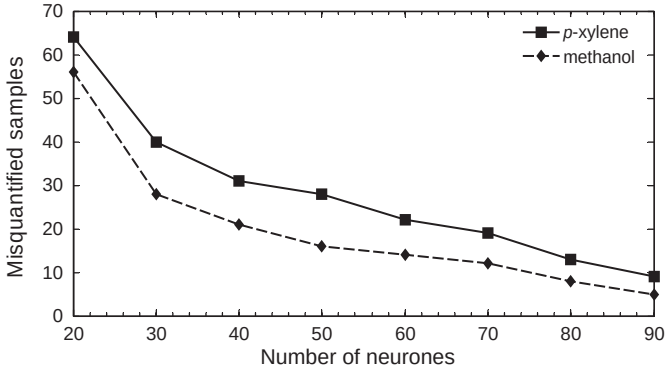

Fig. 4. Evolution of misquantified samples according to the number of neurons utilized.

Table 8
Quantification performances (132-spectra validation dataset) of the MLR, PCR and PLS models employing the first 80 features selected by mRMR and a 16-neuron SOM[a] for feature compression. The number of PCs or LVs employed by PCR or PLS models, respectively is indicated.

Product	Model	Factors	Corrcoef.	Slope	Intercept	MSEPRED
Methanol	MLR	–	0.999	1	−0.012	0.0038
	PCR	3	0.995	1	−0.10	0.0202
	PLS	2	0.997	1	−0.09	0.0130
p-Xylene	MLR	–	0.997	1	−0.25	0.0050
	PCR	3	0.994	1	−0.31	0.0221
	PLS	2	0.995	1	−0.29	0.0202

[a] Increasing the number of neurons within the SOM (i.e., more than 16) does not improve results.

quantitative models built are asked to predict concentrations that do not occur during their training, which clearly demonstrates the generalization ability of the method.

3.4. The combined mRMR and SOM approach

Our goal is to design efficient algorithms to select a compact set of features. In the previous section, we proposed two dimensionality reduction schemes namely, the mRMR feature selection method and our new SOM-based feature compression scheme. Having established the strong capabilities of these both schemes working separately to quantitatively analyze the binary mixtures, our third contribution in the remainder of this paper is to assess how much these two criteria when coupled together contribute to the prediction of the analytes' concentration (i.e., analyte quantification) of our dataset described above (see Table 1). Thus, we present here a two-stage feature selection algorithm. In particular, our intention is to utilize mRMR feature selection in the first-stage to find a small set of candidate features, in which the SOM feature compressor can be applied as second stage. By doing this, we investigate how much the space of candidate features selected by mRMR facilitates the integration of the proposed feature compression scheme to find a compact, yet superior, subset of features that will potentially lead to a more effective way to address the quantification instance problem presented here at a much lower computational cost.

To implement our two-stage feature selection scheme and quantify, thus, the performance of the created model, we apply the following procedure: (i) selection of the optimal set of candidate features. In doing this, we first compute the mRMR incremental feature selection scheme, which leads to n sequential subsets of features, thus $S_1 \subset S_2 \subset \cdots \subset S_n$.[6] We then compare all the n sequential subsets of features $S_1, \ldots, S_n$, $(1 \leq k \leq n)$ to find the range k of small prediction error (here denoted as X), within which the respective cross-validation classification error (calculated via the LOO-CV process described in the previous sub-section) is consistently small. Then, within X (i.e., a relatively stable range of small prediction error), the optimal size of the candidate feature set, n^*, is chosen as the smallest number of features that corresponds to the smallest prediction error found. And (ii), given a more compact optimized set of features n^*, we implement the SOM network as a feature compressor scheme. In doing so, we take the set of n^* features and compute our SOM-based feature compression scheme in an incremental manner, leading to creating different n sequential models according to the number of neurons considered. We then estimate the cross-validation quantification prediction error on each created model. Once again we conduct our validation process by measuring the quantification performance yielded by our regression mechanisms following the training-validation procedures described in the previous subsection, and select the model that leads to the smallest prediction error among them. Finally, we reproduced the whole process 50 times to generalize, utilizing different randomized training-test/validation splits on each trial.

Table 9
Mean squared error of prediction of the MLR, PCR and PLS models for validation samples in the second[a] training/validation approach employing the first 80 features selected by mRMR and a 16-neuron SOM[b] for feature compression. The number of PCs or LVs employed by PCR or PLS models, respectively is indicated.

Product	Model	Factors	MSEPRED
Methanol	MLR	–	0.0914
	PCR	3	0.0615
	PLS	2	0.0598
p-Xylene	MLR	–	0.1023
	PCR	3	0.0621
	PLS	2	0.0615

[a] In this validation approach the model is validated with measurements of a single concentration (i.e., 1%) that was not found in the training set (more challenging since the model is asked to interpolate).
[b] Increasing the number of neurons within the SOM (i.e., more than 16) does not improve results.

The average prediction rates (MSEPRED) for the validation samples of each comprised analyte over the 50 trials utilizing the two-stage feature selection scheme are presented in Tables 8 and 9. Table 8 shows, on the one hand, the quantification performances (132-spectra validation dataset) of the MLR, PCR and PLS models employing the first 80 features selected by mRMR and a 16-neuron SOM for feature compression. As the results indicate, coupling both pre-processing methods together enhance the prediction ability of the different calibration models considered. In all cases the prediction errors are about 10% lower than those obtained when each feature reduction pre-processing methods are used individually (see Table 8 vs. Tables 2 and 3 for comparison). These results clearly indicate that our two-stage reduction method improve the estimation of the analytes' concentrations. Table 9, on the other hand, illustrates the prediction results of the second validation approach in which spectra of 1% concentration of methanol and *p*-xylene integrated the validation set and all remaining spectra integrated the training set. With this more challenging validation procedure the error in the estimation of methanol or *p*-xylene concentration increased to about 5% for both PCR or PLS models. This is understandable, though, since in this case the quantitative models built are asked to predict concentrations that do not occur

[6] For a full description of the mRMR feature selection process the reader is referred to Section 3.1 of this manuscript.

during their training stage. Yet it still compares favorably to the 10% error obtained when no feature reduction step was present (see also Table 4 for comparison) or the 6% error obtained when each feature reduction is used separately. All these errors are relative to the actual species concentration. In this second validation approach, the prediction errors associated to MLR models were significantly higher than those of PCR or PLS models and this could be due to MLR being more prone to over fitting the data.

4. Conclusions

FTIR-spectrophotometry is becoming increasingly utilized by the chemical industry to online monitor the concentration of reagents and by-products. FTIR spectra are characterized by a high number of variables, some of which are highly correlated and others are affected by noise. Here, we have reported the significant benefits of utilizing two simple methods to pre-process spectra prior to use standard calibration models such as MLR, PCR or PLS. On the one hand, we have stressed that a well-designed filter, such as the minimum-Redundancy Maximum-Relevance (mRMR), can be utilized to select a subset of relevant features from FTIR spectra that best characterizes the statistical properties of a target quantification variable. The particularity of the mRMR scheme studied here, is that is does not intend to select features that are independent of each other. In steady, it tries to select the features that minimize the redundancy and simultaneously maximize their relevance with respect to the targeted analyte. A self organizing map, on the other hand, was utilized to perform a feature compression step, in which the number of features input to the different calibration models is reduced and fetch to the classifier/regressor. As demonstrated, this method avoids the burden of envisaging a time-consuming variable selection procedure and does not require previous knowledge about the regions of the spectra that contain relevant variables or information. However, both methods make it possible to build more parsimonious models (i.e. using fewer variables), which are more accurate and more robust (i.e., they generalize better than less parsimonious models). In addition to this, we have demonstrated how combining both the mRMR and SOM-based network approaches as a two-stage feature selection scheme provides a better way to maximize the performance of our methods. In particular, our experimental results have shown the benefit of utilizing these methods in the quantification of methanol and *p*-xylene mixtures dissolved in Trichlorotrifluoroethane. As a final remark of this paper, we want to emphasize that these techniques could be readily exported to other multi-sensor paradigms, in which a high number of features per measurement are available, such as, for example, the identification and localization of chemical analytes in the wind-tunnel utilizing a multi-dimensional sensor array, gas distribution mapping, or gas plume tracking with a robotic platform. We believe, though, that due to the inherent complexity of these examples, those analyses could be the object of a completely new piece of research that we seek to address in further works.

Acknowledgments

This work was funded in part by NATO under the Science for Peace Program grant number CBP.MD.CLG 983914. A. Vergara is funded by the U.S. Office of Naval Research (ONR), contract number N00014-07-1-0741; by the Jet Propulsion Laboratory, contract number 2010-1396686; and by the US Army Medical Research and Materiel Command and the United States Army Research Institute of Environmental Medicine (USARIEM), contract number W81XWH-10-C-0040 in collaboration with Elintrix Inc. E. Llobet is supported by the Spanish Ministry of Science and Innovation and the Catalan Agency for Research under the grant numbers TEC 2009-07107 and 2009 SGR 789, respectively. The authors thank D. Vargas and J. Martín for performing the measurements with the FTIR-spectrophotometer, J. Ferre-Borrull for a helpful discussion and to Joanna Zytkowicz for reading and revising the manuscript.

References

[1] J. Chalmers, P.R. Griffiths (Eds.), Handbook of Vibrational Spectroscopy, vols. 1–5, Wiley & Sons, Chichester, 2001.
[2] E. Smidt, M. Schwanninger, Spectrosc. Lett. 38 (2005) 247.
[3] E. Smidt, K.U. Eckhardt, P. Lechner, H.R. Schulten, P. Leinweber, Biodegradation 16 (2005) 67.
[4] P. Zaccheo, G. Ricca, L. Crippa, Compost Sci. Util. 10 (2002) 29.
[5] M. Grube, J.G. Lin, P.H. Lee, S. Kokorevicha, Geoderma 130 (2006) 324.
[6] H. Günzler, H.M. Heise, H.-U. Gremlich, IR Spectroscopy, Wiley-VCH, Weinheim, 2002.
[7] P.R. Griffiths, P.J.A. de Hasseth (Eds.), Fourier Transform Infrared Spectrometry, 2nd ed., Wiley-Blackwell, 2007.
[8] T. Naes, H. Martens, J. Chemometr. 2 (1988) 155.
[9] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1.
[10] A.S. Barros, D.N. Rutledge, Chemometr. Intell. Lab. Syst. 40 (1998) 65.
[11] U. Depczynski, V.J. Frost, K. Molt, Anal. Chim. Acta 420 (2000) 217.
[12] H. Martens, T. Naes, Multivariate Calibration, Wiley, London, 1993.
[13] R. Leardi, R. Boggia, M. Terrile, J. Chemometr. 6 (1992) 267.
[14] R. Leardi, J. Chemometr. 8 (1994) 65.
[15] C.B. Lucasius, M.L.M. Beckers, G. Kateman, Anal. Chim. Acta 286 (1994) 135.
[16] D. Jouan-Rimbaud, D.L. Massart, R. Leardi, O.E. Noord, Anal. Chem. 67 (1995) 4295.
[17] J. Sun, J. Chemometr. 9 (1995) 21.
[18] N.R. Draper, H. Smith, Applied Regression Analysis, 2nd ed., Wiley, New York, 1981.
[19] K. Sasaki, S. Kawata, S. Minami, Appl. Spectrosc. 40 (1986) 185.
[20] J.H. Kalivas, N. Roberts, J.M. Sutter, Anal. Chem. 61 (1989) 2024.
[21] J.K. Amamcharla, S. Panigrahia, C.M. Logue, M. Marchello, J.S. Sherwood, Biosyst. Eng. 107 (1) (2010) 1.
[22] K. Meissl, E. Smidt, M. Schwanninger, Talanta 72 (2) (2007) 791.
[23] E. Llobet, J. Brezmes, O. Gualdrón, X. Vilanova, X. Correig, Sens. Actuators B 99 (2004) 267.
[24] I. Guyon, J. Mach. Learn. Res. 3 (2003) 1157.
[25] A.L. Blum, P.P. Langely, Artif. Intell. 97 (1997) 245.
[26] R. Bekkerman, N. Tishby, Y. Winter, I. Guyon, A. Elisseeff, J. Mach. Learn. Res. 3 (2003) 1183.
[27] G. Forman, CIKM'08: Proceeding of the 17th ACM Conference on Information and Knowledge Mining, New York, NY, USA, 2008, p. 263.
[28] G. Forman, J. Mach. Learn. Res. 3 (2003) 1289.
[29] G.H. John, R. Kohavi, K. Pfleger, Machine Learning: Proceedings of the Eleventh International Conference, San Francisco, 1994.
[30] R. Kohalvi, G.H. John, Artif. Intell. 97 (1–2) (1997) 273.
[31] P. Langley, Proceedings of the AAAI Fall Symposium on Relevance, AAAI Press, 1994, p. 140.
[32] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Chapman & Hall/CRC, 1984.
[33] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik, Advances in Neural Information Proceeding Systems (NIPS) 13, MIT Press, 2001, p. 668.
[34] D. Jouan-Rimbaud, D.L. Massart, O.E. Noord, Chemometr. Intell. Lab. Syst. 35 (1996) 213.
[35] Bruker Optics Inc. http://www.brukeroptics.com/vertex.html.
[36] SPECAC Inc. http://www.specac.com/products/liquid-transmission-cell/ft-irliquid-transmission-cell/530.
[37] Bruker Optics Inc. http://www.brukeroptics.com/opus.html.
[38] Eigenvector Research, Inc., PLS_Toolbox. Version 5.2.1, 2009.
[39] Matlab User's Guide, The Mathworks Inc., 2009.
[40] H. Peng, F. Long, C. Ding, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226.
[41] C. Di Natale, A. Macagnano, A. D'Amico, F. Davide, Meas. Sci. Technol. 8 (1997) 1236.
[42] M. Zuppa, C. Distante, P. Siciliano, K.C. Persaud, Sens. Actuators B 98 (2004) 305.
[43] E. Llobet, M. Anaimi, A. Pruñonosa, E. Gras, Sens. Actuators B 158 (2011) 252.

Alexander Vergara (PhD, 2006 – Universitat Rovira i Virgili, Tarragona, Spain) is Postdoctoral Scientist Associate at the BioCircuits Institute, UC San Diego (USA). His work mainly focuses on the use of dynamic methods for the optimization of micro gas-sensory systems and on the building of autonomous vehicles that can localize odor sources through a process resembling the biological olfactory processing. His areas of interest also include signal processing, pattern recognition, feature extraction, chemical sensor arrays, and machine olfaction.

Eduard Llobet (PhD, 1997 – Universitat Politecnica de Catalunya, Barcelona, Spain) is full Professor at the Electronic Engineering Department of the Universitat Rovira i Virgili, Tarragona (Spain), and Director of the Research Centre on Engineering of Materials and micro/nano-systems (EMaS). His main areas of interest are in the design of nano-structured semiconductor and carbon nanotube based gas sensors and in the application of intelligent systems to complex odor analysis. Prof. Llobet is also a Senior Member of the IEEE.