Incremental model learning for spectroscopy-based food analysis

Diaz, K., Georgouli, K., Koidis, A., & Martinez Del Rincon, J. (2017). Incremental model learning for spectroscopy-based food analysis. *Chemometrics and Intelligent Laboratory Systems*, *167*, 123-131. https://doi.org/10.1016/j.chemolab.2017.06.002

Published in:
Chemometrics and Intelligent Laboratory Systems

Document Version:
Peer reviewed version

Incremental model learning for spectroscopy-based food analysis

Katerine Diaz[a], Konstantia Georgouli[b], Anastasios Koidis[b], Jesus Martinez Del Rincon[c,*]

[a] *Computer Vision Centre, Universidad Autonoma de Barcelona, Spain*
[b] *Institute for Global Food Security, Queens University Belfast, UK*
[c] *Institute of Electronics, Communications and Information Technology, Queens University Belfast, UK*

Abstract

In this paper we propose the use of incremental learning for creating and improving multivariate analysis models in the field of chemometrics of spectral data. As main advantages, our proposed incremental subspace-based learning allows creating models faster, progressively improving previously created models and sharing them between laboratories and institutions without requiring transferring or disclosing individual spectra samples. In particular, our approach allows to improve the generalization and adaptability of previously generated models with a few new spectral samples to be applicable to real-world situations. The potential of our approach is demonstrated using vegetable oil type identification based on spectroscopic data as case study. Results show how incremental models maintain the accuracy of batch learning methodologies while reducing their computational cost and handicaps.

Keywords: Incremental model learning, IGDCV technique, Subspace based learning, Identification, Vegetable oils, FT-IR spectroscopy

2010 MSC: 00-01, 99-00

*Corresponding author

Email addresses: kdiaz@cvc.uab.es (Katerine Diaz), kgeorgouli01@qub.ac.uk (Konstantia Georgouli), t.koidis@qub.ac.uk (Anastasios Koidis), j.martinez-del-rincon@qub.ac.uk (Jesus Martinez Del Rincon)

1. Introduction

In the last decade the use of chemometrics in food analysis is steadily growing. This is caused because the output of most analytical methods is nowadays multivariate data matrices (spectroscopic, chromatographic/mass spectrometry data, isotopic, sensorial, etc) which cannot be manually analysed and demand appropriate chemometric analysis in order to process and capture the most important and relevant information in the data. Selection of multivariate methods (e.g. classification methods) however is often limited to a set of well known standard methods (e.g. PLS-DA and SIMCA classification methods) and researchers are faced with some persisting problem with the chemometric models that they generate [1].

Among these problems that must be addressed, the generality of the models created to new conditions is the most important one. While extensive research has been done to create models under controlled conditions, for a small problem or dataset, the applicability of those models in real world -e.g. in food testing in the food industry or in routine analysis in a regulated testing laboratory- is very scarce. This is due to the overfitting of the model to the calibration set when only one instrument, one analytical laboratory or, in general, one set of assumptions are taken into consideration to create the models. Thus, when these models are tested in other slightly different conditions, they report much lower performances than the expected one. Recalibrating or recreating similar models to work in those situations may be an extremely arduous task, with a similar time and effort scale to the design, and tuning of the first model.

To avoid a full recalibration, model updating and calibration transfer techniques have been proposed to cover the transfer of multivariate classification models between different spectrometers [2, 3], temperatures [3, 4], harvesting seasons [4] and even different geographical regions [5]. Calibration transfer techniques [2] allow mapping the new spectra to the primary model spectra domain by calculating a transformation matrix from one domain to the other. Different calibration transfer techniques have been recently explored in chemical sensor

arrays to overcome inherent sensor variability [6, 7, 8]. Only a small set of samples are required to be measured in both the primary and secondary conditions. However, in many applications it is not realistic that exactly the same sample can be measured, e.g. the same food sample from two different geographical locations. More interesting are methods based on model updating by augmenting sample spectra from a new condition. While many sample would normally be required to span to the new conditions [4], which amounts to a full recalibration, approaches based on Tikhonov regularisation (TR) [3, 5] only needs a few samples to update the model. As disadvantage, TR still requires access to the initial samples to recompute the updated model, with the consequent computational cost of involving all samples in the optimisation, and its performance heavily relies on a meta-parameter that controls the balance between the initial model and the augmented samples, and which can only be tuned empirically. Finally, some recursive learning approaches [9, 10] propose a framework where both incremental and decremental stages are used to improve the initial model. However, to fully exploit their potential and being able to remove old samples, access to the initial samples is also required.

Moreover, new samples are analysed on a routine basis and new data is generated including cases when new component classes are needed to be created (in authentication/adulteration studies, in traceability, proximate analysis prediction etc). As a result, existing and validated models may stop being useful and/or applicable. It is then necessary to retrain them. However, this requires access to the original samples, which may be lost or unavailable. Similarly, if an external laboratory, or other third party such as a company or an institution wishes to improve an existing model, the access to the original samples may be tricky or impossible, with privacy or confidentiality issues playing a role. In all these previously described situations, it is clear that evolving a chemometric model may be a better solution than recreating or retraining it as a full new batch. This will only require access to the existing models and the new samples. It will also be a more efficient manner to store the information, reducing the memory and physical space required and it can potentially decrease the time to

create an improved model.

While incremental learning has been used and proposed in other fields [11, 12, 13, 9, 10], its intrinsic advantages have been scarcely exploited in the field of food analysis and chemometrics [14, 15, 16, 17, 18, 19]. Bhattacharyya et al. [14, 15] applied neural networks for identification of seven different black tea classes. Their incremental approach allow to add new classes of black tea to the original set. In Tudu et al. (2009) [16], the same researchers applied incremental fuzzy logic to the black tea identification. Cernuda et al. [17, 18, 19] proposed a flexible fuzzy inference system for the monitor of the concentration of sulphuric acid (H_2SO_4), sodium sulfate (Na_2SO_4) and zinc sulfate ($ZnSO_4$) in viscose production and in the melamine resin production process, which allows online adaptation of parameters and structural changes in the model. However, techniques based on neural networks and fuzzy logic are scarcely used in food science, reducing the impact of these incremental approaches, and they require huge amounts of calibration samples to generate the calibration models, which is unlikely for most food analysis scenarios.

In this paper we aim to extend the use of incremental learning in the field of food analysis and chemometrics. Among the variety of incremental learning techniques, we have chosen subspace based learning as the family of machine learning to apply due to their proved ability to evolve online [13], the ability to generate efficient models using a reduced number of calibration samples, and the extensive use of some of the basic subspace based methods such as Principal Component Analysis (PCA), and Soft independent modelling of class analogies (SIMCA)- in food science [20, 21], both for exploratory analysis [22] and classification [23, 24, 25]. Thus, the present work introduces the use of an incremental subspace based learning technique, called Incremental Generalized Discriminative Common Vectors (IGDCV), which allows efficiently adding new data samples and classes to a knowledge base. In this way, our methodology is able to update the model to the new scenario without recalculating the full projection or accessing the previously processed calibration data, while retaining the previously acquired knowledge. Our approach is evaluated using vegetable

oil type identification [22, 26, 27, 28] as case study and results are compared against a non incremental learning technique, i.e. an equivalent batch method. Three different incremental scenarios are tested in this application area: when new samples are available to improve the model, when new classes must be identified by the model, and when new instruments are used in the identification process.

2. Incremental Learning Framework

Several incremental feature extraction based on linear subspace methods have been proposed and used on many practical applications. Among them, we find the Incremental approaches of the PCA [29], Linear Discriminant Analysis (LDA) [30] and DCV [31]. While PCA-based incremental approaches are simple and versatile, they are not optimal for discrimination and classification purposes since no class information is used to obtain principal components which may lead to unsuited subspaces. On the contrary, LDA is a supervised technique which makes use of the class information to obtain the most discriminative space by maximizing the distance between classes while minimizing the distance between the samples within the same class. However, LDA-based approaches cannot be applied when the dimension of the sample space is larger than the number of samples in the calibration set, since the within-class scatter matrix will be singular. This problem is known as the *Small Sample Size* SSS problem [32], and it is frequent in spectroscopic and chromatographic application, where the number of variables per sample is in the order of thousands while the total number of samples used for calibration rarely goes above the hundreds [22].

Among the approaches that have been proposed to solve the SSS problem, the Generalized Discriminative Common Vectors (GDCV) has been proved [13] to provide discriminative subspaces for classification regardless of the SSS assumption. GDCV is a variation of LDA [33, 34] which introduces the idea of approximate extended null and reduced range subspaces of the within-class scatter matrix. Given the good performance of GDCV batch approaches, we

proposed the use of Incremental GDCV [13] as the base of our online learning framework for food analysis, where new information is added while retaining the previously acquired knowledge, without accessing the previously processed calibration data.

2.1. IGDCV

Formally, let the calibration set X be composed of c classes, where every class j has m_j samples. The total number of samples in the calibration set is $M = \sum_{j=1}^{c} m_j$. Let x_j^i be a d-dimensional column vector which denotes the i^{th} sample from the j^{th} class. The within-class scatter matrix, S_w^X, is defined as,

$$S_w^X = \sum_{j=1}^{c} \sum_{i=1}^{m_j} (x_j^i - \overline{x}_j)(x_j^i - \overline{x}_j)^T = X_c {X_c}^T \tag{1}$$

where $\overline{x}_j$ is the average of the samples in the j^{th} class, and the centered data matrix, X_c consists of column vectors $(x_j^i - \overline{x}_j)$ for all $j = 1 \dots c$ and $i = 1 \dots m_j$.

The extension of the null space of S_w^X (which implies restricting the corresponding range space) is done from the Eigen-Value Decomposition (EVD) of S_w^X.

$$EVD(S_w^X) : U_r \Lambda_r U_r^T \tag{2}$$

where $U_r \in \mathbb{R}^{d \times r}$ are the eigenvectors associated to the nonzero eigenvalues Λ_r. The scattering added to the null space can be measured as the trace $tr(U_\alpha^T S_w^X U_\alpha)$. This quantity is up to $tr(S_w^X)$ when no directions are removed, $U_\alpha = U_r$, and decreases as more and more important directions disappear from U_r. Consequently, the scattering preserved after a projection, U_α, can be written as follows

$$\alpha = 1 - \frac{tr(U_\alpha^T S_w^X U_\alpha)}{tr(S_w^X)} \tag{3}$$

The projection basis fulfilling the above conditions for a given value of α can be obtained through U_r, such that r is reassigned. The α value is the main parameter of GDCV, which can be tuned by using cross-validation over the training set. The GDCV method can be the summarized as

1. Obtain U_α such that $S_w^X = U_r \Lambda_r U_r^T$. where Λ_α contains the smallest eigenvalues in Λ_r and $tr(\Lambda_\alpha) = \alpha \cdot tr(\Lambda_r)$
2. Project class means as $x_{gcv}^j = \overline{x}_j - U_\alpha U_\alpha^T \overline{x}_j$. These are the so-called generalized common vectors of each class.
3. Define $X^{com} = [x_{gcv}^1 \dots x_{gcv}^c]$ and let X_c^{com} be its centered version with regard to the mean, $\overline{x}_{gcv} = \frac{1}{c} \sum_{j=1}^{c} x_{gcv}^j$
4. Obtain the projection $W \in \Re^{d \times (c-1)}$ such that $tr(W^T X_c^{com} X_c^{com^T} W)$ is maximum.

Thus, by using the projection matrix W, any sample x_i can be projected in the discriminative subspace *gdcv* for an easier classification, according to

$$x_i^{gdcv} = W^T \cdot (xi - \overline{x}_{gcv}) \tag{4}$$

In an incremental learning scenario, once an initial dataset X has been used to obtain U_α, Λ_α and W, a new set of sample Y will be available in a later stage to improve the learned projection. This new set of data Y may be composed of a single sample or several ones that may belong to pre existing classes or to fully new categories. In the general case, the new dataset Y consists of n_j samples from each class, resulting in a total of $N = \sum_{j=1}^{c} n_j$ new samples to be considered in the learning process.

The IGDCV method allows obtaining $U'_\alpha, \Lambda'_\alpha$ and W' corresponding to the new complete dataset, $[X \;\; Y]$, without having to reapply the GDCV algorithm to $[X \;\; Y]$. Instead, they will be obtained incrementally by adding the effect of new data, Y, into the previous solution corresponding to X, such that

$$S_w^Z = S_w^X + Y_c Y_c^T + A A^T \tag{5}$$

where Y_c consists of column vectors $(y_j^i - \overline{y}_j)$ for all $j = 1 \dots c$ and $i = 1 \dots n_j$. $A = [a_1 \dots a_c]$ is a matrix whose columns are the c weighted average differences given by

$$a_j = \sqrt{\frac{m_j n_j}{m_j + n_j}}(\overline{x}_j - \overline{y}_j), \qquad j = 1 \dots c \tag{6}$$

The IGDCV algorithm is summarized as

Algorithm 1. *IGDCV Algorithm*

Parameter: $\alpha,\ 0 < \alpha \leq 1$

Input: $Y \in \mathbb{R}^{d \times N},\ \{n_j\}_{j=1}^c, \quad N = \sum_{j=1}^c n_j$

From previous iteration: $U_\alpha \in \mathbb{R}^{d \times r},\ \Lambda_\alpha \in \mathbb{R}^{r \times r},\ \overline{x}_j \in \mathbb{R}^d,\ \{m_j\}_{j=1}^c$

Output: $U'_\alpha \in \mathbb{R}^{d \times r'},\ \Lambda'_\alpha \in \mathbb{R}^{r' \times r'},\ \overline{x}'_j \in \mathbb{R}^d,\ \{m'_j\}_{j=1}^c$

Method:

1. Compute $\overline{y}_j,\ Y_c,\ A$
2. Compute $V = orth([Y_c\ A] - U_\alpha U_\alpha^T [Y_c\ A])$
3. Build $M_\alpha = \begin{bmatrix} \Lambda_\alpha & 0 \\ 0 & 0 \end{bmatrix} + [U_\alpha\ V]^T Y_c {Y_c}^T [U_\alpha\ V] + [U_\alpha\ V]^T A A^T [U_\alpha\ V]$
4. Compute R and Λ' by eigendecomposing M_α
5. Compute $\beta = (1 - \alpha)\frac{tr(\Lambda_r)}{tr(\Lambda')} + \alpha$
6. Split R and Λ' in R_β and Λ_β by β
7. Let $U'_\alpha = [U_\alpha\ \ V]R_\beta$ and $\Lambda'_\alpha = \Lambda_\beta$
8. Update: $m'_j = m_j + n_j,\ j = 1, \dots, c$
   $\overline{x}'_j = (m_j \overline{x}_j + n_j \overline{y}_j)/m'_j$
9. Project class means as $x_{gcv}^j = \overline{x}'_j - U'_\alpha {U'_\alpha}^T \overline{x}'_j$.

Figure 1: Incremental Generalized Discriminant Common Vector (IGDCV) algorithm.

If some of the data vectors in Y correspond to new classes which are not present in X, the expressions of the IGDCV algorithm are valid by extending the value of c and setting $m_j = 0$ in X for all new classes. Both if m_j or n_j are zero for any class j, the corresponding mean is undefined and the corresponding column in A, a_j, should be set to zero. If all data vectors in Y correspond to

new classes, then the whole matrix A is the zero matrix and can be removed from all expressions.

The overall cost of the IGDCV is dominated by the cost of step 7 in Fig. 1, $O(dr'^2)$ where r' is the expected rank of the range space preserved that heavily depends on the parameter α.

2.2. Classification

After applying IGDCV, samples can be projected into a discriminative subspace where meaningful conclusion can be extracted, if used as exploratory analysis, or an automatic classification can be achieved. The performed supervised learning ensures that the different classes to be recognized are as separate as possible, making the classification problem very simple, since the complexity of the problem has been moved to the previous stage. Thus, we have coupled our incremental subspace learning with a k-Nearest Neighbors (kNN) classifier in order to provide this functionality.

Two advantages are derived from the use of the KNN classifier. First, given its simplicity, the performance of IGDCV will be directly reflected in the experiments, which could otherwise be masked by a more complex classifier. Second, since no calibration is required in KNN, the online learning of the classifier will be automatic when the subspace is updated.

3. Case of study

In order to evaluate the potential and advantages of incremental learning, the problem of identifying vegetable oil types using spectroscopic analysis was chosen as case study. This is a relevant case of study [22, 26, 27, 28, 25] brought into attention due to European Regulation 1169/2011, which requires producers of foods that contain refined vegetable oil blends to label the oil types. In this context, deliberate or accidental errors in the label are common, leading to consumer misinformation [21], so automatic identification and verification of the provided information is required. From an analytical point of view, testing an

unknown vegetable oil to identify its origin and composition is a very difficult task [27, 35], but where spectroscopy -such as FTIR- and subspace-based methods have demonstrated their capabilities [22]. However, the performed single lab validation [22, 36, 37] of current approaches, which is common but undesirable in the field, indicates that the real performance in realistic conditions may be far from the reported accuracy.

4. Materials and methods

4.1. Samples

A data set of 630 vegetable oil samples was used in this study. Two different classification problems are considered with respect to the number of classes. Calibration models were developed for 6 classes and 12 classes of vegetable oils (see Table 1). For the 6-class problem, the classes to be predicted are labelled as PO: palm oil /palm stearin /palm olein, RS: sunflower /rapeseed oil and their mixtures, PKOC: palm kernel oil /coconut oil and binary mixtures of the above. For the 12-class problem, the classes are PO: palm oil /palm stearin /palm olein, RO: rapeseed, SO: sunflower, PKO: palm kernel, CCO: coconut, and all the binary combinations of the above oils. The 12-class model provides more resolution because it clearly distinguishes between the individual botanical origins, and it is therefore a more complex problem, while the 6-class modeld groups some origins together according to their similarities. This allows us to test our approach at to different levels of complexity, which are related to the expected level of resolution to be detected.

4.2. FT-IR spectral acquisition

The acquisition of most FT-IR spectra samples was performed using a Nicolet iS5 Thermo spectrometer (Thermo Fisher Scientific, Dublin, Ireland) equipped with a DTGS KBr detector and a KBr beam splitter. Spectra were acquired from 4000 to 550 cm-1 co-adding 32 interferograms at 4 cm-1 resolution with

Table 1: Different oil types for the 6 and 12-class problem.

	Class	Samples
1	PO	104
2	RS	114
3	PKOC	36
4	RS-PKOC	83
5	RS-PO	181
6	PO-PKOC	112

	Class	Samples
1	PO	104
2	RO	36
3	SO	23
4	PKO	26
5	CCO	10
6	RO-PO	98
7	SO-PO	83
8	RO-PKO	51
9	SO-PKO	32
10	RO-SO	55
11	PO-PKO	66
12	PO-CCO	46

a diamond attenuated total reflectance (iD5 ATR) accessory. Absorbance values were recorded at each spectrum point. The final sample spectrum was the average of three replicates with initial 7157 data points.

Through an interlaboratory experiment sixteen extra FT-IR instruments were used to acquire several extra oil spectra, as shown in Table 2. A total of nine samples including pure oils and oil admixtures were prepared in our lab and sent to each of the instruments participated to collect spectra representatives of most classes with all instruments. The acquisition parameters have been harmonized so that they are compatible with every FT-IR instrument. Linear interpolation was applied to spectra from different instruments in order to get the desirable number of variables.

4.3. Data pre-treatment

The resulting FT-IR spectral profiles underwent some typical preprocessing techniques in order to reduce or remove any random or systematic variation in the data [38]. Five steps are involved in this phase. Specifically, prior to the application of the multivariate models, Standard Normal Variate (SNV) [39], first order derivative [40], S-Golay filter [41] [polynomial order=2,frame size=9] and

Table 2: Instruments for the interlaboratory experiment. (Note: N/a - not available)

Id	Participant	FT-IR Instrument	Detector	Year	Samples
1	Our lab (Institute for global food security, QUB)	Thermo Fisher Scientific Nicolet iS5	DTGS	2012	486
2	Teagasc, Food Research Centre	Bio-Rad Excalibur FTS 3100	DTGS	2001	9
3	PerkinElmer Ltd	PerkinElmer Spectrum 2	DTGS	2012	9
4	PerkinElmer Ltd	PerkinElmer Frontier	DTGS	2013	9
5	Brennan and Co.	Bruker Alpha	DTGS	2013	9
6	Public Analyst Scientific Services	PerkinElmer Spectrum 100	LiTaO3	2007	9
7	LGC Limited (UK)	PerkinElmer Spectrum One	DTGS	2001	9
8	Premier Analytical Services (Premierfoods)	Bio-Rad Excalibur FTS300MX	DTGS	2002	9
9	Institute of Food Research (IFR)	Nicolet MagnaIR 860	DTGS	1998	9
10	Institute of Food Research (IFR)	Bio-Rad FTS6000	DTGS	1996	9
11	Institute of Food Research (IFR)	Thermo Fisher Scientific Nicolet iN10MX/iZ10	DTGS	2011	9
12	Shimadzu (Mason Technology)	Shimadzu IRA nity-1S	DLaTGS	n/a	9
13	Antech(IRE)	Thermo Fisher Scientific TruDefender FTX	DLaTGS	n/a	9
14	Agri-Food and Biosciences Institute (AFBI)	PerkinElmer Spectrum One	MIR TGS	n/a	9
15	Walloon Agricultural Research Centre (CRA-W)	Bruker Vertex 70	DLaTGS	2007	9
16	Walloon Agricultural Research Centre (CRA-W)	Bruker Vertex 70	DLaTGS	2012	9
17	Walloon Agricultural Research Centre (CRA-W)	Bruker Vertex 70	MCT	2012	9

Pareto scaling [42] were applied for removing the scatter, correcting the baseline, smoothing the data points and scaling the data for preventing the dominance of high absorbances respectively. At the end of this preprocessing procedure, the irrelevant spectra area was cut out by selecting only the wavelengths between 654.23 and 1875.43 cm-1 and between 2520.02 and 3120.74 cm-1, corresponding to relevant fatty acid involved in oil identification [22, 25]. In total, 3781 variables are resulted. All chemometric data preprocessing was performed by means of in-house Matlab routines (The MathWorks Inc., USA).

5. Results

Using our case of study, three scenarios where the potential of incremental learning is relevant will be tested. In the first scenario, an oil type identification model is trained with a few calibration samples. After this initial calibration, new samples for each of the oil types to identify become available and are added to the model for improving the initial performance. In the second scenario, a simple model is initially trained to distinguish between just two oil types, and then extended to identify new oil types, up to 12. In the third scenario, the oil type identification model created by a single lab and using a single spectroscopy analyser is extended and enhanced to be effective when used in other laboratories and instruments.

For comparison purposes, the batch version of IGDCV, batch GDCV, is used as a baseline. By using the exact batch equivalent version, we ensure the comparison is performed in the same conditions. The batch version requires to recreate the model every time that several, or even one single sample is available and added to the calibration set and therefore, access to the original samples is always obliged. The aim is then to ensure the same or similar classification performance to the batch method while reducing the computational time and removing the requirement of having access to the original calibration samples by the incremental approach. The α parameter was empirically optimised in the range $(0, 0.3]$ with steps of 0.01 for each scenario, so that the batch GDCV

provided the best accuracy result prior to any incremental step or addition of any new data. Then, the same value of α is used for both GDCV and IGDCV and keep constant over all the iterations. Thus, we aim to simulate a carefully fined-tuned initial pre-existing model to be further evolved.

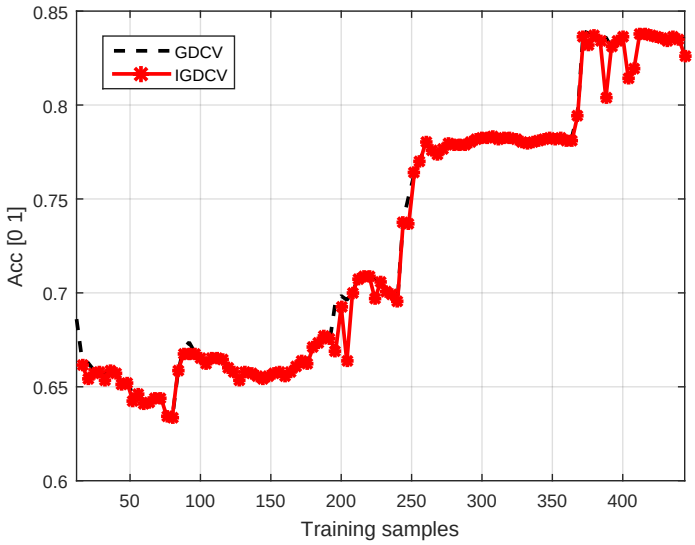
5.1. First scenario: New samples

In this experiment, we simulate a scenario where, for a given problem, an initial dataset is captured and the corresponding model is created. Then, new samples become available for calibration at different stages that can be used to improve the initial model and its performance. To do so, the 6 classes dataset is used. Cross validation is applied as evaluation protocol to avoid bias regarding the chosen samples. Ten iterations are performed, each with a random 70/30 split, i.e. the dataset is divided in 70% for calibration and 30% for validation in each iteration with no overlap between calibration and validation sets to avoid bias in the results. Results are then averaged over the splits to generate the final value. From the calibration samples, initially only 12 samples with representatives of all classes are used to generate a model. Then, in incremental step of 4 samples each, the model is evolved.

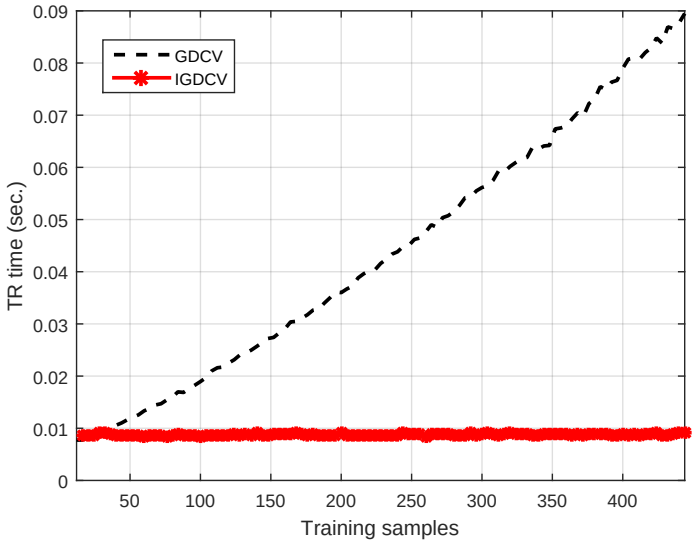
Fig. 2 shows the results of both incremental and batch methods, with the preserved scattering parameter set to $\alpha = 0.13$. As expected, models perform better as more calibration samples are available for learning from. Regarding the incremental learning, it can be observed how the accuracy of the incremental approach does not suffer, when compared with the batch algorithm, from not having access to the initial samples but only to the previous model. Moreover, when comparing the computational time required to generate the models (see Fig. 2b), one can notice the great difference in efficiency of using an incremental method regarding regenerating larger and larger models from scratch.

5.2. Second scenario: New classes

In this experiment, we simulate a scenario where a model has been created for a simpler identification problem that is then extended to cope with a more



(a) Accuracy (ACC)



(b) Training (TR) time

Figure 2: Batch GDCV and incremental IGDCV methods regarding new samples. Scattering parameter $\alpha = 0.13$.

complex problem. In the initial model, only 2 different oils are expected to be distinguished (Oil 1, 2) and this is incrementally evolved to identify more and more classes up to the total of the 12 species.

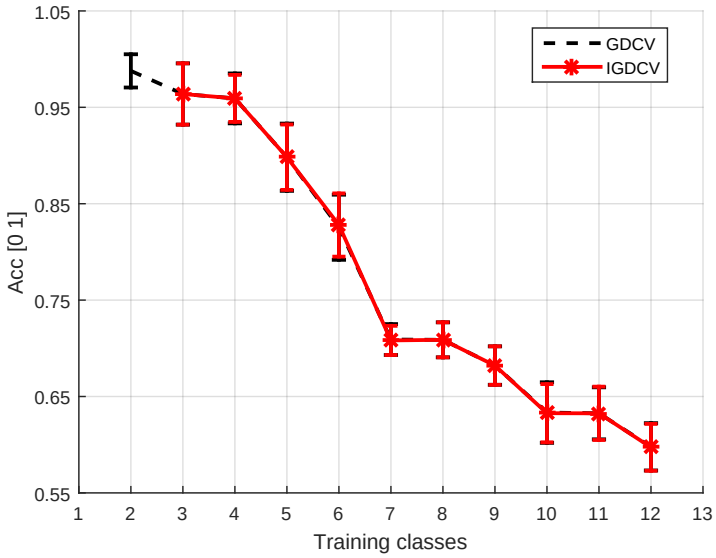
Similarly to the previous scenario, cross validation is also used as evaluation protocol, where 10 iterations are performed, each with a random 70/30 split, i.e. the 70% of the samples from each class is used for calibration and 30% are reserved for validation. Results are then averaged over the splits to generate the final value and the dispersion bar. In each iteration step, all calibration samples for a new class are added to the previous model.

Fig. 3 shown the results of both incremental and batch method. As expected, the more classes must be identified, the more complex the problem and, therefore, the accuracy decreases. Similarly to scenario 1, the potential of incremental learning is stated again by conserving the accuracy of the batch approach while reducing drastically the computational time and the access to the initial samples.

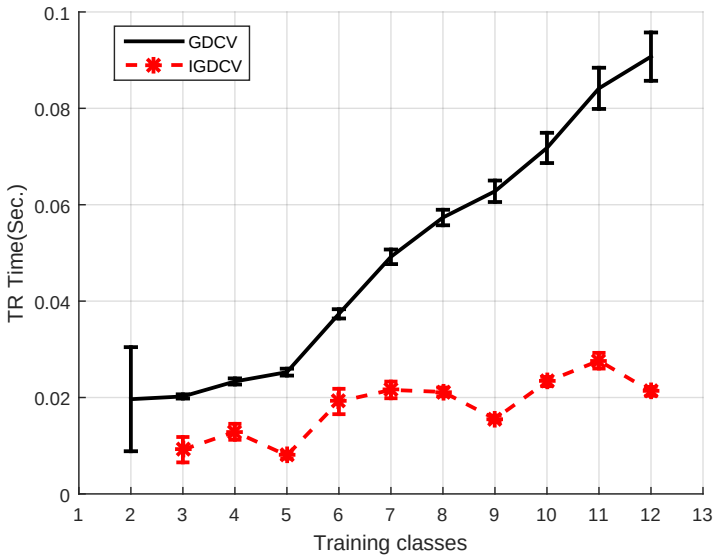
5.3. Third scenario: New instruments

In this scenario, we demonstrate the potential of the incremental learning to generalise previously existing models so that they can then be used by others laboratories using different instruments.

It has been shown that models created under controlled conditions, e.g. from a single calibration set when only one instrument was used, perform poorly when operating in real world conditions and report much lower performances than what it is expected from them. This can be corroborated by generating a model trained with 70% of the samples from instruments 1 (see Table 2). This model is first tested with the remaining 30% of the samples belonging to the very same instruments, and then tested with the samples from all other instruments. Similarly to previous scenarios, cross validation is used as evaluation protocol, where 10 random iterations are performed. Results for the 6 and 12 classes problems are depicted in Table 3.



(a) Accuracy (ACC)



(b) Training (TR) time

Figure 3: Batch GDCV and incremental IGDCV methods regarding new classes. Scattering parameter $\alpha = 0.07$.

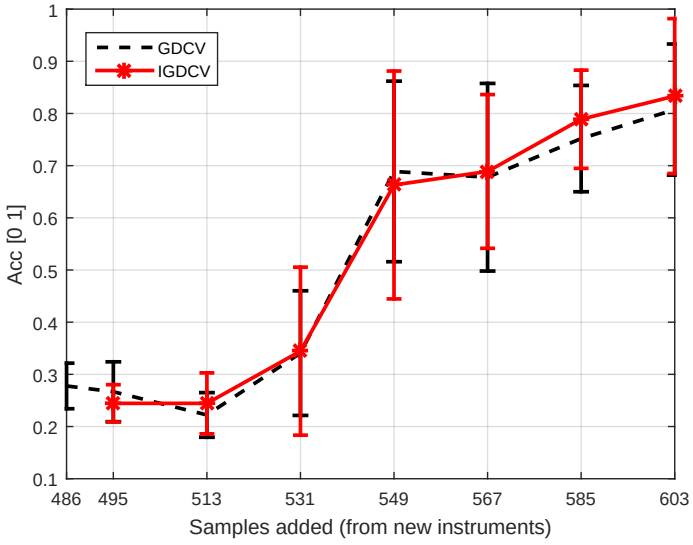
Table 3: Accuracy of GDCV model when using (2nd column) samples of the same instrument in the test set, and (3rd column) samples of different instruments in the test set to the instrument used in calibration.

Classes	Same Inst. in Test	New Inst. in Test
6	0.72 ± 0.04	0.28 ± 0.06
12	0.62 ± 0.03	0.14 ± 0.03

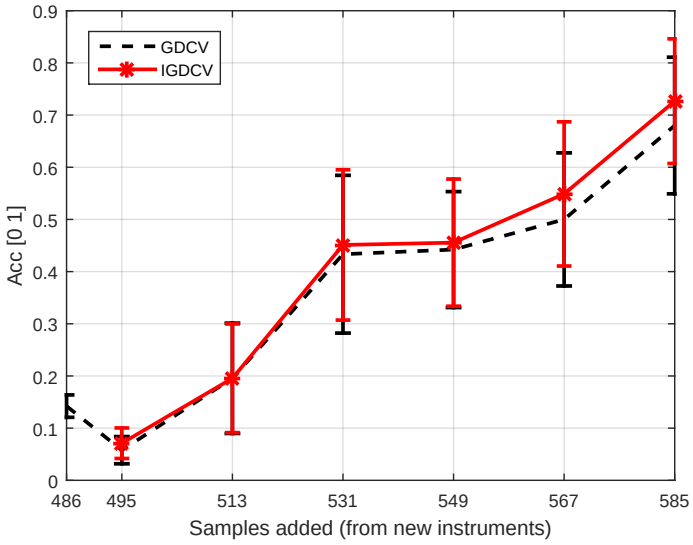
It can be noticed how an apparently good model, with reported accuracies 60-70%, underperforms dramatically under more complicated environments or conditions. It is therefore clear the necessity of improving an existing model in order to operate more broadly.

We simulate this situation in this third scenario, where we evaluate the potential of incremental learning to improve the generality of a previously created model initially created in a single laboratory. An initial model is trained with all samples from a single instrument. Then, the samples of a new instrument are added in a first step to evolve the model, followed by incremental steps of all samples belonging to new 2 instruments in each step. Two experiments are performed, one where the model has to identify 6 classes and the other one with 12 classes. Cross validation is used, repeating the experiment 10 times, where different instruments are randomly left out for the validation. In the 6 classes experiment, all samples from 3 different instruments are reserved for evaluating the system and up to 14 instruments are used in calibration. In the 12 classes experiment, all samples from 5 different instruments are kept for evaluating the system and up to 12 instruments are used in calibration.

Fig. 4 shown the results of both incremental and batch method. It can be noticed how using more instruments and collaborating between different labs allows to radically improved the performance of a given method. Both 6 and 12 class experiments behave similarly with slightly lower performance in the 12 classes due to the higher difficulty of the problem. We can see how



(a) 6 Classes



(b) 12 Classes

Figure 4: Accuracy (ACC) rate of the batch GDCV and the incremental IGDCV regarding new samples from new instruments. Scattering parameter $\alpha = 0.02$.

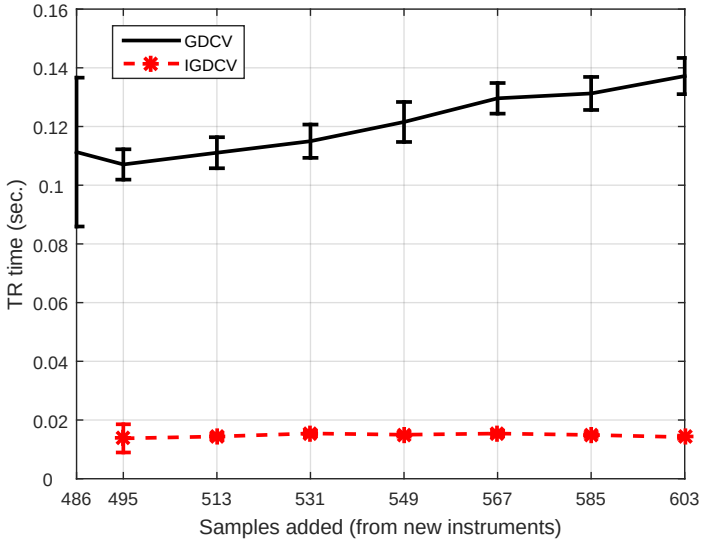
the incremental learning allows not only replicating the batch results but also it improves them regarding computational time, Fig. 5. It is also important to notice, how only a few samples from new instruments are needed (only 9 samples are available, see Table2) in our approach to improve significantly the final accuracy.

5.4. IGDCV as exploratory analysis tool

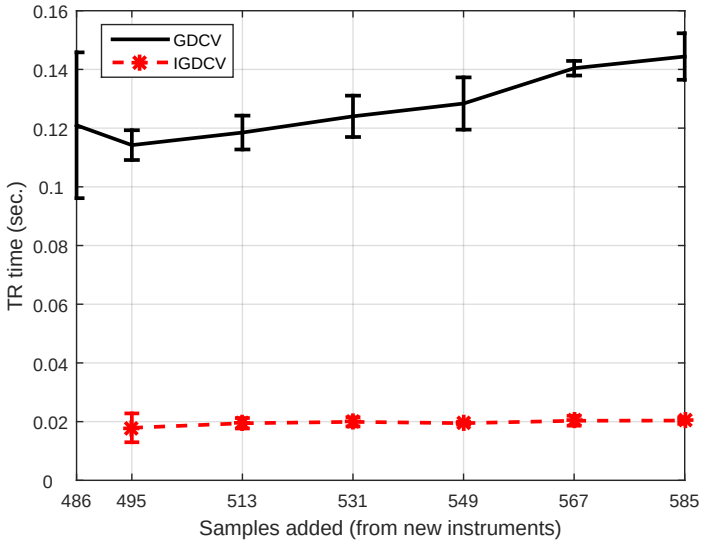
Apart from the benefits of using the IGDCV that were described earlier, IGDCV can also be used as an exploratory analysis tool, similarly to PCA [22]. In this regard, projecting the samples in the learned IGDCV can provide valuable information regarding the complexity of the problem, the likelihood of the model to accurately predict the correct answer and the quality of the samples. Furthermore, its incremental nature provides an extra functionality not available in PCA, GDCV or other batch methods, since once a model is created, a specific new sample(s) can be assessed in terms of its adequacy to be included in the analysis and/or in the calibration set of the following iteration of the model.

Figure 6 shows the evolution of the model for the 6 class problem in the first scenario, i.e. when samples are incrementally added. It can be observed how, while in the first space it is not very clear what are pure or admixture oil samples due to lack of data, this relationship is clearer the more online learning iterations occurs and more relevant samples are added.

Figure 7 shows the evolution of the model for the 12 class problem in the second scenario, i.e. when samples belonging to new classes are incrementally added. It can be observed how the complexity of the problem grows: while in the first space the 3 classes could be easily identified and separated, the space is more cluttered when the number of classes increases. This visualization can be used to decide which classes could not be resolved, and therefore should be excluded, due to their similar properties which are translated in their overlap in the space.

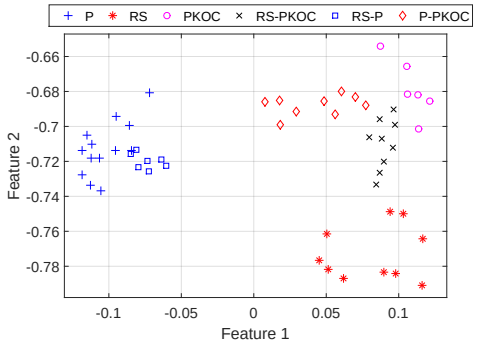


(a) 6 Classes

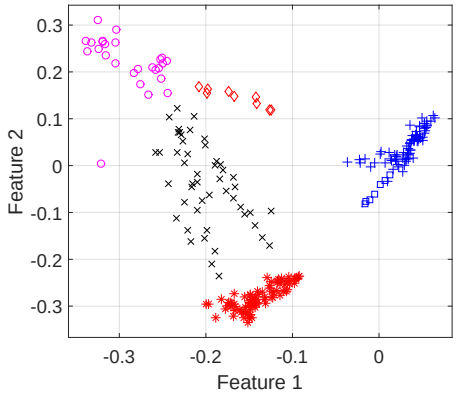


(b) 12 Classes

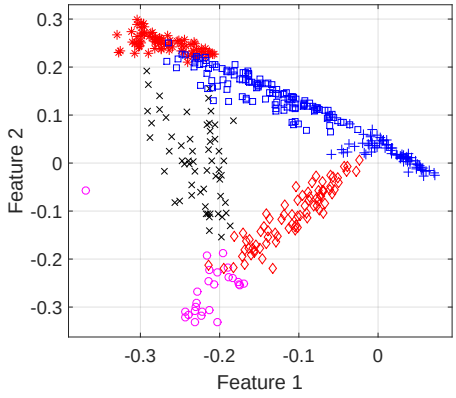
Figure 5: Training (TR) time of the batch GDCV and the incremental IGDCV regarding new samples from new instruments. Scattering parameter $\alpha = 0.02$.



(a) Initial model with 44 samples

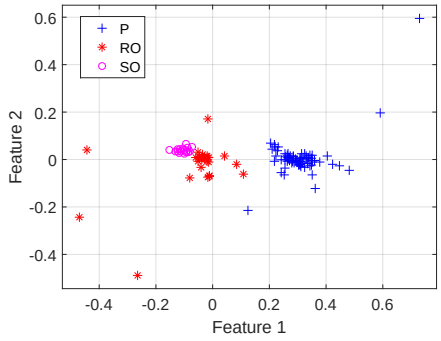


(b) After 200 samples have been incrementally learned

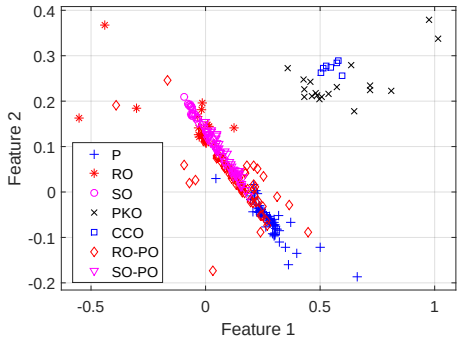


(c) After 400 samples have been incrementally learned

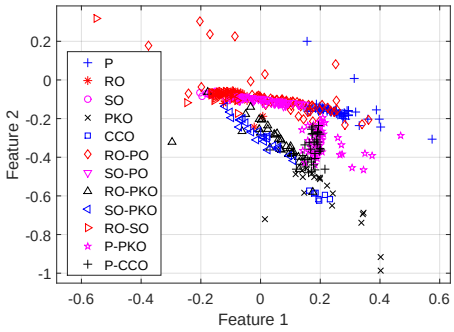
Figure 6: Samples projected into the two discriminant dimensions of the learned subspace, for the first scenario where samples are added incrementally (6-class problem)



(a) Initial model with 3 oil classes



(b) After 4 classes have been incrementally learned



(c) After 9 classes have been incrementally learned

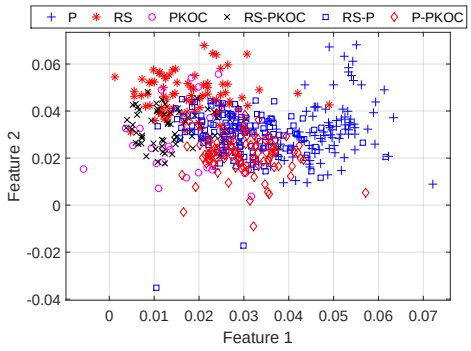
Figure 7: Samples projected into the two discriminant dimensions of the learned subspace, for the second scenario where classes are added incrementally (12-class problem)

Finally, Figure 8 shows the evolution of the model for the 6 class problem in the third scenario, i.e. when samples belonging to new instruments are incrementally added. It can be seen how the initial model is clearly insufficient to solve the problem and how adding more and more instruments seems a good idea to improve discrimination between classes. It could also be used to decide in which moment adding more instruments may not be convenient anymore, since the subspace will not evolve further, as seen between the second and third projections. Please notice how this visualization correlates with the quantitative results in Figure 4, where accuracy improvement reduces after 3 iterations, i.e. 6 instruments.

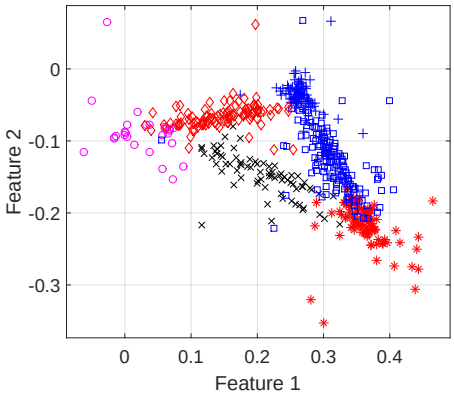
As can be seen, by using a incremental method for exploratory analysis, relevant information is provided to food scientist such as the detection of errors in the sample preparation or data generation, or the likelihood of an improved model by using a new batch of samples. Furthermore, this experiments were performed in a fraction of the time required by the batch method GDCV. Thus, Figure 6.b) and c) were generated in 28% and 13% of the batch time respectively, Figure 7.b) and c) in 40% and 22% of the batch time and Figure 8.b) and c) in 16% and 14% of the batch time.

6. Conclusion

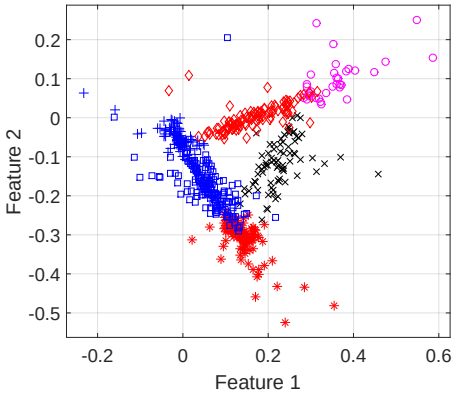
In this paper we apply the concept of incremental learning in food science and proposed the use of a subspace based learning method, both in its incremental and batch method as a new chemometric analysis tool. GDCV and IGDCV can be used as both classification and exploratory techniques, without some of the constraints that PCA or LDA exhibits, such as requiring large number of samples. The potential of incremental learning to improve and share models between analytical laboratories using different acquisition equipment is demonstrated through three different scenarios. By adding a very small number of samples to a preexisting model, our approach allows improving significantly the accuracy as well as to adapt the model to a new problem or scenario. The



(a) Initial model using samples from 2 instruments



(b) After samples from 6 instruments have been incrementally learned



(c) After samples from 10 instruments have been incrementally learned

Figure 8: Samples projected into the two discriminant dimensions of the learned subspace, for the third scenario where instruments are added incrementally (6-class problem)

IGCV incremental approach presented here has the advantage of maintaining or improving the accuracy while reducing the computational and spatial cost, and removing the hassle and privacy issues associated to share raw samples and wasting time and effort reproducing the models and tuning the analytical tools. As future work, we aim to extend our incremental subspace learning method to other cases of studies in chemometrics as well as integrating IGDCV as part of a new version of SIMCA. We also aim to study the use of decremental learning in chemometrics and add a decremental stage to our online learning framework.

Acknowledgements

The authors would like to thank all the participants (research centres, public services and private food testing labs) that helped to perform the interlaboratory experiment. This research was supported with funding from The Department Learning and Employment Northern Ireland (DELNI) (PhD studentship block grant) and the Department of Environment, Food and Rural Affairs (DEFRA) of the UK (Grant no. FAO 157).

References

[1] S. Lohumi, S. Lee, H. Lee, B.-K. Cho, A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration, Trends in Food Science & Technology 46 (2015) 85–98.

[2] A. J. Myles, T. A. Zimmerman, S. D. Brown, Transfer of multivariate classification models between laboratory and process near-infrared spectrometers for the discrimination of green arabica and robusta coffee beans, Applied spectroscopy 60 (10) (2006) 1198–1203.

[3] M. R. Kunz, J. H. Kalivas, E. Andries, Model updating for spectral calibration maintenance and transfer using 1-norm variants of tikhonov regularization, Analytical chemistry 82 (9) (2010) 3642–3649.

[4] M. Golic, K. B. Walsh, Robustness of calibration models based on near infrared spectroscopy for the in-line grading of stonefruit for total soluble solids content, Analytica Chimica Acta 555 (2) (2006) 286–291.

[5] M. R. Kunz, J. Ottaway, J. H. Kalivas, C. A. Georgiou, G. A. Mousdis, Updating a synchronous fluorescence spectroscopic virgin olive oil adulteration calibration to a new geographical region, Journal of agricultural and food chemistry 59 (4) (2011) 1051–1057.

[6] Evaluation of calibration transfer strategies between metal oxide gas sensor arrays, Procedia Engineering 120 (2015) 261 – 264.

[7] Calibration transfer and drift counteraction in chemical sensor arrays using direct standardization, Sensors and Actuators B: Chemical 236 (2016) 1044 – 1053.

[8] L. Fernandez, S. Guney, A. Gutierrez-Galvez, S. Marco, Calibration transfer in temperature modulated gas sensor arrays, Sensors and Actuators B: Chemical 231 (2016) 276 – 284.

[9] Y. Liu, N. Hu, H. Wang, P. Li, Soft chemical analyzer development using adaptive least-squares support vector regression with selective pruning and variable moving window size, Industrial and Engineering Chemistry Research 48 (12) (2009) 5731–5741.

[10] Y. Liu, H. Wang, J. Yu, P. Li, Selective recursive kernel learning for online identification of nonlinear systems with narx form, Journal of Process Control 20 (2) (2010) 181–194.

[11] D. Ross, J. Lim, R. Lin, M. Yang, Incremental learning for robust visual tracking, IJCV 77 (1–3) (2008) 125–141.

[12] Y. Peng, S. Pang, G. Chen, A. Sarrafzadeh, T. Ban, D. Inoue, Chunk incremental idr/qr lda learning, in: The 2013 International Joint Conference on Neural Networks (IJCNN), 2013, pp. 1–8.

[13] K. Diaz-Chito, F. Ferri, W. Díaz-Villanueva, Incremental generalized discriminative common vectors for image classification, IEEE Trans. Neural Networks and Learning Systems 26 (8) (2015) 1761–1775.

[14] N. Bhattacharyya, A. Metla, R. Bandyopadhyay, B. Tudu, A. Jana, Incremental pnn classifier for a versatile electronic nose, in: 3rd International Conference on Sensing Technology, 2008, pp. 242–247.

[15] B. Tudu, A. Metla, B. Das, N. Bhattacharyya, A. Jana, D. Ghosh, R. Bandyopadhyay, Towards versatile electronic nose pattern classifier for black tea quality evaluation: An incremental fuzzy approach, IEEE Transactions on Instrumentation and Measurement 58 (9) (2009) 3069–3078.

[16] B. Tudu, A. Jana, A. Metla, D. Ghosh, N. Bhattacharyya, R. Bandyopadhyay, Electronic nose for black tea quality evaluation by an incremental {RBF} network, Sensors and Actuators B: Chemical 138 (1) (2009) 90 – 95.

[17] C. Cernuda, E. Lughofer, L. Suppan, T. Röder, R. Schmuck, P. Hintenaus, W. Märzinger, J. Kasberger, Evolving chemometric models for predicting dynamic process parameters in viscose production, Analytica chimica acta 725 (2012) 22–38.

[18] C. Cernuda, E. Lughofer, G. Mayr, T. Röder, P. Hintenaus, W. Märzinger, J. Kasberger, Incremental and decremental active learning for optimized self-adaptive calibration in viscose production, Chemometrics and Intelligent Laboratory Systems 138 (2014) 14–29.

[19] C. Cernuda, E. Lughofer, P. Hintenaus, W. Märzinger, T. Reischer, M. Pawliczek, J. Kasberger, Hybrid adaptive calibration methods and ensemble strategy for prediction of cloud point in melamine resin production, Chemometrics and Intelligent Laboratory Systems 126 (2013) 60–75.

[20] E. Szymanska, J. Gerretzen, J. Engel, B. Geurts, L. Blanchet, L. Buydens,

Chemometrics and qualitative analysis have a vibrant relationship, TrAC Trends in Analytical Chemistry 69 (2015) 34–51.

[21] X. Zhang, X. Qi, M. Zou, F. Liu, Rapid authentication of olive oil by raman spectroscopy using principal component analysis, Analytical Letters 44 (2011) 2209–2220.

[22] M. Osorio, S. Haughey, C. Elliott, A. Koidis, Identification of vegetable oil botanical speciation in refined vegetable oil blends using an innovative combination of chromatographic and spectroscopic techniques, Food Chemistry 189 (SI) (2015) 67–73.

[23] S. Wold, M. Sjostrom, Simca: A method for analyzing chemical data in terms of similarity and analogy, Chemometrics Theory and Application (1977) 243–282.

[24] M. Osorio, S. Haughey, C. Elliott, A. Koidis, Evaluation of methodologies to determine vegetable oil species present in oil mixtures: Proposition of an approach to meet the eu legislation demands for correct vegetable oils labelling, Food Research International 60 (2013) 66–75.

[25] K. Georgouli, J. Martinez-Del-Rincon, A. Koidis, Continuous statistical modelling for rapid detection of adulteration of extra virgin olive oil using mid infrared and raman spectroscopic data, Food Chemistry 217 (2017) 735–742.

[26] R. Maggio, L. Cerretani, E. Chiavaro, T. Kaufman, A. Bendini, A novel chemometric strategy for the estimation of extra virgin olive oil adulteration with edible oils, Food Control 21 (2010) 890–895.

[27] B. Ozen, L. Mauer, Detection of hazelnut oil adulteration using ft-ir spectroscopy, Journal of Agricultural and Food Chemistry 50 (2002) 3898–3901.

[28] A. Koidis, M. Osorio, Identification of oil mixtures in extracted and refined vegetable oils, Lipid Technology 25 (2013) 247–250.

[29] S. Ozawa, S. Pang, N. Kasabov, A modified incremental principal component analysis for on-line learning of feature space and classifier, in: PRICAI: Trends in Artificial Intelligence, Vol. 3157, 2004, pp. 231–240.

[30] S. Pang, S. Ozawa, N. Kasabov, Incremental linear discriminant analysis for classification of data streams, IEEE Trans. Systems, Man, and Cybernetics (Part B) 35 (5) (2005) 905–914.

[31] K. Diaz-Chito, F. Ferri, W. Díaz-Villanueva, Image recognition through incremental discriminative common vectors, in: Advanced Concepts for Intelligent Vision Systems - ACIVS, 2010, pp. 304–311.

[32] P. Howland, J. Wang, H. Park, Solving the small sample size problem in face recognition using generalized discriminant analysis, Pattern Recognition 39 (2) (2006) 277–287.

[33] L.-F. Chen, H.-Y. Liao, M.-T. Ko, J.-C. Lin, G.-J. Yu, A new lda-based face recognition system which can solve the small sample size problem, Pattern Recognition 33 (10) (2000) 1713–1726.

[34] M.-B. Zhao, Z. Zhang, T. W. S. Chow, Z. Wu, On the theoretical and computational analysis between trace ratio lda and null-space lda, in: The 2012 International Joint Conference on Neural Networks (IJCNN), 2012, pp. 1–7.

[35] D. mejkalov, A. Piccolo, High-power gradient diffusion nmr spectroscopy for the rapid assessment of extra-virgin olive oil adulteration, Food Chemistry 118 (2010) 153–158.

[36] A. Christy, S. Kasemsumran, Y. Du, Y. Ozaki, The detection and quantification of adulteration in olive oil by near-infrared spectroscopy and chemometrics, Analytical Sciences 20 (2004) 935–940.

[37] E. Lopez-Diez, G. Bianchi, R. Goodacre, Rapid quantitative assessment of the adulteration of virgin olive oils with hazelnut oils using raman

spectroscopy and chemometrics, Journal of Agricultural and Food Chemistry 51 (2003) 6145–6150.

[38] O. Devos, G. Downey, L. Duponchel, Simultaneous data preprocessing and svm classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils, Food Chemistry 148 (2014) 124–130.

[39] R. J. Barnes, M. S. Dhanoa, S. J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc 43 (5) (1989) 772–777.

[40] B. G. Osborne, T. Fearn, P. H. Hindle, Practical NIR spectroscopy with applications in food and beverage analysis, Longman scientific and technical, 1993.

[41] A. Savitzky, M. J. Golay, Smoothing and differentiation of data by simplified least squares procedures, Analytical chemistry 36 (8) (1964) 1627–1639.

[42] H. C. H. R. A. Berg and, J. A. Westerhuis, A. K. Smilde, M. J. Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC genomics 7 (1).