Incremental model learning for spectroscopy-based food analysis


Published in:
Chemometrics and Intelligent Laboratory Systems

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2017 Elsevier B. V. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/, which permits distribution and reproduction for noncommercial purposes, provided the author and source are cited.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen’s institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person’s rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
Incremental model learning for spectroscopy-based food analysis

Katerine Diaz\textsuperscript{a}, Konstantia Georgouli\textsuperscript{b}, Anastasios Koidis\textsuperscript{b}, Jesus Martinez Del Rincon\textsuperscript{c,*}

\textsuperscript{a}Computer Vision Centre, Universidad Autonoma de Barcelona, Spain
\textsuperscript{b}Institute for Global Food Security, Queens University Belfast, UK
\textsuperscript{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 \cite{11, 12, 13}, its intrinsic advantages have been scarcely exploited in the field of food analysis and chemometrics \cite{14, 15, 16, 17, 18, 19}. Bhattacharyya et al. \cite{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) \cite{16}, the same researchers applied incremental fuzzy logic to the black tea identification. Cernuda et al. \cite{17, 18, 19} proposed a flexible fuzzy inference system for the monitor of the concentration of sulphuric acid (H$_2$SO$_4$), sodium sulfate (Na$_2$SO$_4$) 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 \cite{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 \cite{20, 21}, both for exploratory analysis \cite{22} and classification \cite{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 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, Linear Discriminant Analysis (LDA) and DCV. 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, 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.

Among the approaches that have been proposed to solve the SSS problem, the Generalized Discriminative Common Vectors (GDCV) has been proved to provide discriminative subspaces for classification regardless of the SSS assumption. GDCV is a variation of LDA 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 \cite{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 $x^i_j$ be a $d$-dimensional column vector which denotes the $i^{th}$ sample from the $j^{th}$ class. The within-class scatter matrix, $S^X_w$, is defined as,

$$S^X_w = \sum_{j=1}^{c} \sum_{i=1}^{m_j} (x^i_j - \bar{x}_j)(x^i_j - \bar{x}_j)^T = X_c X_c^T \quad (1)$$

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

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

$$EVD(S^X_w) : U_r \Lambda_r U_r^T \quad (2)$$

where $U_r \in \mathbb{R}^{d \times r}$ are the eigenvectors associated to the nonzero eigenvalues $\Lambda_r$. The scattering added to the null space can be measured as the trace $tr(U^T_{\alpha} S^X_w U_{\alpha})$. This quantity is up to $tr(S^X_w)$ 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, $U_{\alpha}$, can be written as follows

$$\alpha = 1 - \frac{tr(U^T_{\alpha} S^X_w U_{\alpha})}{tr(S^X_w)} \quad (3)$$
The projection basis fulfilling the above conditions for a given value of $\alpha$ can be obtained through $U_r$, such that $r$ is reassigned. The $\alpha$ 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 $U_{\alpha}$ such that $S_w^X = U_{\alpha} \Lambda_{\alpha} U_{\alpha}^T$, where $\Lambda_{\alpha}$ contains the smallest eigenvalues in $\Lambda_r$ and $tr(\Lambda_{\alpha}) = \alpha \cdot tr(\Lambda_r)$

2. Project class means as $x_{gcv}^j = \bar{x}_j - U_{\alpha} U_{\alpha}^T \bar{x}_j$. These are the so-called generalized common vectors of each class.

3. Define $X_{com} = [x_{gcv}^1 \ldots x_{gcv}^c]$ and let $X_{com}^c$ be its centered version with regard to the mean. $\bar{x}_{gcv} = \frac{1}{c} \sum_{j=1}^{c} x_{gcv}^j$

4. Obtain the projection $W \in \mathbb{R}^{d \times (c-1)}$ such that $tr(W^T X_{com}^c X_{com}^{cT} W)$ 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

$$x_i^{gdcv} = W^T \cdot (x_i - \bar{x}_{gcv})$$

(4)

In an incremental learning scenario, once an initial dataset $X$ has been used to obtain $U_{\alpha}, \Lambda_{\alpha}$ 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

$$S_w^{Z} = S_w^X + Y_c Y_c^T + AA^T$$

(5)
where \( Y_c \) consists of column vectors \((y_j^i - \bar{y}_j)\) for all \( j = 1 \ldots c \) and \( i = 1 \ldots n_j \).

\( A = [a_1 \ldots a_c] \) is a matrix whose columns are the \( c \) weighted average differences given by

\[
a_j = \sqrt{\frac{m_j n_j}{m_j + n_j}} (\bar{x}_j - \bar{y}_j), \quad j = 1 \ldots c
\]

The IGDCV algorithm is summarized as

**Algorithm 1. IGDCV Algorithm**

**Parameter:** \( \alpha, \ 0 < \alpha \leq 1 \)

**Input:** \( Y \in \mathbb{R}^{d \times N}, \ \{n_j\}^c_{j=1}, \ N = \sum^c_{j=1} n_j \)

**From previous iteration:** \( U_\alpha \in \mathbb{R}^{d \times r}, \ \Lambda_\alpha \in \mathbb{R}^{r \times r}, \ \bar{x}_j \in \mathbb{R}^d, \ \{m_j\}^c_{j=1} \)

**Output:** \( U'_\alpha \in \mathbb{R}^{d \times r'}, \ \Lambda'_\alpha \in \mathbb{R}^{r' \times r'}, \ \bar{x}'_j \in \mathbb{R}^d, \ \{m'_j\}^c_{j=1} \)

**Method:**

1. Compute \( \bar{y}_j, \ y_c, \ A \)
2. Compute \( V = \text{orth}( [Y_c A] - U_\alpha U_\alpha^T [Y_c A] ) \)
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 \( \Lambda' \) by eigendecomposing \( M_\alpha \)
5. Compute \( \beta = (1 - \alpha) \frac{\text{tr}(\Lambda_\alpha)}{\text{tr}(\Lambda')} + \alpha \)
6. Split \( R \) and \( \Lambda' \) in \( R_\beta \) and \( \Lambda_\beta \) by \( \beta \)
7. Let \( U'_\alpha = [U_\alpha V] R_\beta \) and \( \Lambda'_\alpha = \Lambda_\beta \)
8. Update: \( m'_j = m_j + n_j, \ j = 1, \ldots, c \)
   \[
   \bar{x}'_j = (m_j \bar{x}_j + n_j \bar{y}_j) / m'_j 
   \]
9. Project class means as \( x'_{gcv} = \bar{x}'_j - U'_\alpha U'^T_\alpha \bar{x}'_j \).

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 $\alpha$.

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 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, 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, 30, 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 model 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.

<table>
<thead>
<tr>
<th>Class Samples</th>
<th>Class Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 PO</td>
<td>1 PO</td>
</tr>
<tr>
<td>2 RS</td>
<td>2 RO</td>
</tr>
<tr>
<td>3 PKOC</td>
<td>3 SO</td>
</tr>
<tr>
<td>4 RS-PKOC</td>
<td>4 PKO</td>
</tr>
<tr>
<td>5 RS-PO</td>
<td>5 CCO</td>
</tr>
<tr>
<td>6 PO-PKOC</td>
<td>6 RO-PO</td>
</tr>
<tr>
<td>7 SO-PO</td>
<td>7 SO-PO</td>
</tr>
<tr>
<td>8 RO-PKO</td>
<td>8 RO-PKO</td>
</tr>
<tr>
<td>9 SO-PKO</td>
<td>9 RO-SO</td>
</tr>
<tr>
<td>10 RO-SO</td>
<td>10 RO-PKO</td>
</tr>
<tr>
<td>11 PO-PKO</td>
<td>11 PO-PKO</td>
</tr>
<tr>
<td>12 PO-CCO</td>
<td>12 PO-CCO</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>Id</th>
<th>Participant</th>
<th>FT-IR Instrument</th>
<th>Detector</th>
<th>Year</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Our lab (Institute for global food security, QUB)</td>
<td>Thermo Fisher Scientific Nicolet iS5</td>
<td>DTGS</td>
<td>2012</td>
<td>486</td>
</tr>
<tr>
<td>2</td>
<td>Teagasc, Food Research Centre</td>
<td>Bio-Rad Excalibur FTS 3100</td>
<td>DTGS</td>
<td>2001</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>PerkinElmer Ltd</td>
<td>PerkinElmer Spectrum 2</td>
<td>DTGS</td>
<td>2012</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>PerkinElmer Ltd</td>
<td>PerkinElmer Frontier</td>
<td>DTGS</td>
<td>2013</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>Brennan and Co.</td>
<td>Bruker Alpha</td>
<td>DTGS</td>
<td>2013</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>Public Analyst Scientific Services</td>
<td>PerkinElmer</td>
<td>LiTaO3</td>
<td>2007</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>LGC Limited (UK)</td>
<td>PerkinElmer</td>
<td>Spectrum 100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Premier Analytical Services (Premierfoods)</td>
<td>Bio-Rad Excalibur FTS300MX</td>
<td>DTGS</td>
<td>2002</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>Institute of Food Research (IFR)</td>
<td>Nicolet MagnaIR 860</td>
<td>DTGS</td>
<td>1998</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>Institute of Food Research (IFR)</td>
<td>Bio-Rad FTS6000</td>
<td>DTGS</td>
<td>1996</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>Institute of Food Research (IFR)</td>
<td>Thermo Fisher Scientific Nicolet iN10MX/iZ10</td>
<td>DTGS</td>
<td>2011</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>Shimadzu (Mason Technology)</td>
<td>Shimadzu IRA nity-1S</td>
<td>DLaTGS</td>
<td>n/a</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>Antech(IRE)</td>
<td>Thermo Fisher Scientific TruDefender FTX</td>
<td>DLaTGS</td>
<td>n/a</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>Agri-Food and Biosciences Institute (AFBI)</td>
<td>PerkinElmer</td>
<td>MIR</td>
<td>n/a</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>Walloon Agricultural Research Centre (CRA-W)</td>
<td>Bruker Vertex 70</td>
<td>DLaTGS</td>
<td>2007</td>
<td>9</td>
</tr>
<tr>
<td>16</td>
<td>Walloon Agricultural Research Centre (CRA-W)</td>
<td>Bruker Vertex 70</td>
<td>DLaTGS</td>
<td>2012</td>
<td>9</td>
</tr>
<tr>
<td>17</td>
<td>Walloon Agricultural Research Centre (CRA-W)</td>
<td>Bruker Vertex 70</td>
<td>MCT</td>
<td>2012</td>
<td>9</td>
</tr>
</tbody>
</table>
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 \(\alpha\) 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 $\alpha$ 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
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 ex-
pected, the more classes must be identified, the more complex the problem and,
therefore, the accuracy decreases. Similarly to scenario 1, the potential of incre-
mental 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.
Figure 3: Batch GDCV and incremental IGDCV methods regarding new classes. Scattering parameter $\alpha = 0.07$. 

(a) Accuracy (ACC) 

(b) Training (TR) time
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.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Same Inst. in Test</th>
<th>New Inst. in Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.72 ± 0.04</td>
<td>0.28 ± 0.06</td>
</tr>
<tr>
<td>12</td>
<td>0.62 ± 0.03</td>
<td>0.14 ± 0.03</td>
</tr>
</tbody>
</table>

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

(a) 6 Classes

(b) 12 Classes
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 Table 2) 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.
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) 6 Classes

(b) 12 Classes
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)
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 an 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
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


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


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


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