Brazilian Journal of Analytical Chemistry
2021, Volume 8, Issue 32, pp 45-61
doi: 10.30744/brjac.2179-3425.AR-38-2021

BrJAC

ARTICLE

Data Mining, Machine Learning, Deep Learning, Chemometrics

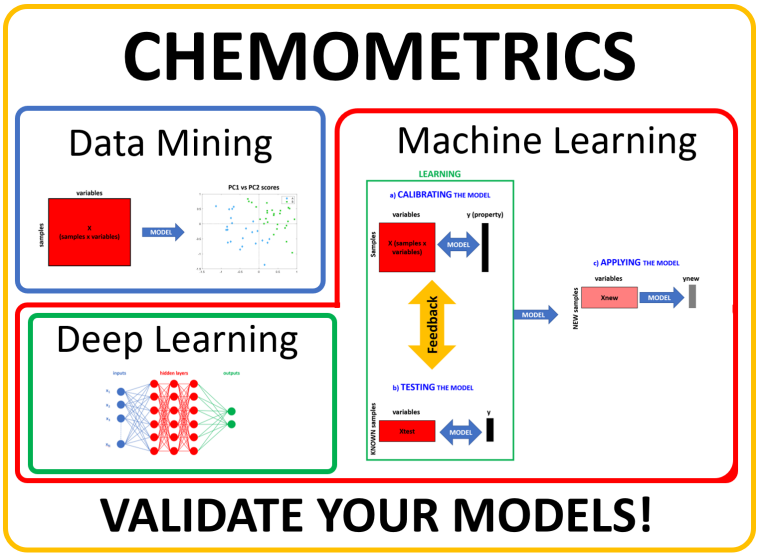
Definitions, Common Points and Trends

(Spoiler Alert: VALIDATE your models!)

José Manuel Amigo

IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain

Department of Analytical Chemistry, University of the Basque Country UPV/EHU, P.O. Box 644, 48080 Bilbao, Basque Country, Spain



Concepts like Machine Learning, Data Mining or Artificial Intelligence have become part of our daily life. This is mostly due to the incredible advances made in computation (hardware and software), the increasing capabilities of generating and storing all types of data and, especially, the benefits (societal and economical) that generate the analysis of such data. Simultaneously, Chemometrics has played an important role since the late 1970s, analyzing data within natural science (and especially in Analytical Chemistry). Even with the strong parallelisms between all of the abovementioned terms and being popular with most of us, it is still difficult to clearly define or differentiate the meaning of Machine Learning, Data Mining, Artificial Intelligence, Deep Learning and Chemometrics. This manuscript brings some light to the definitions of Machine Learning, Data Mining, Artificial Intelligence and Big Data Analysis, defines their application ranges and seeks an application space within the field of analytical chemistry (a.k.a. Chemometrics). The manuscript is full of personal, sometimes probably subjective, opinions and statements. Therefore, all opinions here are open for constructive discussion with the only purpose of Learning (like the Machines do nowadays).

MOTIVATION

I have spent the last 20 years analyzing data from many different analytical sources and many different scientific fields. From hyphen and hypernated chromatographic data [1] to all types of spectroscopies and different analytical metrics, they are specialized in hyperspectral image analysis [2]. Curiously, in recent years, I have been asked increasingly more often if what I do is Machine Learning, Data Mining, or even

Cite: Amigo, J. M. Data Mining, Machine Learning, Deep Learning, Chemometrics – *Definitions, Common Points and Trends* (Spoiler Alert: VALIDATE your models!) *Braz. J. Anal. Chem.*, 2021, *8* (32), pp 45–61. doi: http://dx.doi.org/10.30744/brjac.2179-3425.AR-38-2021

Deep Learning. My answer is always the same: I apply the mathematical procedure that I need to solve the problem I am dealing with if I need one at all. I completely acknowledge that this answer is quite open to interpretation. However, it could be summarized by one small sentence. **What I do is Chemometrics**.

During the last few years, there has been quite a lot of confusion in the literature with the terms Machine Learning, Data Mining, Deep Learning and Artificial Intelligence. I have observed a dangerous trend in the scientific literature towards their usage, highlighting the powerful benefits that these methodologies have in the data, using them as if the most important part of applying Machine Learning, for instance, was the algorithm we use. I have also observed that every analytical data issue could be solved by just constructing more complex algorithms, including more non-linear parameters and without placing more attention on the quality of the data being used.

Therefore, I thought that writing a manuscript such as this one could serve to 1) differentiate the terms mentioned beforehand, 2) highlight the most important facts of building a multivariate model and 3) give some tips and tricks to new students who want to start applying any mathematical model in their analytical data.

This manuscript is divided into 4 sections. The first will discuss the three most important aspects of data analysis: the data, the reference values and the model. I will comment on the structure of the data, the importance of relying on yor reference values, and the meaning of a mathematical model. The second and third parts define the terms Data Mining, Machine Learning, Deep Learning and Artificial Intelligence and puts them into the framework of Chemometrics. The manuscript will finish with a fourth section containing “take-home” messages. This is, arguably, the most important section of this manuscript, as it collects a series of advice that I have been using during my career and advised my students.

The manuscript is written in informal English to arrive at the audience in a more straightforward way. The manuscript is primarily addressed to Chemometricians; nevertheless, it would be a mistake not to open it up to anyone who analyzes data of any kind. It does not explain every term exhaustively or make a comprehensive revision of the literature on those matters. For that, some (but not many) keynote references are provided. Therefore, the reader is encouraged to check those references to find a more comprehensive explanation of specific models.

This manuscript contains several critical (therefore, subjective) opinions and suggestions that are completely open for discussion. Most of them give the impression of being too obvious; they are so obvious that sometimes we forget to pay attention to them. That is why I consider that highlighting them is important. I hope the readers find this manuscript interesting, bearing in mind that constructive criticism is more than welcome, always with the purpose of learning (as machines do nowadays).

THE DATA (X), THE REFERENCE (Y) AND THE MODEL. GI-GO

The reader would probably expect the manuscript to start with the definitions of Machine Learning, Data Mining, etc. Nevertheless, let me start the manuscript with the most important part in applying mathematical models to data: the data.

The data are just a mere collection of information containing relevant information and noise (i.e. not relevant information). Nevertheless, the most important aspect for success when applying any data analysis strategy is to have GOOD data **X** and, if needed, GOOD reference values **Y**. We know it as the GI-GO (Garbage In – Garbage Out) truism [3]:

IF THE DATA DO NOT CONTAIN ANY INFORMATION RELATED TO WHAT YOU WANT TO MEASURE AND/OR IF THE REFERENCES ARE NOT RELATED TO WHAT YOU WANT TO MEASURE, YOU WILL NOT OBTAIN GOOD RESULTS REGARDLESS OF THE MODEL USED

The GI-GO truism is, by far, the major cause of the frustration in the data analysis procedures. We blame the algorithms most of the time, forgetting that the algorithm will not find the information if it is not in the data. One of the biggest mistakes that we can commit is hypothesizing that the model will give

the solution we are looking for. On the contrary, the solution must be in the data and its correlation with the reference values. The model is, and will always be, the tool that helps us to find data patterns or the correlation between the data and the reference (if it exists).

The major issue here is that sometimes those patterns/correlations are difficult to find because they represent a small amount of variance/co-variance in the data. This is there where different algorithmic approaches can be tested, but always after being completely sure that we fully understand the nature (structure) and origin of our data.

KNOW YOUR DATA AND, IF ANY, YOUR REFERENCE

The data

One of the premises to start with is the fact that the analytical information that we measure is (or can be seen as) multivariate. That is, for one sample, many variables/observations are normally collected. We usually want to compare samples assuming that the differences or similarities between them will be found in the variables (or groups of variables) that we measure. In other words, we want to obtain useful information and get rid of the noise.

DATA = INFORMATION + NOISE

For this to be accomplished, we need to know three important data features: the nature (structure), amount and quality.

The nature (structure) of the data

Knowing the nature of our data will help us to choose 1) the appropriate pre-processing methods and 2) the subsequent models. A normal arrangement of the data is in the shape of a matrix (Figure 1a-d). This matrix **X** is normally composed of samples in rows and variables in columns (even though some disciplines prefer the transposed version with samples in columns and variables in rows, we will keep the classical nomenclature that is commonly used in Chemometrics) [3]. Nevertheless, there are many ways in which **X** can be constructed; even with the same apparent dimensions, different scientific instruments might provide data with a completely different structure. For instance, the cases presented in Figure 1a-d, where four matrices **X** are presented with the same dimensions but a completely different structure. The data in Figure 1a represents the typical data coming from any spectroscopic device, where the spectrum at *N* variables has been collected for each of the *m*th samples. Instead, the data in Figure 1b represents a single sample where the intensity level of a property has been measured in the pixels located in positions *M* and *N* (i.e. a monochannel picture). Even the data coming from the same instrument can be arranged/handled in different ways. For instance, when the chromatogram of a sample is measured, we can study the chromatogram variation between samples (Figure 1c) or construct a table where the integrated area of the peaks composes the variables.

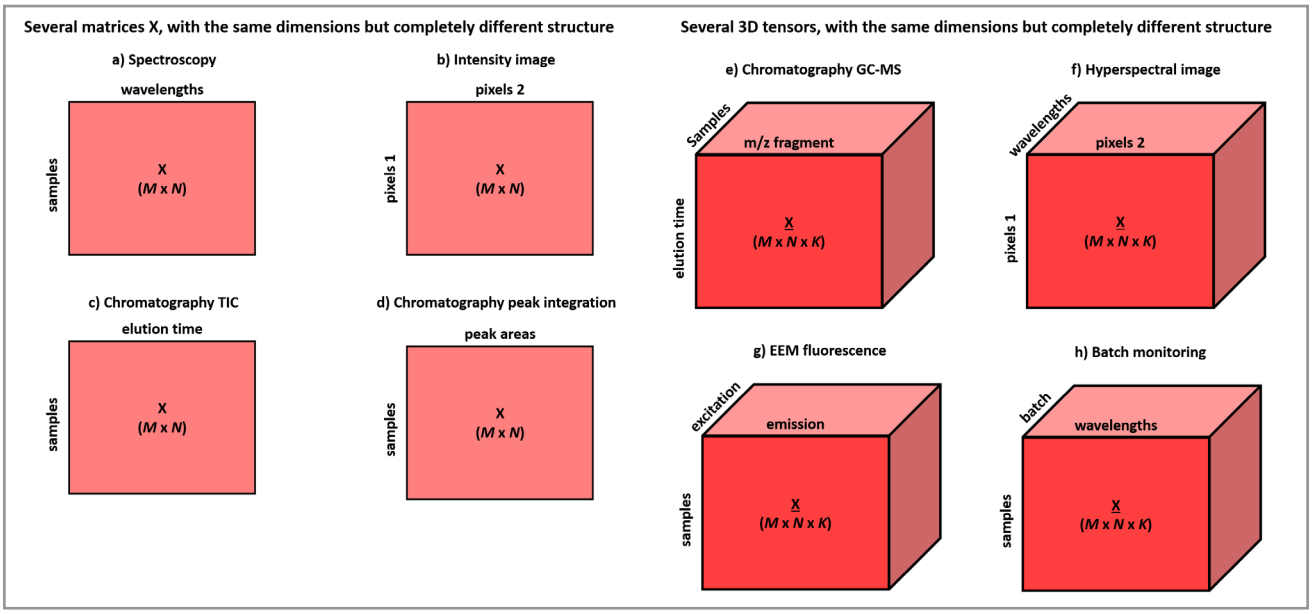


Figure 1. Different structure of datasets having the same dimensions. TIC, total Ion Chromatogram. GC-MS, Gas Chromatography-Mass Spectroscopy.

This difference in the nature of the signal has a strong impact, especially in the pre-processing and variable normalization steps that need to be applied before data analysis, where the relationship between the variables plays a fundamental role. Apart from being continuous, discrete or categorical in the columns, the variables can also be correlated in the rows. A toy example constructs a matrix considering of two continuous variables like pH and temperature. Those two variables, in principle, are independent. Therefore, there is no relevance in constructing a [pH Temperature] matrix or [Temperature pH] matrix. Now, we should think about spectral variables. When a spectrum is measured, for instance, in Near Infrared, there is a correlation between wavelength1, wavelength2 and wavelength3. Therefore, constructing a matrix such as [wavelength1 wavelength3 wavelength2] will have an impact on the pre-processing applied to that matrix because the correlation/continuity between variables has been broken.

Of the four examples, the image (Figure 1b) does not fulfil the requirement of (sample x variables) structure. Indeed, that matrix only represents information relative to one sample. Therefore, and depending on the aim, several steps should be taken beforehand (vectorization, features extraction, etc.). Considering the spectral and chromatographic profile matrices (Figure 1a and c), one can argue that the data structure is the same (or, at least, very similar). Nevertheless, the data source is completely different, and, therefore, the issues coming from the measurement must be addressed properly. The issues normally found in chromatograms (baseline drifts, peak misalignment, normalization by standard peaks, etc.) are not the same as the issues normally found in spectroscopy. Consequently, both scenarios need different pre-processing methods.

The issue can be further complicated if we consider that we might have data that contain more than 2 directions. Figure 1e-h show the situation dealing with 3D datacubes (a.k.a. tensors). There is a third direction in the data that expresses either another analytical variable or another measurement condition. For instance, the data coming from chromatographic measurements where the detector is a multichannel detector normally leads to datacubes where there is an extra spectral dimension. In a hyperspectral measurement, one sample is normally visualized as a datacube. The measurements in one sample give a matrix composed of the emission spectra at different excitation wavelengths in multiway fluorescence [4,5].

Another typical example is the data collected from different batches in factories (Figure 1h). All of these are datacubes. Nevertheless, their structure is completely different, and here we might have to take the initial decision to treat the data as such (by using multiway models [4,5]) or unfold the data in such a way that we set it as a matrix that fulfils the previous statement in the relationship between the samples and the variables.

In the case of chromatographic datacubes, there are normally two alternatives: either use multiway methods with a cube **X** (samples x elution time x spectral direction) (and thus assuming specific relationships between the samples and now the two variable directions) or unfolding the cube in different ways: (samples*elution time x spectral direction) or (samples x elution time*spectral direction). The same issue is found in, for example, Excitation – Emission fluorescence. In the case of hyperspectral images, a previous step of unfolding is normally preferred. Therefore, the datacube **X** (X x Y x wavelengths) is normally treated as a matrix where each pixel is considered a sample, obtaining **X** (X*Y x wavelengths).

The reader will find in the literature a plethora of options for arranging data regarding its structure, the aim of the analysis and the benefits/drawbacks of arranging the data in different ways, which makes this a bit cumbersome. Unfortunately, there is no clear answer in this regard. Nevertheless, I could recommend that the reader perfectly knows the data, the nature of the relationship between the samples and the variables, and the plausibility of applying different strategies that might arrive at the same result.

The amount and quality of data

One recurrent question during my lessons is how many samples are needed to build a good model (either for exploring data or building a regression or a classification model). The answer is easy:

YOUR MODEL NEEDS AS MANY SAMPLES AS NECESSARY TO CERTIFY THE RELIABILITY AND REPRESENTATIVENESS OF YOUR MODEL AND WHATEVER YOU WANT TO EXPRESS WITH YOUR MODEL

There are scenarios where what we want to measure is clearly expressed in the data that are being measured. For instance, measuring proteins in barley with Near Infrared (NIR) has been a classical tool, almost since NIR instruments were invented. The signal of the protein band is normally well expressed in the NIR region, the concentration of protein is high enough to have a good signal-to-noise (S/N) ratio, and the classical interferences in NIR like water (moisture) are not present in such high amounts, so they do not affect the NIR signal to a great extent. Therefore, it is the amount of data that plays the game and affects quality.

The quality of the data can be assessed by controlling different parameters of the measurement scenario in the measured signal:

- The instrumental noise: Data will contain noise. It is a fact. Nevertheless, even though different methodologies can minimize that noise, it must not be higher than the signal being measured.
- The composition of the sample and the plausible interferences in the signal that we are looking for: Measuring multivariate data means that not all of the variables are useful for answering the analytical question. Moreover, the signals that will give us the answer (spectral bands, chromatographic peaks, etc.) may be strongly influenced by another chemical (or physical) compound that could be in the sample matrix.
- The influence of the environmental conditions: Measuring in laboratory conditions is sometimes completely different from measuring in more uncontrolled conditions.
- The correlation with the property that we want to measure: In regression and classification, the close connection of the data with the reference value is essential.

Many of the previously mentioned issues can be partially minimized by developing a proper protocol in the calibration of the instrument and the inclusion of reference compounds to normalize the data. Also, having a good Design of the Experiment [6] and a perfectly optimized analytical method will help us to understand the plausible confounding factors that we might have when designing our experiment.

The reference values

When regression or classification is needed, the role of the reference values is essential to identifying reliable correlations between them and the data **X** through the model, since having good data **X** is not sufficient to obtain a good model. The reference values **Y** can be obtained in many different ways, depending on the aim of the experiment. However, they can be summarized into two major blocks: regression and classification.

Regression

When a regression model is made, the reference value normally comes from a standardized analytical procedure that the proper regulatory agencies have approved; sometimes, it comes by being the most accepted procedure by the analytical community. Regardless of the precedence, we must be aware that the reference values contain an analytical error and a calibration range, together with a limit of detection and quantitation that is utterly linked to the data **X**.

Sometimes, the reference value error can be neglected if the error of the data **X** is larger. However, assuming this statement without verifying it might lead to a wrong interpretation of the result. The normal procedure to verify the error of the reference values is to make repeated measurements of the same sample and ascertain that the variance (standard deviation of the mean) between the replicates is within certain confidence levels. These concepts come from classical analytical chemistry procedures. Nevertheless, developing such protocols is sometimes time-consuming, or there is not enough budget to perform as many as we would like.

Classification

Classification is directly linked to the assignation of the belonging of one sample to one class, several classes, or none, depending on the classification strategy [7]. Therefore, the reference values are normally given by a categorical indexation of **Y**, where an inter-correlation between the different columns in **Y** is expected or, at least, assumed. In many scenarios, the **Y** class is imposed by applying certain thresholds to continuous variables (Temperature < 20 is cold, 21 < Temperature < 30 is mild, Temperature > 30 is hot). Alternatively, there are also occasions where the classes are assigned by using sensory panels. Then, the final class is assessed by an average of the grades given by a certain number of judges. In these cases, the assignation of classes comes with a similar analytical error defined beforehand in the regression scenario, being crucial to understanding the performance of the future model.

Many different strategies can be implemented when dealing with classification problems. Let me start with the simplest one; that is, the two-class problem. Figure 2a shows a case where the samples want to be separated by color. There are two plausible arrangements for the matrix **Y**. The first arrangement is to construct **Y** in the shape of a column vector containing arbitrary values assigned to each class. In this case, the number assigned to each class is completely irrelevant since all of the statistic parameters, and figures of merit will be constructed based on the relative difference between the predictions. The second strategy consists of constructing a matrix with two columns, where 1 and 0 are normally used to denote belonging and not belonging, respectively. Both strategies are equally good in the two-class case, and the results obtained from both strategies will be the same.

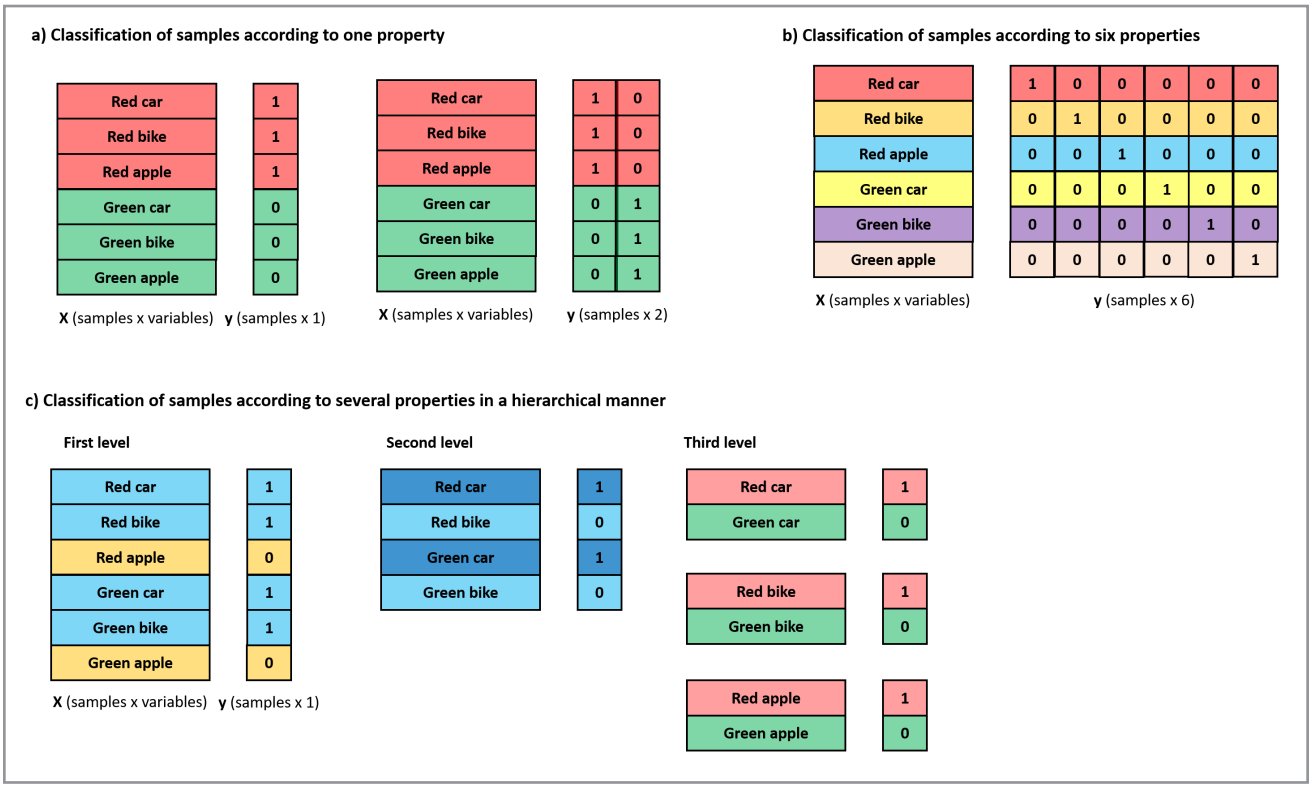


Figure 2. Different classification strategies that can be adopted to classify the same samples in different classes.

Nevertheless, it is important to make a statement that will be important for any classification strategy (not only for the two-class case). As the reader has probably noticed, the toy example in Figure 2 does not seem to be as realistic as we could expect. The samples are, indeed, quite different. There are red and green cars, red and green bikes and red and green apples. This is, indeed, an example that makes it clear that samples can be classified as we need them to be classified. It will depend on the strong correlation between what we want to classify and what we are actually measuring. Obviously, if we want to separate the samples by a different color, we must be sure that the data **X** that we obtain/measure is directly or indirectly related to the color of the samples. We normally hypothesize that our classification problem will be solved by performing certain measurements. Also, as indicated, this is just a hypothesis. Despite the many efforts made with our models, we cannot classify the samples by color measuring the number of wheels.

IF DATA X DO NOT REFLECT OR CONTAIN INFORMATION RELATED TO THE Y CLASS, NO MODEL WILL BE ABLE TO CLASSIFY THEM CORRECTLY

The previous statement is, of course, valid for the regression problem. Despite seeming obvious, it is one of the major sources of frustration; this is quite understandable. We construct our classification/regression hypothesis based on the problem. Nevertheless, the final decisions to solve the problems are normally drafted by the instruments that we have available. Also, we have to adapt, hoping to find those little pieces of information in our data that correlate with the class.

When the classification problem involves more than two classes, we might follow different strategies. The first is to build a classification model with all of the plausible classes. In our bike-car-apple example, this is represented in Figure 2, where we have 6 different and somehow independent classes. In this case, the matrix **Y** is normally constructed following a logical pattern of a correlated/not correlated (a.k.a. yes/no) strategy where there are as many columns as classes. Here, it is extremely important to follow the methodology indicated in the figure since we have to ascertain the equality of the influence for each class

in the model. There might be the temptation to assign a number to each class (if we have 6 classes, **Y** will be a vector with a number from 1 to 6). This is a big mistake. All classification models are normally built following a strategy in which a regression model is first developed, and then some thresholds are applied. Therefore, we will be committing the incorrect assumption that the distance between class 1 and class 4 is larger than the distance between class 1 and class 2, just because 4 is larger than 2.

Apart from that, and after constructing the correct **Y** matrix, normally filled with zeros and ones, we must be completely sure, as before, that the data **X** will contain enough information in the variables that will make the generation of a classification model for 6 classes at the same time possible.

It might be that the variance given in the data **X** by the different classes is not comparable or evenly distributed. Therefore, there is another strategy that facilitates the task of classifying the samples by using what is called a hierarchical model strategy. In a hierarchical strategy (Figure 2c), the problem is split into three minor classification models to solve the classification issue hierarchically. The samples will be classified first by the most important class (the class giving more variance in the data **X**) and then by other classes. However, this strategy is still based on the fact that no matter how many levels you think of, if the variables in data X are not collected, the information will not work either.

The model

Now yes! After understanding the importance of the data **X** and the reference **Y**, it is the perfect time to talk about the model a little bit. What is a model, and what is the difference between that and the word "modelling"?

A model, in mathematics, is the set of parameters and operations that fits the value of a dependent variable (y) to an independent variable (x). Easy [3]:

$$\mathbf{y = f(x)}$$

That relationship can be different (linear or non-linear) and with varying complexity. Nevertheless, it can be directly extrapolated to the concept of multivariate models. The main difference is that in a multivariate model, we have multivariate data. Therefore, given a set of samples measured by independent variables **X** (M x N) and a property **y** (M x 1), the model that establishes the correlation between **X** and **y** is:

$$\mathbf{y} = b_1x_1 + b_2x_2 + \ldots + b_nx_n = \mathbf{bX + e}$$

where **b** is what we normally call the regression vector and
e is the vector (M x 1) containing the residuals.

Of course, the previous equation is a mere visualization of a multivariate model applied to a regression (or even classification). The multivariate model must be considered the relationship between **X** and **y** and all of the pre-processing strategies, variables selected, and optimization parameters needed to predict/forecast the behavior of new **y** samples. This is the main aim of a multivariate model, to predict the value of y in new samples that have not been included in the construction of the model. This is partially true because there are situations where we do not want to correlate our data **X** with any **y**. We just want to study the patterns (points in common and trends) in the data **X**.

Pattern recognition models are the ones that, given one data matrix **X,** aim to study the correlations and differences (variance) between the samples reflected by the variables or groups of measured variables. The workhorse method is, undoubtedly, the principal component analysis (PCA) model [8,9]. PCA will indeed decompose our data matrix **X** into a set of so-called scores and a set of so-called loadings as indicated below:

$$\mathbf{X} = \mathbf{TP}^{T} + \mathbf{E}$$ (by the way, this is one of the essential equations in data analysis)

where **T** (M x F) is the score, **P** (N x F) is the loading with a superscript denoting the transposed value of the matrix, and **E** (M x N) is the residuals. F denotes the number of principal components or, in other words, the number of different independent sources of variance (variability) of the data.

In PCA, the model is, strictly speaking, the loading and pre-processing and normalization strategies used. Nevertheless, we tend to include the scores in the team, which it is acceptable to do.

Many algorithms obtain relationships between **X** and **y** and determine the trends in a single matrix **X** [10]. Nevertheless, in the end, what counts is that the relationships obtained give us useful information. In this regard, it is of utmost importance to highlight one fact:

THE MODELS WORK IN LOCAL SCENARIOS IF THE DATA ONLY CONTAIN LOCAL INFORMATION

The models that we construct are based on the information that we provide (the data **X** and the **Y** reference values, again!). It all depends on how good and comprehensive this information is. Therefore, what we can normally ensure is that our models work under certain conditions and certain limitations. Let me use a simple example. A classification model can be constructed in order to differentiate the origin of red wines. First of all, we might want to avoid expressions like “my model can classify different red wines”. This could be true, but then we have to ensure that our **X** matrix includes all red wines in the world. The model arrives as far as the richness of the data with which it is built. Also, we must be careful with expressions like “This model is always better than this one”. Well, not in a multivariate perspective and not working in local situations. The correct sentence should be: This model is better than this other in measuring a specific property in a specific sample using a specific instrumental device and under specific instrumental and environmental conditions. Generalization has the risk of giving the wrong impression that certain models will always overcome other models, and that is not true. We work on local situations unless, as I said, we ensure that our data **X** and reference values **Y** are representative enough of the problem.

DEFINITIONS

Having perfectly understood (I hope) that three actors are playing an important role in multivariate data analysis, now is the time to define the terms that bring us to this manuscript. It is also the moment to think back on everything that has been said before and put it into the framework of the definition of each term.

Data Mining

Data Mining can be defined as a set of methods used to extract usable information from large raw data sets. It should be noticed that this definition implies that the usable information is already in the data. Nevertheless, the complexity of the data and the multivariate (multiway) nature of the data means that we are unable to find useful information without powerful mathematical tools. Basically, data mining aims to separate the grain from the hay or find patterns that already exist in the data; however, they are hidden due to a large number of samples and variables, the noise of the data, or the difficulty in linking more than two variables at the same time in a univariate fashion (one variable at a time).

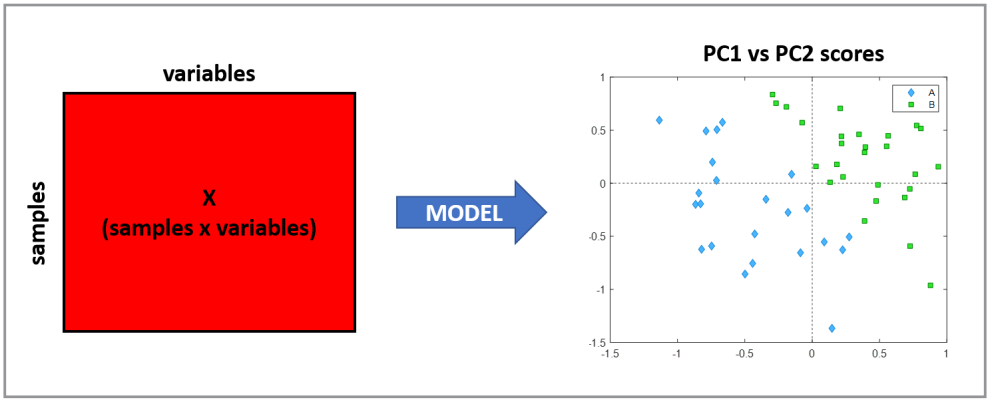


Figure 3. Data Mining applied to a matrix X. A PCA model has been applied, and the score scatter plot between PC1 and PC2 is shown.

One important fact of the Data Mining definition is that it does **NOT** include any procedure of Learning. Therefore, the methods or algorithms used for data mining are of an **unsupervised nature** (Figure 3). This narrows down the choices since any algorithm that involves Learning (i.e. supervision) should be kept aside.

DATA MINING = UNSUPERVISED

Under the umbrella of Data Mining, we can find methods like dendrograms, clustering or PCA (among others). One thing that must be clear is that Data Mining methods are not, *per se*, classification methods. In the literature, we can find works using, for instance, PCA for classification purposes. This might be a common language error since we tend to mention that PCA has classified the samples into some groups when we explain the results of a PCA model (it happens to me quite often). Nevertheless, the word "classify" and the word "group" have two separated meanings. Classification implies supervision (as we will see further). One general action, however, could be applying PCA or dendrograms and then setting some thresholds to group the samples into different classes. In any case, setting the threshold is an operation made *a posteriori* in an attempt to group samples.

Machine Learning

Machine Learning is normally defined as a series of methods that learn from the data to make or construct a model that can make informed decisions based on what is learned. This definition directly implies that the model/algorithm needs to learn. The learning procedure makes the algorithm reliable enough to predict any property in new data that has not been used for Learning. Here is where the supervised methodologies must be used.

MACHINE LEARNING = SUPERVISED

Indeed, the step of Learning could be substituted by the word Training. Therefore, any Machine Learning algorithm (but really, any) is composed of two steps: calibrating and testing the model, that is, the learning (training) step and applying the model to new samples (Figure 4).

Calibrating the model

Given a calibration data matrix **Xcal** and a reference value (in Regression or Classification) **Ycal**, the calibration step involves all of the necessary operations of data pre-processing, data normalization, variable selection, and the removal of outliers, among others, with the only purpose of finding the best conditions for obtaining the highest correlation (co-variance) between **Xcal** and **Ycal** (Figure 4a). That is as simple and as complicated as it appears. Once the optimal conditions have been found, the model is

set, and it could be used to predict whatever property is needed. Nevertheless, this step of calibration is completely useless without the next step, testing the model.

Testing (validating) the model

One of the main drawbacks of all multivariate models is that, since they are based on projection operations and are not parameterized, they need to be tested. Testing a model is used to verify that the model created in the calibration step can predict the outcome.

A GOOD MODEL IS NOT THE ONE THAT BETTER CALIBRATES, BUT THE ONE THAT BETTER PREDICTS. WITHOUT VALIDATION, THE MODEL IS COMPLETELY USELESS. VALIDATE YOUR MODELS!

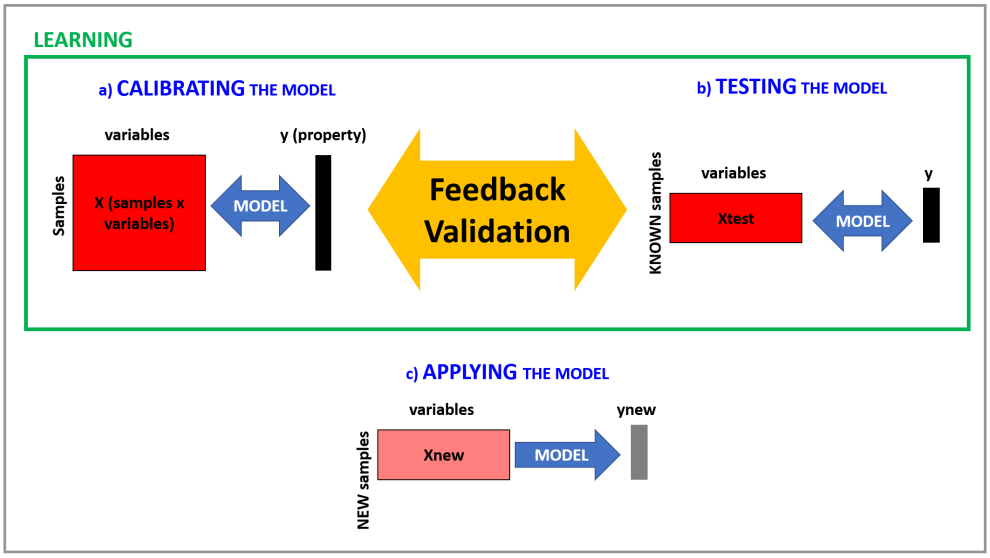


Figure 4. Machine Learning procedure showing the three stages of a) calibration, b) testing validation and c) prediction.

Let us create a matrix **Xcal** (20 x 200) of observations and a **Y** matrix **Ycal** (20 x 1) of a property that we want to measure, both of them composed of normally distributed random numbers. Now, let us apply a common multivariate regression model like Partial Least Squares (PLS). One of the key points of optimizing a PLS model is to ascertain the proper number of Latent Variables (LVs) [10]. This number can be optimized by following how the error (RMSE) and the R2 change when more LVs are included in the models. The results are represented in black in Figure 5. As can be observed, the error of the model drastically decreases with the number of LVs (Figure 5a) to such a point that the error obtained with 4 LVs is zero. Observing the predicted values obtained for different LVs (black lines in Figure 5b-d), it is clear that a perfect regression model is obtained at 4 LVs with the error at zero value and the R2 at 1. Even being the perfect calibration model, it is completely useless.

Let us create another matrix **Xval** (10 x 200) of observations and a **Y** matrix **Yval** (10 x 1) of a known property. Those two matrices are also created with normally distributed random numbers, so we are completely sure that the calibration and the validation matrices span the same space. That is, we can use the calibration models created previously to predict a **val** in such a way that we can calculate the difference between the obtained **val** and the known **Yval**. The results are displayed in red in Figure 5.

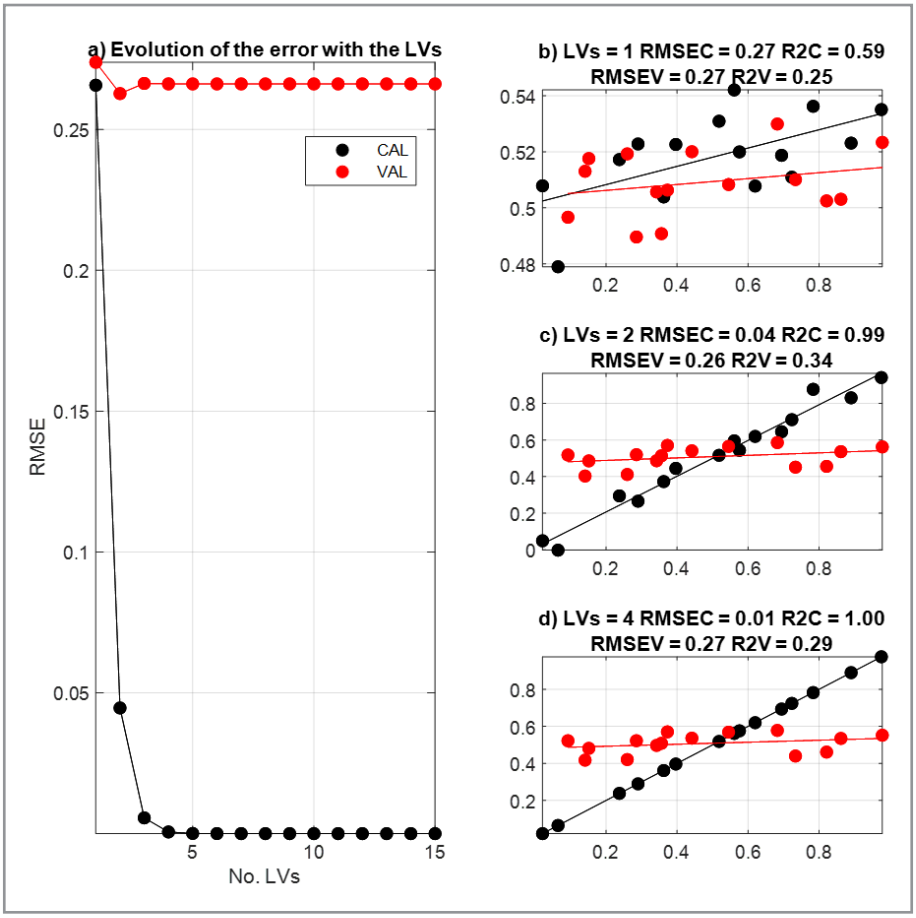


Figure 5. Applying PLS to quantify random Y values from random X values.

Having a completely perfect calibration model does not mean, at all, that this model is optimal for predicting external samples. As shown in Figure 5, it does not matter how many LVs the calibration model contains. What matters is that the real error of my model, regardless of the number of LVs, is around 0.25. This error, compared with the mean and standard deviation values of **Xval** (0.5 and 0.25, respectively), indicates that the model is predicting new samples giving values of **val** like a completely random number. The step of validation/testing the model is crucial to ensure that the model is actually working properly. There are many strategies for validating models (being cross-validation and external test set, the most common ones) [11]. It takes some time to match the most adequate validation methodology since it also involves checking the performance with different pre-processing methods, variable selection, etc. However, one thing is certain; it is necessary to:

VALIDATE YOUR MODELS!

Once the model is properly validated, we can be sure that it will be able to predict external samples with the accuracy and reliability of the validation step (Figure 4c).

Semi-supervised models

There is a family of models that falls between unsupervised and supervised modelling. They are the semi-supervised models. Semi-supervised modelling is a general approach that combines a small set of well-known labelled data (well-known class belonging) with a relatively large amount of unlabeled data (data without a pre-assumption of the class). They are especially useful when labelling data is instrumentally difficult or expensive. Semi-supervised models are normally used in classification scenarios to profit the

similarity between samples in the unsupervised data with the well-labelled samples of the supervised data. In this manner, several assumptions can be made about the belonging of the unsupervised data to different classes reflected in the supervised data. This is done by establishing more or less complicated boundaries to create clusters based on different distances between the samples in the variable space.

Deep Learning

A sub-set of Machine Learning methods is comprised of the Deep Learning (DL) algorithms. Deep Learning algorithms are also a sub-set of the well-known artificial neural networks (ANN) when the usage of multilayer structures (hidden layers) is preferred since they can handle more than one problem at the same time to give a unique answer [12]. Deep Learning algorithms are mostly based on the well-known Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN). ANN and CNN have a basic structure of inputs (the data matrix **X**), hidden layers composed of the so-called neurons and an output layer of responses. As said before, the main difference between DL networks and ANN is the complexity of the connection between the hidden layers (Figure 6). This complexity in the connections allows the feature extraction from the raw data independently, without pre-processing or pre-arranging it.

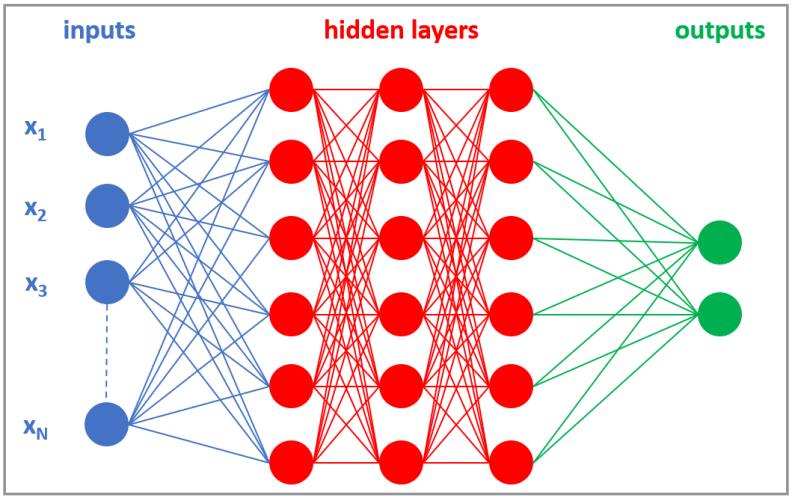


Figure 6. An example of the architecture of a Deep Convolutional Neural Network with three hidden layers.

This statement, as incredible as it appears, is not exempt from drawbacks. It should be remembered that whatever information or feature the Neural Network can extract from the raw data **MUST BE CONTAINED** in the data. The DNN learns from the data. Therefore, DNNs need an overwhelming amount of data (very comprehensive databases) in order to be able to look for the proper information and solve highly non-linear issues. Besides, due to the complexity in the structure and increasing the hidden layers, the validation (Learning) needs to be extensively and exhaustively performed to ensure that the network is not overfitting the solution. DNN trend to overfit since they are designed to model the minimum amount of variance/co-variance in the data.

Artificial Intelligence

Strictly speaking, the definition of Artificial Intelligence is the intelligence demonstrated by machines. Translated into our scenario, Artificial Intelligence is the umbrella that covers all of the previous definitions (Figure 7); that is, the application of Machine Learning, Data Mining and Deep Learning to data. Without entering into more detail, Artificial Intelligence also covers the possibility that the algorithms will be able, in the near future, to perform logical reasoning and interaction in order to improve the outcomes of the models.

However, as always, the major bottleneck in this is the availability of valid data. In a straight analogy, I always say that the most powerful machine that applies Artificial Intelligence continuously is the human brain. The human brain can analyze the analytical information that is continuously received from the analytical instruments (eye, ear, touch, taste and smell) and, in a complex procedure, offer answers or responses that will be as accurate as the information stored in the database (the memory). Together with the senses, the human brain works so well because it is continuously being trained with information and learning procedures. We normally call it education. In data analysis, this is called Training (a.k.a. Learning).

WHERE IS CHEMOMETRICS?

The term Chemometrics was coined in 1972 by Svante Wold and Bruce Kowalsky. The most accepted definition of Chemometrics refers to the chemical discipline that uses mathematical, statistical, and other derived methods employing formal logic to (a) design or select optimal measurement procedures and experiments and (b) provide maximum relevant chemical information by analyzing chemical data [3,13,14].

Another definition could be that Chemometrics is the application of Artificial Intelligence (therefore, Data Mining, Machine Learning, Artificial Neural Networks, and Deep Learning) to data coming from Natural Systems (Figure 7). Therefore, it turns out that we have been talking about Chemometrics from the very first line of this manuscript.

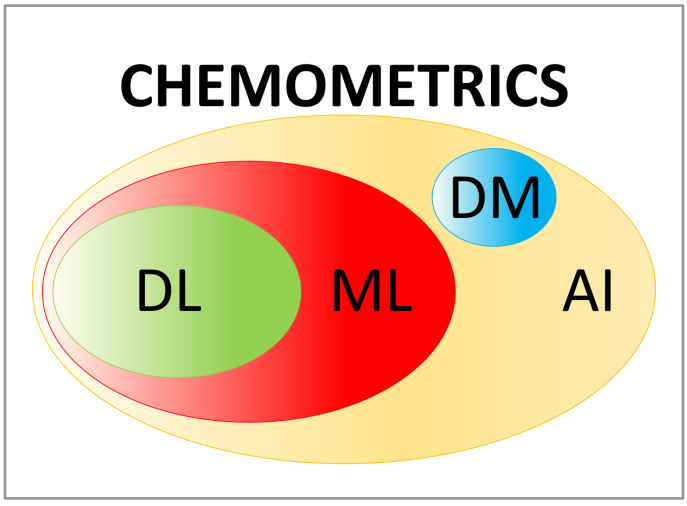


Figure 7. Chemometrics in perspective. AI, Artificial Intelligence; DM, Data Mining; ML, Machine Learning; DL, Deep Learning.

The next question would be what makes us, Chemometricians, different from Machine Learning. The straight answer would be nothing. But this statement might be too straight. As I just said, Machine Learning is part of Chemometrics. Nevertheless, I think that pertinent explanations should be given from a different perspective. Let me say that a Scientist is a person that collects and analyzes data. Therefore, the need to analyze data has been always there. For instance, PCA was proposed at the beginning of the 20th century [15,16], more than 100 years ago. Nevertheless, 100 years ago, there were not the computing capabilities that we have nowadays. We also did not have the advanced analytical instruments that we have in our laboratories or the data storage capabilities that we have now. Nevertheless, there were researchers that needed to extract complex relationships from matrices. This was not exclusively done in the field of Analytical Chemistry, but in all scientific fields. It then merged the -metrics scenarios creating new concepts like Psychometrics [17], Environmetrics [18], Econometrics [19], etc. All of them have one thing in common: the use of mathematical tools to extract relevant information from the data. Nevertheless, all of them have a strong difference: **THE DATA**.

The data coming from the previously mentioned disciplines might have certain similarities, and it could be analyzed with the same methodologies. Nevertheless, the nature, precedence and behavior of the data

are specific to the scientific scenario where the data are obtained. For example, in Chemometrics and Environmetrics, we can study temporal trends to understand the pollution of a river, as, in Econometrics, they can study temporal trends to understand the evolution of certain values in the stock market. The great difference is that a Chemometrician will better understand the environmental data than an Econometrician, and vice-versa.

DATA ARE NOT JUST DATA. THEY HAVE AN ORIGIN, A MEANING, AND A STRUCTURE THAT MUST BE FULLY UNDERSTOOD TO INTERPRET THE OUTCOMES FROM DATA ANALYSIS

Nevertheless, nowadays, there is a dangerous trend to disparage all -metric words favoring the "Data Science" words. I am a bit worried about the vast amplitude of the words "data science" and the little importance that sometimes we give to the word data. Sometimes it seems that the more programming skills you have, the better a data scientist you are. Also, we sometimes forget the "data" on the road.

ONLY BY UNDERSTANDING THE SCIENTIFIC PROBLEM AND THE COMPLEXITY OF THE DATA, WILL WE BE ABLE TO CHOOSE A PROPER DATA ANALYSIS METHODOLOGY, IF IT IS NEEDED. MACHINE LEARNING, DATA MINING, DEEP LEARNING, AND CHEMOMETRICS ARE COMPLETELY USELESS IF WE DO NOT GIVE THE DATA THE DESERVED ATTENTION

TAKE-HOME MESSAGES ABOUT HOW TO APPLY CHEMOMETRICS

During the development of this manuscript, I have included some of the references that I consider a "must-read" to understand what Chemometrics can do in your data and the benefits and major pitfalls of applying Chemometrics in analytical data [20]. To finalize this "short" manuscript, I want to give you some take-home messages about different global aspects that you should consider when applying Chemometrics to your data. They can be seen as a summary of the main concepts and statements made before.

- **A good outcome might not necessarily be a perfect model**. Always remember the GI-GO truism. Before doing anything, check the quality and information that your data can provide, as well as the purpose of the data and how it was obtained. Otherwise, we could finish with any of the following issues:
  a) Wrong hypothesis: Hoping that Chemometrics can find what is not in the data.
  b) Wrong design of the experiment: Hoping that the measurements contain variation that is not contemplated in the design of the experiment.
  c) Overuse of resources: Measuring in one instrument just because it is available in the laboratory.

  There might be temptations to use mathematical models to correct poorly designed experiments or even experimental problems with the data. Even though it might be plausible and useful in some situations, this might not be the best option since biasing the models to correct artifacts that could be easily corrected beforehand might lead to misinterpretations and underfitting situations in the prediction or interpretation of new data with the biased model.

- **Maybe you do not need fancy algorithms to solve the issue. Maybe you do not need an algorithm at all to solve the issue.** Do not even think of having data to apply Chemometrics! Think about whether your problem, with the data you have, requires Chemometrics. Think about the structure of your data, its quality, and the quality of the reference you will use. Sometimes, the reference has an associated error that is larger than the error of the data.

- **Question everything:** Chemometrics does not give you absolute answers. Chemometrics gives you validated answers in local scenarios. How big and comprehensive is the "local" scenario? As big and comprehensive as your data (database). Sentences like "This algorithm is better than this one" are only true in the conditions where those algorithms were tested. Also, you should be careful with "rules of thumb". For example, "the minimum number of PCs is set with PCs whose eigenvalue is larger than 1".

That is not true. There are situations where you will need to check into PCs whose eigenvalue is smaller than 1 but still explain interesting sources of variability in your data.

- **The Illuminati of the software:** Use the software that you can/want/like, but please, use it right. Also, always verify that the software actually does what it claims to do. This goes for commercial Machine Learning packages and for functions/libraries that you can use in software like R, Python or Matlab (among others, of course).

- **The Seven Commandments in Chemometrics:** I always tell my students that Chemometrics (and many aspects of our daily life) is mostly based on the following Commandments:
  1. Think
  2. Be patient
  3. Know your data and your goal
  4. Keep it simple
  5. VALIDATE
  6. Question everything
  7. The chemistry/physics prevails over the algorithm

FINAL COMMENTS

When applying Chemometrics to my data, I normally follow my own standardized protocol, which implies the following steps:

1) Apply the parsimony principle (Occam's razor). The simpler model, the better.
2) Check the raw data. That means to understand the data, plot the data in different manners.
3) Think about plausible artifacts and methods to solve/minimize/avoid them (pre-processing).
4) Start with the simplest methods, and draw figures. If you are using projections methods, plot the score and loading plots, regression lines, etc.
5) Be careful with assessing that a sample is an outlier.
   - In explorative/unsupervised scenarios, consider outliers as extreme samples that do not necessarily have the wrong samples.
   - In Regression and classification, check the proper figures of merit of the model. Using an exploratory method, for instance, PCA, might lead to issues when removing samples whose variance is different than expected, but their co-variance with the property you want to measure is correct.
6) GO TO LINEAR MODELS! At least, at the beginning. If the issues cannot be fixed with linear models, go step by step, increasing the complexity of the model.
7) Optimize your models. Iterate from step 1 until step 5, checking plausible pre-processings, variable selections, etc.
8) **VALIDATE**. A good model is not the one that best calibrates but the one that best predicts.
   - Test your model with the own data (cross-validation).
   - Test your model with completely external data (test, external prediction, etc.).
   - Tune/change/re-parameterize your model if needed (go from Step 2 to step 6, optimizing the figures of merit).
9) Test all plausible (and coherent) combinations of pre-processing, normalization, and variable selection methods applied to the models and then, if needed, re-parameterize the model.
10) Save all of the results. It is extremely important to keep track of your steps and to understand what is wrong or right in the obtained models.

CALL IT MACHINE LEARNING, CHEMOMETRICS OR WHATEVER YOU WANT. USE WHATEVER SOFTWARE YOU WANT OR NEED. BUT PLEASE, ALWAYS BE AWARE OF WHAT YOU ARE DOING TO YOUR DATA AND WHY YOU ARE DOING IT. AND, OF COURSE, VALIDATE YOUR MODELS!

REFERENCES

1. Amigo, J. M.; Skov, T.; Bro, R. *Chem. Rev*., **2010**, *110*, pp 4582–4605 (https://doi.org/10.1021/cr900394n).
2. Amigo, J. M. *Hyperspectral Imaging*, Elsevier, **2019**.
3. Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verberke, J. (Eds.). *Handbook of Chemometrics and Qualimetrics: Part A*, **1997** (https://doi.org/10.1016/S0922-3487(97)80056-1).
4. Bro, R. *Chemom. Intell. Lab. Syst*., **1997**, *38* (2), pp 149–171 (https://doi.org/10.1016/S0169-7439(97)00032-4).
5. Bro, R. *Crit. Rev. Anal. Chem*., **2006**, *36* (3-4), pp 279–293 (https://doi.org/10.1080/10408340600969965).
6. Leardi, R. *Anal. Chim. Acta*, **2009**, *652*, pp 161–172 (https://doi.org/10.1016/j.aca.2009.06.015).
7. Ballabio, D.; Consonni, V. *Anal. Methods*, **2013**, *5*, pp 3790-3798 (https://doi.org/10.1039/c3ay40582f).
8. Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst*., **1987**, *2* (1-3), pp 37-52 (https://doi.org/10.1016/0169-7439(87)80084-9).
9. Smilde, A. K.; Bro, R. *Anal. Methods*, **2014**, *6*, pp 2812–2831 (https://doi.org/10.1039/c3ay41907j).
10. Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst*., **2001**, *58* (2), pp 109–130 (https://doi.org/10.1016/S0169-7439(01)00155-1).
11. Westad, F.; Marini, F. *Anal. Chim. Acta*, **2015**, *893*, pp 14–24 (https://doi.org/10.1016/j.aca.2015.06.056).
12. Marini, F.; Bucci, R.; Magrì, A. L.; Magrì, A. D. *Microchem. J*., **2008**, *88* (2), pp 178-185 (https://doi.org/10.1016/j.microc.2007.11.008).
13. Brereton, R. G.; Jansen, J.; Lopes, J.; Marini, F.; Pomerantsev, A.; Rodionova, O.; Roger, J. M.; Walczak, B.; Tauler, R. *Anal. Bioanal. Chem*., **2017**, *409*, pp 5891-5899 (https://doi.org/10.1007/s00216-017-0517-1).
14. Brereton, R. G.; Jansen, J.; Lopes, J.; Marini, F.; Pomerantsev, A.; Rodionova, O.; Roger, J. M.; Walczak, B.; Tauler, R. *Anal. Bioanal. Chem*., **2018**, *410*, pp 6691-6704 (https://doi.org/10.1007/s00216-018-1283-4).
15. Pearson, K. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **1901**, *2* (11), pp 559-572 (https://doi.org/10.1080/14786440109462720).
16. Hotelling, H. *Journal of Educational Psychology*, **1933**, *24* (6), pp 471-441 (https://doi.org/10.1037/h0071325).
17. Borsboom, D. *Psychometrika*, **2006**, *71*, Article Nº 425 (https://doi.org/10.1007/s11336-006-1447-6).
18. Hunter, J. S. 1 Environmetrics: An emerging science. In: *Handbook of Statistics*. Elsevier, **1994**, *12*, pp 1-7 (https://doi.org/10.1016/S0169-7161(05)80003-3).
19. Farebrother, R. W.; Common, M. S. *Journal of the Royal Statistical Society. Series A (General)*, **1978**, *141* (3), pp 417-418 (https://doi.org/10.2307/2344828).
20. Kjeldahl, K.; Bro, R. *J. Chemom*., **2010**, *24* (7-8), pp 558-564 (https://doi.org/10.1002/cem.1346).