Chemometrics and Intelligent Laboratory Systems 127 (2013) 7–16

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Applications of a new empirical modelling framework for balancing model interpretation and prediction accuracy through the incorporation of clusters of functionally related variables

Marco S. Reis *

CIEPQPF, Department of Chemical Engineering, University of Coimbra, Rua Sílvio Lima, 3030-790 Coimbra, Portugal

ARTICLE INFO

Article history:
Received 29 January 2013
Received in revised form 26 April 2013
Accepted 12 May 2013
Available online 23 May 2013

Keywords:
Network-Induced Supervised Learning
Partial correlation
Clustering
Generalized topological overlap measure
Interpretation
Partial least squares

ABSTRACT

Current classification and regression methodologies are strongly focused on maximizing prediction accuracy. Interpretation is usually relegated to a second stage, after model estimation, where its parameters and related quantities are scrutinized for relevant information regarding the process and phenomena under analysis. Network-Induced Supervised Learning (NI-SL) is a recently proposed framework that balances the goals of prediction accuracy and interpretation [1], by adopting a modelling formalism that matches more closely the dependency structure of variables in current complex systems. This framework computes interpretable features that are incorporated in the final model, which effectively constrain the predictive space to be used. However, this restriction does not compromise prediction ability, which quite often is enhanced. Both classification and regression problems can be handled. Four widely different real world datasets were used to illustrate the main features claimed for the NI-SL framework.

1. Introduction

Quantitative models are pervasive in chemical and related sciences, being required in a broad range of applications, either tacitly, as in classification, process monitoring or fault detection, where they are usually encoded in the method's structure and parameters estimated from reference data, or in an explicit way, such as in process optimization, advanced control and product/process design, where a model must be entirely specified before such tasks can be implemented. In this context, a wide spectrum of modelling approaches can be adopted, ranging from those relying on a priori knowledge about the specific processes and phenomena going on, where models are derived from the application of the fundamental laws of nature (conservation of mass, energy and momentum), to pure data-driven approaches able to "infer" or "induce" process knowledge from data available in abundance, as happens when the predictive space of the problem under analysis is densely covered by reliable observations.

However, in practice, the analyst is often confronted with situations where the amount of a priori knowledge available is limited and, given the number of variables involved, the predictive space is also not densely populated with observations (a common consequence of the well-known "curse of dimensionality"). In these scenarios, empirical modelling approaches emerge as adequate solutions, by combining elements of the two extreme paradigms: they use some data to develop models, but the wide "gaps" in the multidimensional space are filled using the model structure postulated, based on previous information about the process and/or resulting from successive model refinements and accumulated experience. The way empirical modelling frameworks are currently developed and implemented falls into two possible categories. On one side, one finds approaches that consider each variable one-at-a-time and the model is built in a stagewise fashion. This process may be entirely sequential or involving iterations, but the distinguishing feature is the consideration of a single variable in each step. Examples include the several methodologies for constructing ordinary least squares (OLS) models (forward addition, backward removal, forward/backward stepwise) [2], classification and regression trees (CART) [3], k-nearest neighbour classification and regression methods (k-NN) [4]. On the other side, we find multivariate methods that consider simultaneously all variables involved. Even though in the end they will weight each variable differently, all variables are considered together in the analysis of the problem. Examples include partial least squares (PLS) [5–11], principal component regression (PCR) [9,10], partial least squares for discriminant analysis (PLS-DA) [12], soft independent modelling by class analogy (SIMCA) [13], linear discriminant analysis classifiers (LDA) [14,15], etc. However, neither of these two paradigms for the development of empirical models match the underlying structure of variables found in complex systems, irrespectively of their artificial (e.g., industrial facilities) or natural origin (cells, tissues, cultures of microorganisms, etc.). It is well known today that complex systems present high levels of *modularity*, *hierarchy* and *specialization* [16–19]. Systems' input variables do not operate altogether at the same time, nor do they act in an entirely isolated fashion. They are

* Tel.: +351 239 798 700; fax: +351 239 798 703.
E-mail address: marco@eq.uc.pt.

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.05.007

organized in modules, with a certain degree of specialization, that operate in the scope of one or several relevant functions in the system. The modules or cluster of variables may sometimes work simultaneously, in a synergistic way, or some of them may be silent under certain circumstances. In this context, what the real nature of systems shows, is that the basic modelling elements should be modules or clusters of variables, instead of isolated variables or the whole set of them. Thus, empirical modelling frameworks should adapt to this reality, in order to enable the development of mathematical descriptions that are closer to the fundamental nature of complex systems. Therefore, one of the features of the methodology described in this article is to construct and handle clusters of functionally related variables, instead of isolated variables or variates containing all variables under analysis (a variate is a linear combination of variables).

But developing an empirical modelling framework that better matches the systems' inner mechanism is not the only issue to be improved in current empirical methods. Another problem arises from the full priority attributed by these methods, to the maximization of prediction ability, leaving model interpretation concerns to a subsequent stage, after the model is established and validated. Examples include OLS (maximization of quality of fit), PCA (maximization of explained variation), PLS (maximization of prediction ability) and LDA or LQA (maximization of true classification rate). The tacit premise is, apparently, "the best we can predict, the best we will be able to explain". However this is often not the case. Excessive focus on prediction usually leads to situations where all degrees of freedom available are used to maximize the amount of explained variability of the response, resulting in complex and ambiguous combinations of variables that raise many difficulties to interpretational queries. Of course, the interpretation difficulties do not constitute a serious problem when the goal of the analysis is strictly centred on the fitting or prediction ability of the model, as happens for instance in calibration and soft-sensor applications. However, the current challenges for analysts and engineers involve, with an increasing frequency, the analysis and operation of progressively more complex systems where the central task is more often focused in interpreting the nature of the relationships between all the variables involved, their interaction and specific roles in the process, than on producing accurate estimates for some system properties or output variables. For instance, the goal can be to gain insight in the way the systems work for the purposes of process improvement or development of new products, in which case the information about the structure of relationships involved can be of great value for defining the next sequence of experiments. Examples of applications where this scenario can be found, include (but are not restricted to): Quality by Design in the pharmaceutical sector (seeking a suitable design space where more efforts can then be devoted in order to find a proper formulation solution), cosmetic and food products, analysis of biosystems (gene regulation, proteomics, metabolomics), analysis of formulated products, or products with complex matrices (wine, paints), and reduction of fluctuations in chemical processing industries. This problem is also found in other scenarios involving the analysis of complex systems, such as the discovery of mechanisms for complex chemical reactions, inference of the molecular origin of a disease, maximization of the throughput from metabolic reactions, etc. In all these cases, the focus in prediction ability is overtaken by the need to collect information about the structure of the system and to know which modules have an active role on the phenomena under study. In this context, Network-Induced Supervised Learning (NI-SL) addresses from the very onset of the analysis, the issue of improving the interpretational value of the results [1]. This is done by building, in the first stage, modules of variables that potentially share the same function. These modules or clusters are then used to construct the final model, whilst keeping their integrity. Therefore, in the end, it is possible to analyse which modules are playing an active role in driving the variability of the response. It is also quite easy to extract the way variables interact in the scope of each module that was found relevant, by analysing their associated weights in the selected variates.

With these two goals in mind (matching complex systems underlying mechanism and balancing interpretation and prediction accuracy), this article presents the NI-SL approach, as follows. In the next section, we describe the basic stages of the framework, and refer how the method is implemented in each one of its stages. Then, in the third section, the results achieved from the implementation of NI-SL in four real world case studies from widely different scenarios, are presented. Two of them involve classification problems (addressed with Network-Induced Classification, NI-C) and the other two, regression problems (where we apply Network-Induced Regression, NI-R). In the final section, we conclude by summarizing the main contributions of this work, and point out some interesting areas of future research in the continuation of the developments reported here.

2. Methods

In this section, we introduce the empirical modelling framework and describe its modules and methods employed. Being a supervised learning framework, it addresses both classification and regression problems. We assume that a suitable training set is available, [**X**,*y*], where **X** represents the $(n \times m)$ matrix with m variables disposed in columns, side-by-side, containing n observations, and y is the $(n \times 1)$ vector of the response variable, which can be either quantitative or categorical depending on the type of problem addressed. The NI-SL framework consists of two stages (Fig. 1). In the first stage, clusters of variables potentially involved in the same function are formed, by analysing the amount of unique information shared by pairs of variables as a measure of their direct interaction. This will be evaluated through the computation of partial correlations and the whole process is robustified by a neighbourhood analysis of all pairs of variables in question. The methodology followed is called Network-Induced Clustering. In the second stage, the basic modelling elements formed in Stage I – the clusters of functionally related variables – , are processed, selected and combined, in order to derive the required classification or regression model, leading to Network-Induced Classification (NI-C) or

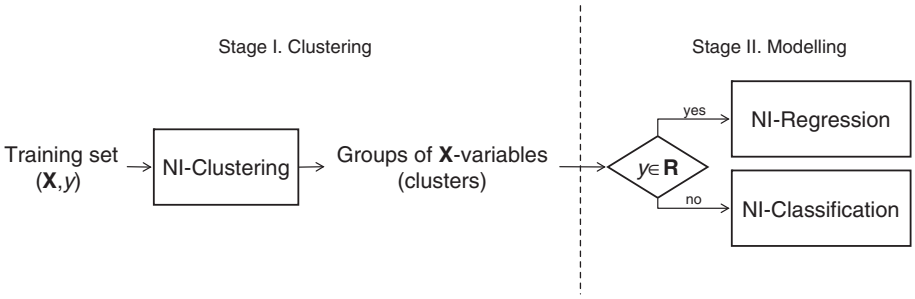

Fig. 1. The modular and stagewise structure of the empirical modelling framework NI-SL.

Network-Induced Regression (NI-R), respectively. In summary, the NI-SL framework is composed by three modules and organized in two stages: Network-Induced Clustering (Stage I), Network-Induced Classification (Stage II) and Network-Induced Regression (Stage II). In the next subsections, these three modules will be briefly described.

2.1. Network-Induced Clustering (Stage I)

The goal of this module is to construct clusters of related variables, that contribute as a group to a given system function. In order to do so in a data driven way, the direct associations between pairs of variables are first assessed using their mutual partial correlation, while controlling for others. This means that the indirect influence of other variability inducing factors is being removed, while assessing the association between each pair of variables. The order of the partial correlation reflects the number of controlled variables. In this study we have considered controlling for one and two variables, as well as for all variables but the pair whose partial correlation is being assessed. We call these 1st order (*1oPC*), 2nd order (*2oPC*) and full order (*foPC*) partial correlations, respectively. These coefficients can be computed either using analytical formulas derived specifically for the 1st order and 2nd order coefficients [1], or through a more general procedure based on regression analysis, valid for coefficients of any order. The regression-based procedure consists of the following steps. Let us assume that one wants to compute the partial correlation between variables A and B, controlling for a finite set of other variables, **S** (please note that **S** represents a set of variables and not a single variable). This partial coefficient, designated by $r_{AB\cdot\mathbf{S}}$, can then be obtained as follows:

1) Compute a regression model, where variable A is the response variable and variables in the set **S** are the regressors (or input variables), and save the regression residuals thus obtained, ε_A;
2) Do the same for variable B, i.e., estimate another regression model, where variable B is now the response variable and variables in the set **S**, the corresponding regressors, and again save the regression residuals, ε_B;
3) The partial correlation coefficient between A and B, controlling for **S**, is just the Pearson correlation coefficient between the residuals computed in steps 1) and 2), i.e., ε_A and ε_B, respectively: $r_{AB\cdot\mathbf{S}} = r_{\varepsilon_A,\varepsilon_B}$ (where $r_{\varepsilon_A,\varepsilon_B}$ stands for the Pearson correlation coefficient between ε_A and ε_B).

Partial correlations provide a finer map of the direct and causal interconnectivity of a system, something that marginal correlations cannot achieve, as they are affected by induced associations caused by other variables besides the pair under analysis. The flow of computations in the NI-Clustering module is schematically presented in the block diagram of Fig. 2.

In the first step, the partial correlation coefficients are computed for all pairs of variables, and a thresholding operation is applied in order to identify those pairs with a significant amount of direct association. The thresholding operation involves a transformation of the partial coefficients to a variable following approximately a Gaussian distribution, and the definition of a significance level to set the cut-off (more details can be found in reference [1]). The relevant directed associations are identified as 1's in a symmetric ($m \times m$) Adjacency matrix, **Adj**, while for all others 0's are inserted. More specifically, this step is accomplished as follows. If, the value of the partial correlation between variables x_i and x_j, when controlling for one variable, x_k ($r_{x_i x_j \cdot x_k}$) or two variables, x_k and x_m ($r_{x_i x_j \cdot x_k x_m}$) is found to be bellow the significance threshold for some k or m (indices of the variables under control), then a zero is immediately inserted in the corresponding entry of the adjacency matrix, $\mathbf{Adj}(i,j) = \mathbf{Adj}(j,i) = 0$, irrespectively of the value of this coefficient when controlling for other variables. The rational is to remove a direct link between two variables, when evidence is found that they do not correlate significantly when a given variable (or pair of variables) is kept constant, meaning that such pair is not directly connected.

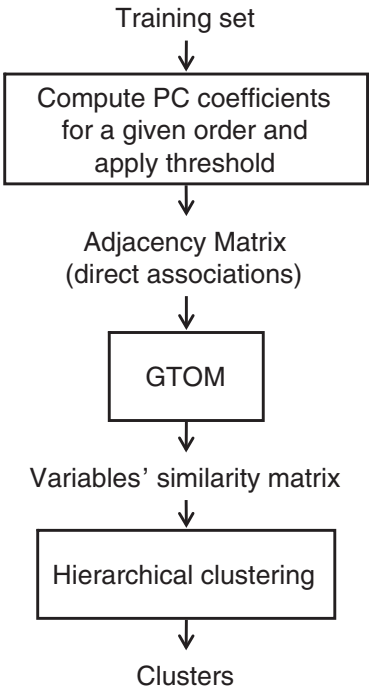

Fig. 2. Block diagram for the NI-Clustering methodology.

However, this procedure is a bit prone to spurious identifications of relevant direct associations, and therefore it was robustified in a second step, by extending the analysis to the neighbourhood of each pair of variables. In complex systems, if two variables are strongly associated, such connection is extended to their direct neighbours. In other words, variables involved in the same function tend to share common neighbours. A degree of similarity can then be computed through the topological overlap (similarity) measure (TOM), proposed by Ravasz et al. [16], which is defined by:

$$TOM(i,j) = \frac{l(i,j) + \mathbf{Adj}(i,j)}{\min\left\{k_i, k_j\right\} + 1 - \mathbf{Adj}(i,j)} \tag{1}$$

where $l(i,j) = \sum_{u \neq i,j} \mathbf{Adj}(i,u) \times \mathbf{Adj}(u,j)$, i.e., is the number of neighbours shared by nodes "*i*" and "*j*" and $k_i = \sum_{u \neq i} \mathbf{Adj}(i,u)$, the number of links in node "*i*". The above definition of the topological overlap measure was also extended to higher-order neighbourhoods in order to improve the robustification process, leading to the generalized topological overlap measure (GTOM) [20]. GTOM has a parameter, l, defining the order of the neighbours contemplated in the analysis: $l = 0$ is equivalent to consider only the direct links of the adjacency matrix, and $l = 1$ corresponds to the definition of TOM. In this work, we have considered $l = 2$ in the computation of the GTOM measure of pairwise interconnectedness.

In the third step the GTOM similarity matrix is converted into a distance matrix, $\mathbf{D}_l$ (the subscript l indicates the order of the neighbourhood considered in the GTOM computations), through, $\mathbf{D}_l(i,j) = 1 - \mathrm{GTOM}_l(i,j)$, in order to apply an hierarchical clustering algorithm (linkage criteria set to the unweighted average distance). This algorithm produces a hierarchical tree (dendrogram) with the clustering distance for increasingly larger groups, which can then be analysed in order to identify the natural clusters that are present. In this step, we suggest analysing not only the dendrogram and the variables' clustering distances, but also the TOM plot (a matrix-like plot, where the topological overlap measure is codified by colours reflecting the similarity degree between variables, after they are reordered according to the results of the clustering algorithm) and the "silhouette" values for each variable,

that provides a measure of how close a variable belonging to one cluster is from the variables assigned to other clusters. Analysing the outcomes from all these different analysis tools, it is possible to define the number of natural clusters present in the dataset (**NCLUST**) and to obtain their composition. With the variable clusters obtained in this way, the empirical modelling framework proceeds to Stage II, where they will be employed in the development of a supervised model for regression or classification.

2.2. Network-Induced Classification (Stage II)

In this case, a classification model is derived using the clusters of variables formed in Stage I. As the basic modelling elements are the clusters of variables identified during the first stage, Stage II begins by constructing the linear combinations of each cluster of variables that present more predictive power regarding the response variables in question (in this work only one response variable is considered, but the procedures can be easily extended to multiple response variables). These linear combinations will be here referred as variates. Thus, for the case of a categorical response variable, the variates to be computed in each cluster are those providing the best separation possible for the observations with distinct class labels. We have used the first Fisher linear discriminants of each cluster to accomplish this goal. The maximum number of variates (or linear discriminants) to extract in each cluster is a user defined parameter, called **NVC**. Usually this parameter does not have to be higher than two, as each cluster of functionally-related variables is not expected to contribute to the definition of many classes, but this may change from application to application. The actual number of variates computed for each cluster may however be lower than **NVC**, as the theory of Fisher discriminant analysis establishes an upper-bound on the number of linear discriminants to extract: $\min\{m_i, g-1\}$, where m_i is number of variables in cluster i and g the number of classes present. Therefore, the total number of variates computed for cluster i is given by: $\min\{m_i, g-1, \mathbf{NVC}\}, i = 1,\ldots,\mathbf{NCLUST}$.

Once computed the discriminant variates for each cluster using the training dataset, all of them are gathered, side-by-side, and subjected to a selection process. In fact, some of these variates may significantly contribute to the prediction of the class labels, isolated or in combination with others, but others contain no relevant prediction power regarding the phenomena under analysis. Therefore, these new variables (the discriminant variates), that summarize the best each cluster has to contribute to the final predictive model, will be selected and combined in the best way possible, using a given classifier (in our case, the linear classifier was adopted for combining the variates). The selection procedure consists in exploring, in an exhaustive way, the predictive classification power of all variates individually, then all the combination of 2, 3, 4, etc., until a pre-defined maximum is achieved. The classification power associated with each particular combination is assessed through Monte Carlo cross-validation. This procedure consists of splitting samples into train and test sets, according to a user specified split fraction (in the present case the test set always contained 20% of the total number of samples available), after which a classifier is trained and tested. This classifier is then applied to the test set, and a measure of classification ability is computed and saved. In this work, measures of global accuracy (the overall mean of the percentage of correct class predictions obtained in each Monte Carlo cross-validation trial) and class-mean accuracy (average of the mean accuracies for each class, given by the percentages of correct class attributions in each Monte Carlo trial, for each class label) were considered. The final performance is determined from the classification results obtained in several successive Monte Carlo cross-validation runs (in this case, 20).

The analysis of the results obtained for all the different combinations is made by analysing the best combinations of variates for each size (1, 2, 3,…) and the evolution of the associated prediction performance scores. The final number of combinations to use in the model, is the one leading to a minimum in this curve, or after which the evolution of the prediction scores levels off and a plateau is achieved. This will lead to the definition of the maximum number of variates to consider in the construction of the final model, **NVmod**.

With the three method's parameters defined (**NCLUST**, **NVC**, **NVMod**), one can estimate the best model using the entire training set, and apply it directly to any new test set. This procedure is called Network-Induced Classification (NI-C) and is schematically presented in Fig. 3, corresponding to the branch on the left-side of this figure, in Stage II. An interesting feature of this procedure is that some variates may not be selected at all and therefore the corresponding variables will not be considered in the final model. This variable selection feature leads to parsimonious models, eliminating blocks of variables that are not significantly contributing to the phenomena under analysis. On the other hand, from the variates that were selected and their weights, it is possible to infer and interpret the role of each cluster of related variables in the classification task.

2.3. Network-Induced Regression (Stage II)

A regression model is a function of the input variables that provides an estimate of the population mean of the response variable. In the proposed empirical modelling framework, and following the same reasoning of NI-C, the inputs will be variates computed from the clusters of variables identified during Stage I. These variates are now composed according to their prediction power for the response. In this work, the variates are the first PLS components (or latent variables) in a PLS model involving each cluster of variables and the response. The next steps of the network-induced regression (NI-R), also mimic the operations performed in NI-C (Fig. 3). Together, NI-C and NI-R constitute a consistent and coherent framework for Network-Induced Supervised Learning, able to deal with outputs of both categorical and quantitative nature. In NI-R the variates are also collected and organized side-by-side, in order to select those carrying relevant prediction power regarding the response. The basic variate selection procedure for NI-R is similar to the one described before for NI-C, but now one adopts a regression methodology for combining each set of variates, which can be OLS or PLS (the variates do not usually show strong collinearity problems, as happens with the original variables), and the prediction performance metrics are based on the prediction accuracy of the models, namely root mean square error of cross-validation (RMSECV). The number of variates to include in the model (**NVMod**) can be decided based on the graphical analysis of the evolution of the RMSECV obtained for the best variate combinations of size 1, 2, 3, etc. If this number is such that $\mathbf{NVMod} < 5$, the procedure of exhaustively considering all possible combinations is still feasible and can be applied. However, for higher number of combinations of variates to explore, it may encompass a significant computational load, as in each stage, a number of cross-validation trials are involved. Thus, for $\mathbf{NVMod} \geq 5$, an alternative variate selection methodology was adopted, that consists in implementing a "forward stepwise" selection methodology. This approach successively selects variates as long as they bring a statistically significant contribution to the amount of y-variability explained by the model (evaluated with a partial F-test). Variates that have been selected before may also be discarded later on, if their contribution becomes redundant or not significant, after the introduction of other variates [2,21]. This procedure runs much faster, and usually leads to good predictive models. A similar accelerating procedure can be also implemented for NI-C, but we have not experienced that need in all the examples analysed so far.

Once the variates to include in the model are established, the final model is estimated using the training set. This model can be explored by looking at i) the composition of clusters formed and, in particular, the ones regarding variates with a dominant role in the final model; ii) the loadings and loading weights of the variables in each variate

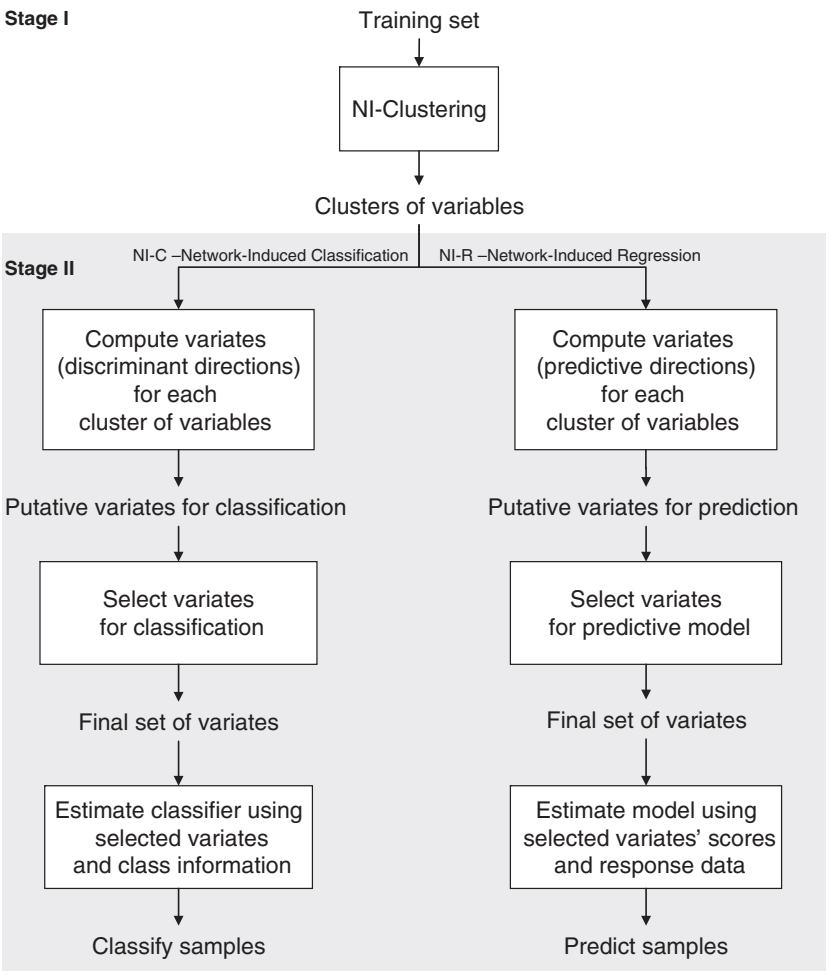

Fig. 3. Schematic representation of the sequence of steps underlying the Stage II methodologies of Network-Induced Classification (on the left-hand side) and Network-Induced Regression (on the right-hand side).

(specially in the dominant ones). It can also be straightforwardly applied to future data, using only the variables included in the selected variates.

In summary, NI-R also operates in a constrained space of regressors or input variables, composed by variates (linear combinations of input variables), each one of them containing variables with topological similarities, i.e., that may share similar functional roles in the system or belong to the same functional modules. The use of such variates instead of the original variables may limit, in principle, the prediction ability of the method, with the goal of bringing interpretation added-value to the estimated model, as the predictive elements (variates), now have a functional meaning by design. However, we have found out that such a compromise does not always result in a significant loss in prediction ability, in fact very rarely it does so, and quite often leads to similar prediction ability scores or even better results, as a consequence of the more parsimonious nature of the models obtained which can lead to more stable predictions, due to a more assertive definition of the predictive space. These features of the NI-SL framework will be illustrated in the four case studies presented in the next section.

3. Results

In this section we will present four real world case studies where the empirical modelling framework for supervised learning, NI-SL, is applied and comparatively assessed. In order to enable a sound comparison, we have adopted for benchmarking methods the unrestricted counterparts of the methodologies used for classification and regression. In other words, as the NI-SL framework operates over the restricted predictive space formed by the variates (computed in the NI-Clustering module) by combining them using a classifier (NI-C) or a suitable regression methodology (NI-R), we have also used the same classifiers and regression methodologies over the original set of variables, as benchmark methods. By doing this, any difference in the results will be strictly attributed to the effects arising from the imposition of interpretational-oriented restrictions (in the form of variates), as other factors were kept fixed. These results will be presented and discussed, after introducing the datasets and the performance indicators on the basis of which the comparative assessment is made.

3.1. Datasets

The real world case studies addressed in this work, are the following:

1) *Waviness dataset*. The dataset has the structure [X(29 × 12),2] (i.e., it is composed by 12 variables and 29 samples from 2 class categories). Input variables regard geometrical-oriented features that summarize different aspects of an accurate profile taken from the surface of paper sheets, at the waviness scale (which is a large length-scale surface phenomenon, corresponding to visible deviations from the flat shape), using high-resolution mechanical stylus profilometry [22,23]. The goal is to predict the sensorial evaluation made by a panel of experts (class labels "Pass"/"Fail"),

in order to develop a tool for the quick inspection of the surface quality of paper, at line and in situ.

2) *Gene expression dataset*. This is a classical dataset containing gene expression data for samples arising from patients with different types of Leukaemia: [24] acute lymphoblastic leukaemia (ALL), subdivided according to its lineage (ALL-B and ALL-T) and acute myeloid leukaemia (AML). We will use a truncated version of the original dataset, where the 50 genes suggested in reference [25] were processed and screened using simple methods specific for this type of measurements, namely (i) a minimum of 100 and a maximum of 16,000 are imposed on the expression levels for each gene *i*, (p_i); and (ii) all genes whose expression range over the samples is lower then 500, $range(p_i) \leq 500$, and whose ratio $\max(p_i)/\min(p_i) \leq 5$, are discarded. All values are also logarithmically transformed (base 10). The final dataset as the following structure: [X(72 × 28),3].
3) *Madeira wine dataset*. This dataset regards a study of the long-term ageing process of the Portuguese Madeira wine [26]. Samples were collected from wines of the same grape variety ("Malvasy") with different ageing times, spanning a range of approximately 20 years of ageing in oak casks. All samples were analysed with a variety of analytical instrumentation techniques, including GC–MS, HPLC–DAD and UV–vis spectrometry. For illustration purposes, the data collected with HPLC–DAD will be used here, which contains the composition of 25 compounds present in the wine matrix (input variables). The output variable is the ageing time of the wine samples. The goal is to develop a predictive model for such an important feature of wine from which the wine's market value is, to a large extent, established, and to analyse which chemical compounds might be closely linked to the ageing process and the chemical phenomena taking place. Such a model may find very interesting applications in process monitoring and fraud detection tasks, for instance.
4) *PAH dataset*. This is a calibration case study [27], where two datasets are available for analysis (a train and a test dataset). Each dataset consists of 25 Electronic Absorption Spectroscopy (EAS) spectra, collected for 27 wavelengths, from 220 nm to 350 nm, in 5 nm intervals, regarding different mixtures of polyaromatic hydrocarbons (PAHs). The train dataset is used for developing the model and estimating its parameters, as well as to compute the MC-CV and LOO-CV prediction quality metrics, while the test dataset is used for independent testing. The outputs to be predicted are the concentrations of the PAHs for each sample. For illustration purposes, from the ten PAHs analysed in the original reference [27], we present here the results regarding five of them, for which the benchmark method, PLS, presents good prediction accuracy, namely: pyrene (Py), anthracene (Anth), chrysene (Chry), benzanthracene (Benz), and fluorene (Fluore).

3.2. Performance indicators

In order to enable a thorough and objective evaluation of the NI-SL framework with its unrestricted benchmarks, a set of quantitative performance indicators are adopted. They will cover both the fitting ability and prediction accuracy of the methodologies, in order to evaluate how NI-R and NI-C perform on these performance dimensions, besides providing higher interpretation power (a feature that cannot be easily quantified). The performance indicators used in this work are briefly described in the following sections.

3.2.1. Re-substitution accuracy

This is just a measure of the ability to estimate the same observations used to derive the model. Therefore what is being assessed is the quality of fit of the methodologies (not prediction), a feature that may have some ambiguousness when presenting high values, but that is quite useful for screening out bad models: if a model is not able to describe well the data used for its development, then it is not an acceptable model and should be discarded. For classification, we have used the *global accuracy* (overall percentage of correct class assignments) as a measure of accuracy. For regression, we report the root mean square error of calibration ($RMSE_C$) and the respective coefficient of determination, R^2:

$$RMSE_C = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{3}$$

where, y_i and $\hat{y}_i$ are the observed and estimated values of the response in the *i*th observation.

3.2.2. Monte-Carlo cross-validation accuracy (MC-CV)

These and the following two measures address the prediction accuracy of the methodologies. Therefore, they always encompass the prediction of samples not included in the datasets used to estimate the model. In Monte-Carlo cross-validation, a random sample with circa 20% of the observations in the training dataset is set aside. The remaining observations are used to estimate a model, and this model is used to predict the values of the response in the observations left aside. In classification, the performance indicator is obtained by averaging the classification accuracy scores obtained in all trials. As for regression, the following quantities are computed in each cross-validation trial (k) (the first corresponds to a measure of the prediction error while the second represents an extension of the coefficient of determination, R^2, to the cross-validation context):

$$RMSE_{CV}(k) = \sqrt{\frac{\sum_{i=1}^{n_{out}}(y_i - \hat{y}_i)^2}{n_{out}}} \tag{4}$$

$$R^2_{CV}(k) = 1 - \frac{\sum_{i=1}^{n_{out}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{out}}(y_i - \overline{y}_{out})^2} \tag{5}$$

where n_{out} represents the number of observations left out in the kth cross-validation run ($k = 1,\ldots,20$) and $\overline{y}_{out}$ the respective mean of the output variable. This procedure is repeated a number of times (we have performed 20 repetitions), in order to mitigate distortions from the random allocation of observations in the two groups. Finally, the overall mean and the associated standard deviation for these two quantities ($RMSE_{CV}$ and R^2_{CV}) are computed and reported in the result tables.

This method has some limitations however, particularly regarding the assessment of the classification accuracy. This happens when the number of samples is small relatively to the number of class labels available. In these circumstances, some classes may not be properly represented in the training sets for some trials, leading to poor classification models and therefore to bad classification performances (a non-stratified sampling cross-validation approach was adopted in this work, as it tends to provide more conservative estimates of the methods' accuracy). This may lead to quite unreliable estimates of the methods' performance. In these situations, the following methodology is more appropriate.

3.2.3. Leave-one-out cross-validation accuracy (LOO-CV)

This cross-validation procedure is similar to MC-CV, but now only one observation is left aside in turn to test the classification or

regression models, and the procedure stops when all observations were left aside once and only once (there are no repetitions in this procedure). The classification performance indicator is the global accuracy of the test samples. The regression measures of performance are based on the following quantities:

$$RMSE_{LOO-CV}(k) = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{(i)}-\hat{y}_{(i)}\right)^2}{n}} \quad (6)$$

$$R^2_{LOO-CV}(k) = 1-\frac{\sum_{i=1}^{n}\left(y_{(i)}-\hat{y}_{(i)}\right)^2}{\sum_{i=1}^{n}\left(y_{(i)}-\overline{y}\right)^2} \quad (7)$$

where, $y_{(i)}$ and $\hat{y}_{(i)}$ stand for the value of the output variable and its estimate, respectively, in the trial where the ith observation is left aside.

3.2.4. External validation with a test set (when available)

When an independent test dataset is available, the prediction ability will also be assessed by computing the performance indicators presented before. For instance, regarding regression, they correspond to the root mean square error of prediction ($RMSE_P$) and the predicted R^2 (R^2_{pred}), defined by:

$$RMSE_P = \sqrt{\frac{\sum_{i=1}^{n_{test}}(y_i-\hat{y}_i)^2}{n_{test}}} \quad (8)$$

$$R^2_{pred} = 1-\frac{\sum_{i=1}^{n_{test}}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n_{test}}(y_i-\overline{y})^2} \quad (9)$$

(n_{test} stands for the number of observations available in the test dataset).

3.3. Results

In this section we present the results regarding the application of the NI-SL framework on the four real world datasets from widely different areas (pulp and paper industry, health molecular biology, wine production and chemistry). The idea is to illustrate the main features of the framework, its flexibility and potential of application in different practical scenarios, as well as to conduct a sound assessment of its performance when compared to the benchmark methodologies. As referred before, the benchmark methods are the counterparts of NI-C and NI-R without the inclusion of the variates formed in the NI-Clustering step. In other words, they are just the unrestricted versions of the methodologies employed in NI-C and NI-R, as no constraints are imposed in the way variables are combined, contrary to the structure imposed in the construction of the variates in the empirical modelling framework. The performance indicators for the assessment study regarding NI-C are presented in Table 1 (datasets 1 and 2) and for NI-R, in Table 2 (datasets 3 and 4).

3.3.1. Results for the waviness dataset

The results obtained for this dataset (Table 1), illustrate some interesting features of the NI-SL approach. Even though NI-C does not present the same quality of fit as the unconstrained approach, as degrees of freedom are being removed with the preliminary construction of interpretational-driven variates, it clearly outperforms the benchmark in the prediction accuracy cross-validation tests. This means that the selected variates carry relevant prediction power, and that the more parsimonious and structured model obtained (therefore more interpretable) is not compromising prediction accuracy goals. One can therefore conclude that interpretational and predictive goals do not have to be competitive or mutual exclusive, but may be achieved simultaneously. Due to space restrictions we will not present all results obtained in the analysis of the case studies, but we would like to point out that, in this case, the most relevant variables were found to arise from one of the clusters analysed (Cluster 1), composed by only five out of the twelve variables available. This cluster provides the best variates among all the combinations of size 1 and 2. The variables of this cluster can thus be analysed in order to find out what common phenomena are they collectively describing, that make them the most useful for describing the several classes in question.

3.3.2. Results for the gene expression dataset

The performance indicator scores for NI-C and the benchmark method are, in this case, quite similar (Table 1), meaning that the introduction of the variates' constraints is once again not compromising the classification accuracy. However, the way the clusters are selected and the prevalence with which they are selected for various combinations sizes, carry important information about which genes are playing a more active role in discriminating the several physiological states in question (Leukaemia types). In this example, one cluster composed by 9 genes, and another cluster, with 3 genes, lead the list of the more interesting groups of variables (Fig. 4). Therefore, if the topological clustering is indeed sensible to their mutual and direct associations in the scope of some common functions, our approach should be able to point out those involved in the development of the physiological conditions. In this context, not only are we achieving good prediction ability of the disease types from gene expression measurements, but also finding out (i.e., discovering) "which genes" are involved in "which common functions".

Table 1
Comparative assessment of NI-C. Global accuracy results (%) computed using the re-substitution approach, Monte-Carlo cross-validation (MC-CV, with standard deviation inside parenthesis) and leave-one-out cross-validation (LOO-CV). The method adjustable parameters used, and the order of the partial correlation adopted for computing the adjacency matrix, are also referred.

Dataset	Partial correlation order	Method's adjustable parameters			Global accuracy (%) Proposed method Benchmark		
		NCLUST	**NVC**	**NVMod**	Re-subs.	MC-CV	LOO-CV
Waviness	*foPC*	3	2	2	96.6	90.8 (10.1)	89.7
					100	75.8 (14.8)	82.8
Gene expression	*foPC*	9	1	3	93.1	83.6 (10.9)	84.7
					98.6	80.7 (9.9)	86.1
	foPC	9	1	4	97.2	80.4 (11.6)	86.1
					98.6	80.7 (9.9)	86.1

Legend: *foPC* — full-order partial correlation. Other parameters are defined in the text.

Table 2
Comparative assessment of NI-R. Results for NI-R and the benchmark methods (PLS, OLS): R^2 (coefficient of determination) and RMSE (root mean square error) for the training set (re-substitution), Monte Carlo cross-validation (MC-CV), Leave-one-out cross-validation (LOO-CV) and, when available, for the independent test set. In the case of MC-CV, both the mean and standard deviation of the results are presented (the standard deviation appears inside round brackets).

Dataset	Partial correlation order[a]	Method's adjustable parameters			R^2/RMSE, NI-R Benchmarks (OLS, PLS)			
		NCLUST	**NVC**	**NVMod**	Re-subs.	MC-CV	LOO-CV	Test set
Madeira wine	*2oPC*	11	2	5	NI-R 1.076/0.963 OLS 0.5110/0.9917 PLS (#LV = 5) 0.8592/0.9766	NI-R 1.148 (0.1823)/0.9452 (0.0345) OLS 1.422 (0.8950)/0.9036 (0.1394) PLS (#LV = 5) 1.154 (0.2146)/0.9484 (0.0268)	NI-R 1.190/0.955 OLS 1.116/0.9605 PLS (#LV = 5) 1.095/0.9619	–
PAH (Py)	*2oPC*	20	2	2	NI-R 0.04/0.965 PLS (#LV = 10) 0.0190/0.9922	NI-R 0.041 (0.0113)/0.921 (0.0616) PLS (#LV = 10) 0.046 (0.0148)/0.927 (0.0602)	NI-R 0.046/0.954 PLS (#LV = 10) 0.046/0.955	NI-R 0.0852/0.8428 PLS (#LV = 10) 0.0870/0.8361
PAH (Anth)	*2oPC*	20	2	7	NI-R 0.015/0.963 PLS (#LV = 14) 0.00783/0.9902	NI-R 0.025 (0.0056)/0.880 (0.0451) PLS (#LV = 14) 0.034 (0.0093)/0.624 (0.6936)	NI-R 0.023/0.913 PLS (#LV = 14) 0.031/0.847	NI-R 0.0254/0.8973 PLS (#LV = 14) 0.0171/0.9532
PAH (Chry)	*2oPC*	20	2	5	NI-R 0.022/0.982 PLS (#LV = 11) 0.00800/0.9975	NI-R 0.036 (0.0111)/0.864 (0.1678) PLS (#LV = 11) 0.035 (0.0223)/0.930 (0.1105)	NI-R 0.038/0.944 PLS (#LV = 11) 0.023/0.980	NI-R 0.0345/0.9526 PLS (#LV = 11) 0.0268/0.9713
PAH (Benz)	*2oPC*	20	2	2	NI-R 0.093/0.985 PLS (#LV = 9) 0.03301/0.9981	NI-R 0.114 (0.0207)/0.963 (0.0227) PLS (#LV = 9) 0.083 (0.0245)/0.983 (0.0139)	NI-R 0.105/0.981 PLS (#LV = 9) 0.074/0.991	NI-R 0.1007/0.9826 PLS (#LV = 9) 0.0548/0.9949
PAH (Fluore)	*2oPC*	20	2	10	NI-R 0.047/0.972 PLS (#LV = 10) 0.04848/0.9706	NI-R 0.087 (0.0308)/0.846 (0.1088) PLS (#LV = 10) 0.145 (0.0840)/0.616 (0.5073)	NI-R 0.094/0.890 PLS (#LV = 10) 0.112/0.843	NI-R 0.1200/0.8201 PLS (#LV = 10) 0.1080/0.8542

[a] *2oPC* — second-order partial correlation. Other parameters are defined in the text.

3.3.3. Results for the Madeira wine dataset

This dataset regards a study focused on the analysis of the long-term ageing process of the Portuguese Madeira wine [26]. Analysing first the NI-clustering outcomes, 9 clusters were selected for generating the variates, from which 5 should be retained in the final model, using the fast procedure for model building based on the forward stepwise variate selection algorithm. From the results presented in Table 2, one can verify that the model obtained shows a remarkable prediction ability (MC-CV and LOO-CV results), well in line with the results exhibited by the unconstrained benchmark methods (PLS and OLS). A specialist can analyse and comment the chemical compounds belonging to each cluster, especially those picked up by the model, but for the purposes of this article we will provide some general comments supported by Table 3, which presents the composition of the clusters picked up by the forward stepwise selection algorithm, in the order they were selected. The associated R^2 values provide an indication of the relative importance of each cluster in predicting wine age. Cluster 1 dominates in importance, explaining by itself 91% of the wine age variability. Among the components that integrate this cluster, we can verify the presence, for instance, of HMF and Furfural, which were already identified as being connected with the ageing process of Madeira wine in particular (namely in sugars degradation processes) and Caffeic acid, p-Coumaric acid, which are associated with microbial activity throughout the ageing process. In cluster 2, the next cluster in importance, one finds Protocatechuic acid, which is also known to be involved in wine ageing in general, not specifically only for Madeira wine and Ferulic acid, also associated with microbial activity in the ageing process. From both clusters 1 and 2, one also finds Gallic acid, Vanillin and Ellagic acid, well known to be involved in reactions with the oak casks along time, where Madeira wine is aged. The interpretational analysis is therefore quite consistent with the background knowledge available in the field. However there also some components in the main clusters that were not identified as important in the past. Their presence should be better scrutinized, as they may be important

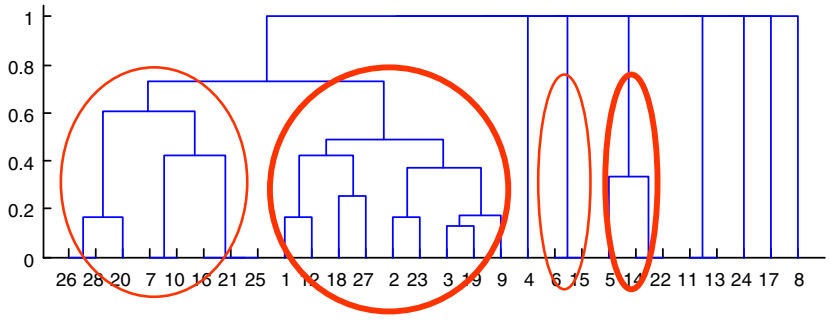

Fig. 4. Dendrogram for the hierarchical clustering of variables using GTOM similarity's distance, for the gene expression dataset. From the results obtained in the variate selection module, the clusters indicated with bold ellipses play a more active role in classification, followed by those identified with finer ellipses.

Table 3
Composition of the most important clusters of variables. The order by which they were selected to incorporate the model is indicated, as well as the associated model R^2 for the training set.

Order	R^2	Composition of the cluster
1	0.91	HMF, Furfural, nd2, nd5, Vanillin, Ferulic acid, Myricetin
2, 3	0.942	Gallic acid, Protocatechuic acid, (−)-Epigallocatechin, Caffeic acid, p-Coumaric acid, Ellagic acid
4	0.955	(+)-Catechin, (−)-Epicatechin
5	0.960	trans-Resveratrol

predictors or just associated with other key components during a part of the ageing process.

3.3.4. Results for the PAH dataset

In this last case study, the NI-R framework is tested in a context where the benchmark method PLS, finds one of its more typical and successful application scenarios: multivariate calibration. In these conditions, only PLS was used as a benchmark method, due to its ability to handle multicollinearity in the input variables (as typically happens in multivariate calibration problems) and the fact that OLS cannot even be employed, given that, in this case, the matrix $\mathbf{X}^T\mathbf{X}$ is singular (the number of input variables available exceeds the number of observations, a situation that is also rather common in multivariate calibration applications). However, OLS was used for combining the selected variates in the NI-R framework.

Analysing the results presented in Table 2 for the 5 PAHs considered in this study, one can verify that the accuracy of the predictions obtained with NI-R is comparable with that presented by the benchmark PLS method, sometimes being slightly inferior, others superior, but always of the same magnitude. The importance of each wavelength can be assessed with the Variable Importance in Projection for NI-R, $VIP_{NI\text{-}R}$, defined as:

$$VIP_{NI-R}(k) = \sum_{j \in \Omega_k} \left\{ \beta_j^2 \times w_{k,j}^2 \right\} \tag{10}$$

where Ω_k stands for the set of variates containing variable k, and $w_{k,j}$ is the corresponding kth entry of the PLS weighting vector used to compute the jth variate. β_j stands for the regression coefficient affecting the jth variate in the final model involving the selected variates (for proper interpretation this parameter requires variables to be previously "autoscaled"). Fig. 5 presents the $VIP_{NI\text{-}R}$ plots obtained for the models developed, where it is clear the variable selection ability of the proposed methodology, leading to parsimonious models that still present good prediction capabilities.

3.4. Discussion

The analysis of the case studies presented in this previous section, arising from widely different application scenarios, illustrates the potential of the NI-SL empirical modelling framework in the development of parsimonious and interpretable models. These models retain much of the prediction ability characteristics of their unrestricted counterparts, and, not rarely, even lead to improvements in the accuracy of the results obtained.

The modular nature of the approach allows for some flexibility on its use. For instance, in Stage I, the NI-Clustering methodology to select the clusters of functionally-related variables in a data-driven way, can be by-passed if the groups of variables can be formed using a priori knowledge about the process. The empirical modelling framework does not require any additional change and it can be directly applied in this context. The important point is that the clusters of variables have

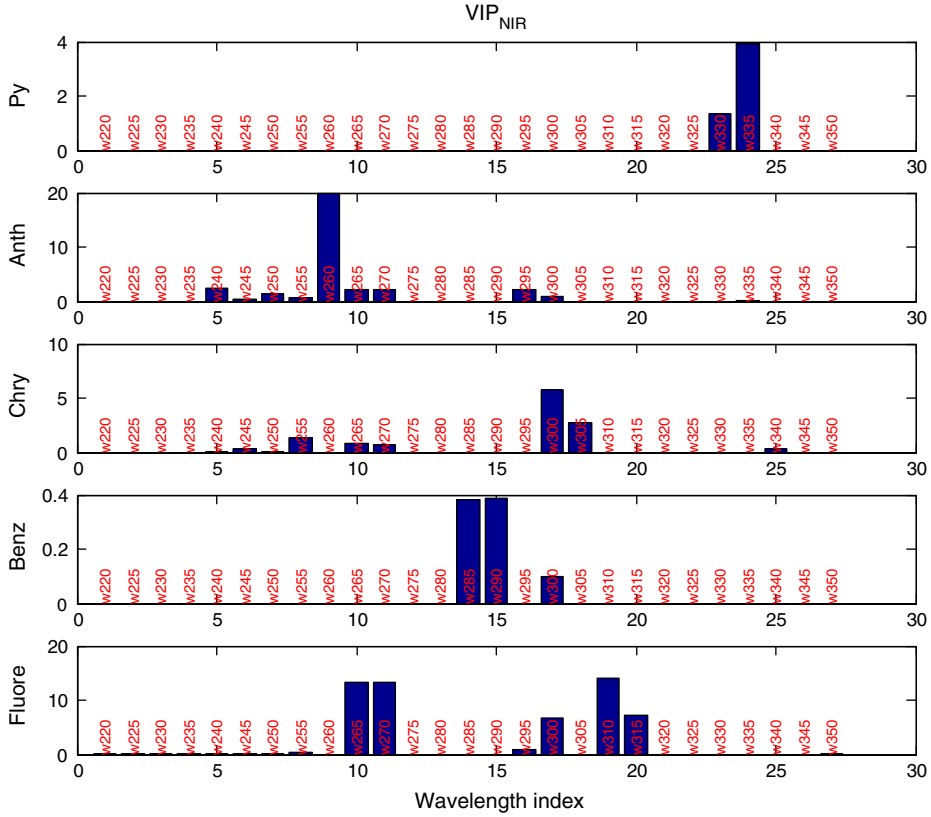

Fig. 5. $VIP_{NI\text{-}R}$ plots for the NI-R models developed for predicting the PAH concentrations in the PAH dataset.

some fundamental meaning, in the scope of a function they perform, in order to bring interpretation potential to the final model.

In the same way, other regression or classification methods can be employed, if necessary, for instance due to the particular distribution of the classes in the feature space or some characteristic that should be present in the regression model. However, from our accumulated experience in testing and using the NI-SL framework, these cases are relatively rare, and most of the times it can be straightforwardly applied in the original form.

4. Conclusions

In this article, we have illustrated the use of an empirical modelling framework, NI-SL, which is able to bring interpretational concerns to the forefront of classification and regression problems, without compromising their prediction ability. The results presented, intentionally involving the analysis of real world datasets from very different application domains (pulp and paper industry, health molecular biology, wine production and chemistry), do illustrate such features, as well as the framework's flexibility and potential added value for data analysis. Therefore, we believe that NI-SL provides a good alternative to current methods in situations where the interpretation of the models obtained is also an important outcome from data analysis, besides the prediction accuracy, as happens, for instance, during process/product development and improvement stages, or in the analysis of metabolic/genomic data, where the aim is centred in finding out the network of relationships between the several metabolites/genes analysed, that might be responsible for a given state, usually abnormal or pathological.

In future work, we aim to contemplate other application scenarios, and achieve a more profound knowledge about the implications of using partial correlation coefficients of different orders, in the NI-clustering algorithm.

Acknowledgements

The author acknowledges the support of Ana Cristina Rebola Pereira in the interpretation of the results for the Madeira wine dataset. The author acknowledges financial support through project PTDC/EQU-ESI/108374/2008 co-financed by the Portuguese FCT and European Union's FEDER through “Eixo I do ProgramaOperacional Factores de Competividade (POFC)” of QREN (with ref. FCOMP-01-0124-FEDER-010397).

References

[1] M.S. Reis, AICHE Journal (2013), http://dx.doi.org/10.1002/aic.13946.
[2] N.R. Draper, H. Smith, Applied Regression Analysis, Wiley, NY, 1998.
[3] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Chapman and Hall/CRC, Boca Raton, 1984.
[4] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, NY, 2001.
[5] P. Geladi, B.R. Kowalski, Analytica Chimica Acta 185 (1986) 1–17.
[6] D.M. Haaland, E.V. Thomas, Analytical Chemistry 60 (1988) 1193–1202.
[7] I.S. Helland, Communication in Statistics — Simulation 17 (2) (1988) 581.
[8] A. Höskuldsson, Prediction Methods in Science and Technology, Thor Publishing, 1996.
[9] J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.
[10] H. Martens, T. Naes, Multivariate Calibration, Wiley, Chichester, 1989.
[11] S. Wold, M. Sjöström, L. Eriksson, Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130.
[12] M. Barker, W. Rayens, Journal of Chemometrics 17 (2003) 166–173.
[13] S. Wold, M. Sjöström, SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy, 1977.
[14] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, Upper Sadle River, NJ, 2002.
[15] F. Van der Heijden, R.P.W. Duin, D. De Ridder, D.M.J. Tax, Classification, Parameter Estimation and State Estimation, Wiley, Chichester, 2004.
[16] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barabási, Science 297 (2002) 1551–1555.
[17] A. Clauset, C. Moore, M.E.J. Newman, Nature 453 (2008) 98–101.
[18] R. Guimerà, L.A.N. Amaral, Nature 433 (24) (2005) 895–900.
[19] J.A.S. Newman, Proceedings of the National Academy of Sciences of the United States of America 103 (23) (2006) 8577–8582.
[20] A.M. Yip, S. Horvath, BMC Bioinformatics 8 (22) (2007) 1–14.
[21] D.C. Montegomery, E.A. Peck, G.G. Vining, Introduction to Linear Regression Analysis, Wiley, Hoboken, NJ, 2006.
[22] M.S. Reis, P.M. Saraiva, Industrial and Engineering Chemistry Research 49 (5) (2010) 2493–2502.
[23] D. Angélico, M.S. Reis, R. Costa, P.M. Saraiva, J. Ataíde, Tecnicelpa — XIX Encontro Nacional 2005, Proceedings of the Conference, Tomar (Portugal), 2005. 165–173 , (Vol. City, Year).
[24] T.R. Golub, D.K. Slonim, P. Tamayo, et al., Science 286 (1999) 531–537.
[25] A.J. Myles, R.N. Feudale, Y. Liu, N.A. Woody, S.D. Brown, Journal of Chemometrics 18 (2004) 275–285.
[26] A.C. Pereira, M.S. Reis, P.M. Saraiva, J.C. Marques, Chemometrics and Intelligent Laboratory Systems 105 (1) (2011) 43–55.
[27] R.G. Brereton, Chemometrics — Data Analysis for the Laboratory and Chemical Plant, Wiley, Chichester, 2003.