Input Selection and Regression using the SOM

Francesco Corona[1,2] and **Amaury Lendasse**[1,*]
[1]Laboratory of Computer and Information Science - Helsinki University of Technology
P.O. Box 5400, FIN-02015 HUT Espoo, Finland
[2]Dipartimento di Ingegneria Chimica e Materiali - Università di Cagliari
Piazza d'Armi 1, I-9123 Cagliari, Italy
[*]**lendasse@hut.fi**

Abstract - *This paper presents a global methodology to build a nonlinear regression when the number of available samples is small compared to the number of inputs. The task is divided in two parts: selection of the best inputs and construction of the approximator. A first SOM is used to compute clean correlations between the inputs and the output. A second SOM is built to link the output to the selected inputs. The good performances of this methodology are illustrated on a spectrometric dataset.*

Key words - Spectrometry, Regression, Variable Selection, Self-Organizing Maps

1 Introduction

The problem of regression is common to many fields of engineering, finance, etc. A set of measurements is available, and one wants to investigate the relationships between them. In the general context of regression from sparse data, we have observations $(\mathbf{x}^i, y_i), i = 1, ..., N$, where $\mathbf{x}^i = (x_{i1}, ..., x_{in})^\top$ and y_i are the input and output variables for the $i-th$ observation, respectively. The task for the data analyst is represented by obtaining a correct representation of the underlying functional relationship $y = f(\mathbf{x})$.
However, a characteristic of the data used in regression problems is represented by redundancy. This means that it is not necessary true that all available input variables are relevant to the output variable to be estimated. Moreover, some inputs may be dependent on other ones: e.g., they may be collinear. In addition, it is not unusual to analyse databases consisting of a number of inputs that is comparable to the number of available observations. Operating in such conditions may led to ill-conditioned solutions and overfitting may also appear. Furthermore, these conditions might bring up the known problem referred as to the curse of dimensionality.
Typical examples can be found in spectrometric problems where the aim is to estimate the content of some chemical component (the output variable) starting from its measured spectra (the input variables). Usually, a large number of input spectral variables is measured (hundreds, up to thousands) and only several dozen of samples are available.
In such a situation, it is necessary to select among all possible candidates only the inputs that truly contribute to a correct representation of the output. More formally, being $\mathbf{x} \in \mathbb{R}^n$ the original set of input variables, the task is to find the subset $\mathbf{x}' \in \mathbb{R}^s$, where $s \ll n$, that

builds the best regression model according to some predefined criteria [4].
In this paper, an application of the self-organizing map (SOM) to spectrophotometric modelling is presented. The application refers to the problem to determining the fat content of meat samples from its near-infrared transmittance spectroscopy. The SOM paradigm is used as an efficient framework to accomplish both the task of variable selection and nonlinear regression. As for the selection of the variables, the problem is approached by evaluating the relevance of the inputs to the outputs using correlation coefficients. The correlations are calculated from the model vectors of a SOM trained with all the input variables and the variable to be estimated. The information extracted from this preliminary SOM is then employed to develop a nonlinear regression model using a new SOM. This second SOM is built in the space of the selected inputs and the output variable.
The paper is organized as follows. In Section 2, the general methodology for variable selection and nonlinear regression using the SOM is illustrated. Section 3 presents the spectrometric application and discusses the results.

2 Methodology

As stated in the introduction, the aim of this paper is to assess the potentialities of the self-organizing map [1] for both variable selection and regression. In this section, the general methodology in reported.

2.1 Variable Selection using the SOM

In practical data analysis and modeling, one of the most common tasks is to search and find dependencies between variables. The self-organizing map can be successfully employed for getting a visual insight of the data and to start the preliminary investigation of potential correlations. From the SOM, dependencies can be searched by looking for similar patterns in identical positions in the component planes and distance matrices visualizations of the map [9]. Despite its inherent appeal, when the dimensionality of the data is large, this qualitative approach is not practical. Moreover, the visual impression of dependency needs to be validated with more rigorous statistical methods [2].
Alternatively, a quantification of the similarities between the variables can be accomplished measuring their correlations from the components of the model vectors or their distance matrices. As proposed in [10], to benefit from the noise filtering performed by the SOM paradigm, *clean* correlations can be calculated directely from the model vectors instead of the original data as:

$$c_{j,k} = \frac{1}{\sigma_j \sigma_k} \sum_{l=1}^{M} (m_{lj} - \mu_j)(m_{lk} - \mu_k) \tag{1}$$

where j and k represent the input and output variables, respectively. With μ and σ are denoted their mean value and the standard deviation, and M is number of the model vectors $\mathbf{m}_l \in \mathbb{R}^{n+1}$ in the SOM.
In principles, the selection is then simply performed by ranking the inputs according to their relevance to the output variable, and selecting an appropriate subset $\mathbf{x}' \in \mathbb{R}^{\mathbf{s}}$. Indeed, when

the input variables are very similar (as in spectrometric problems), to avoid the selection of collinear variables additional informations from *a priori* knowldedge can be considered.

2.2 The SOM as a Nonlinear Regression Model

As for the development of a regression model [7], a set of model vectors $\mathbf{m}_j \in \mathbb{R}^d$ (where, $d = s + 1$) is trained into the selected space of the observation vectors $(\mathbf{x}', y) \in \mathbb{R}^s \times \mathbb{R}$. The estimation of y is accomplished by identifying the winner model for a set of known independent variables $\mathbf{x}'$:

$$\mathbf{m}_w = \arg\min_j \sum_{p=1}^{s} (x'_p - m_{jp})^2 \tag{2}$$

so that $\hat{y} = m_{wp}$, for $p = d$. Using the selected inputs, the accuracy of the model is parametrized by the number of model vectors. In general and given the usual restrictions on generalization, the larger is the number of model vectors the more dense is the quantization of the observations' space and, hence, the better is the estimation accuracy.

3 The Study Case

The task for the Tecator dataset [8] consists of estimating the content of fat in a meat sample starting from its light absorbance spectrum. The spectra are acquired by means of a Tecator Infratec Food and Feed Analyzer operating in the $850 - 1050nm$ wavelength range. The absorbance ($-log_{10}T$, where T is the light transmittance) is measured on the basis of the Near Infrared Transimission (NIT) principle for 100 wavelengths within the mentioned range. The content in fat of the finely chopped meat samples is evaluated in laboratory tests by analytic chemistry.

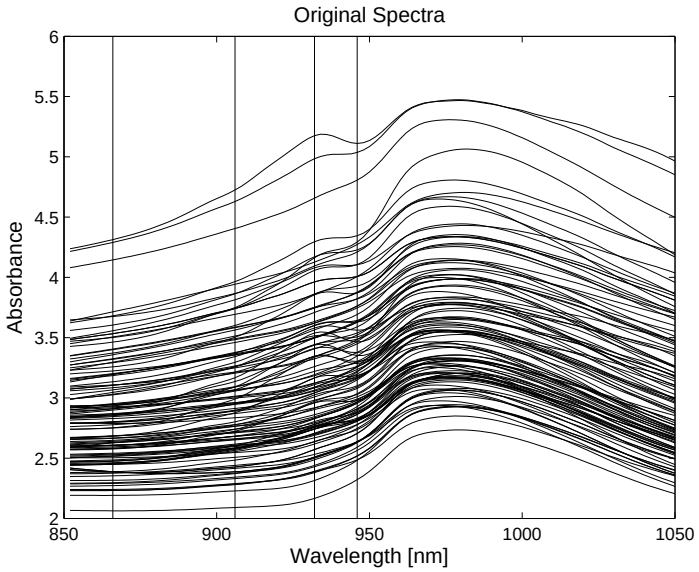

Figure 1: A selection of near-infrared spectra

The dataset is composed of 172 training observations (divided in 129 learning samples, and 43 samples for validation) and 43 observations for testing the final model. For the sake of clarity, in the following the learning set is denoted with L, the validation set with V and the testing set with T. Each observation consists of the 100−channel spectrum of absorbances and the fat content: that is, $\mathbf{x} \in \mathbb{R}^n$ (with, $n = 100$) and $y \in \mathbb{R}$.

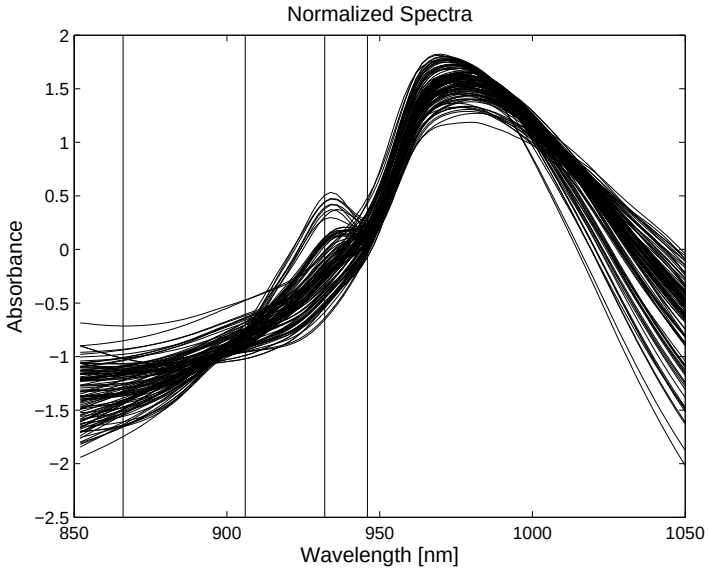

Figure 2: A selection of normalized spectra

Previous studies [5, 6] on this standard dataset, pointed out the relevance of the shape of the spectrum. Indeed, a remarkable property of the Tecator dataset is represented by the criticality to extract interesting structural differences between nearly identical spectra. In spectrometry, particularly in near-infrared, this problem is approached with a widely used technique, referred as to derivative spectrometry, that employes differentiation as a general way to discriminate against broad spectral features in favor of narrow ones. Therefore, to implicitely embed informations on, at least, the first derivative of the spectrum, the original observations were preprocessed so that each spectrum is normalized to zero mean and unit variance:

$$\mathbf{x}^i := \frac{(\mathbf{x}^i - \bar{\mathbf{x}}^i)}{\sqrt{var(\mathbf{x}^i)}}, \forall i \in [1, N] \tag{3}$$

A selection of spectra from the available database is illustrated in terms of both the original (Figure 1) and the normalized variables (Figure 2). In both figures, the vertical lines correspond to the selected variables, as described in the following.

According to the methodology presented in Subsection 2.1, the *clean* correlation between each pair (x, y) of inputs and the output was evaluated using a first bi-dimensional SOM with 9×7 units. The size of this preliminary SOM was evaluated according to the heuristics discussed in [11]. On the basis of the results presented in Figure 3, the selection of the inputs was performed considering the spectral variable that most correlate with the output (the local maxima). In addition, to characterize the complete band of wavelengths, also the

variables corresponding to the local minima were selected. This is equivalent to the following assumption: as the linear dependence decreases, possible nonlinear dependencies might arise. On the basis of these considerations 7 inputs were selected. In order to define a sparser model and validate the assumptions, an exhaustive search for the best combination is performed in the space of the selected inputs: the number of possible combinations is obviously 2^7. For each possible combination of inputs a SOM-based regression model is built (see, Subsection 2.2). The selection of the best model is defined using the normalized mean square error on the validation set ($NMSE_V$) as accuracy criterion:

$$NMSE_V = \frac{1/N_V \sum_{i=1}^{N_V} (\hat{y}_i - y_i)^2}{1/(N_L + N_V + N_T) \sum_{i=1}^{N_L+N_V+N_T} (y_i - \bar{y})^2} = \frac{1/N_V \sum_{i=1}^{N_V} (\hat{y}_i - y_i)^2}{var(y)} \tag{4}$$

where N_L, N_V, N_T are the number of observations in the L, V and T set, respectively. The observed variance $var(y)$ of the output y estimated from all available observations is used as a normalization term common to all the datasets.

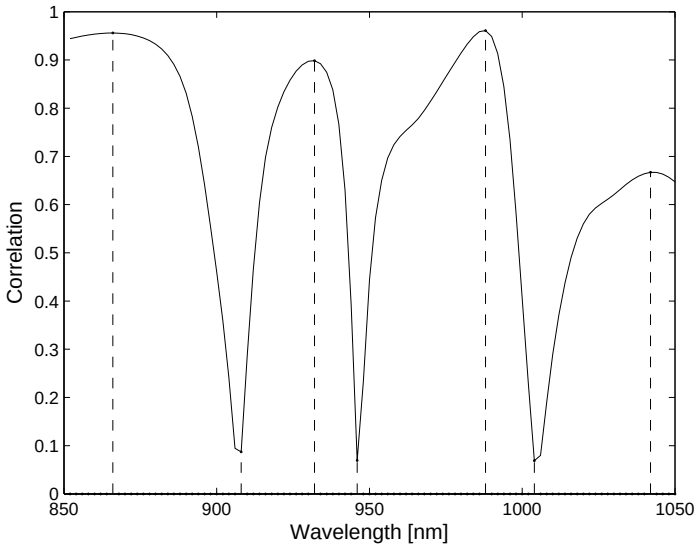

Figure 3: Correlation with the fat content as a function of the wavelength

The best set of inputs is represented by the first 4 spectral variables corresponding to 866, 906, 932 and 946 *nm*. Two of them correspond to local maxima of the *clean* correlations and two of them are local minima. Note that the most correlated variables do not necessary belong to the optimal set.

As for the final regression model, after calculating the eigenvalues of the complete observation matrix $(\mathbf{x}', y) \in \mathbb{R}^m \times \mathbb{R}$, the model vectors of a $2D$ map were initialized along the 2 greatest eigenvalues of the covariance matrix of the given data. The regression of the models into the input space was then performed according to a batch learning algorithm using euclidean metrics and gaussian neighborhood kernel functions.

To optimize the choice of the number of model vectors, the performances of different configurations of the map were compared using the accuracy measure defined in Equation 4.

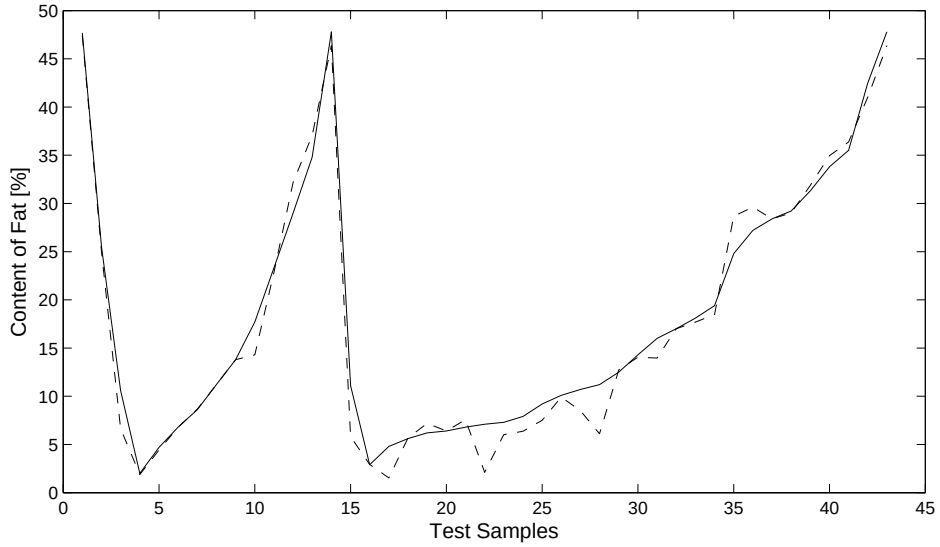

Figure 4: Simulation results: Solid line (Measured values) and Dashed line (Estimated values)

In Figure 4, the results obtained with the 4 selected variables are presented for the test dataset. The corresponding $NMSE_T$ is equal to 0.016. The achived accuracy represents an improvement when compared to a previous work [6] that used mutual information to select the inputs a number of traditional regression models.

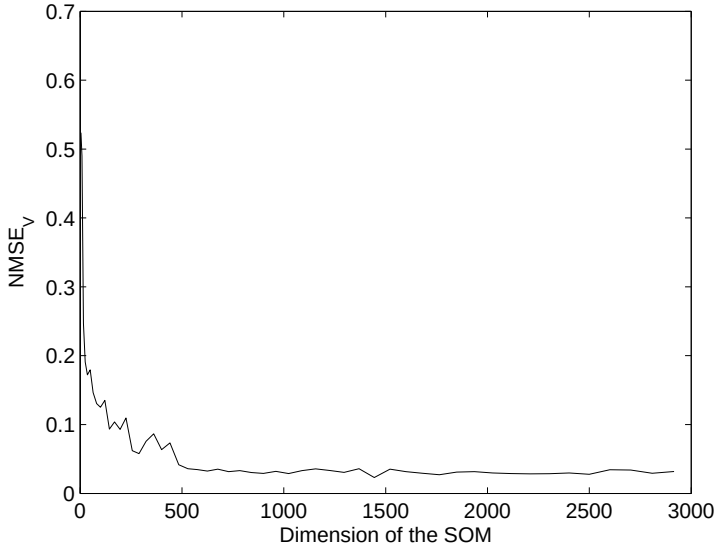

Figure 5: Evolution of the performances with the dimensionality of the SOM

The optimal number of model vectors was found to be 1156 displaced in a 34×34 grid, see Figure 5. Since the validation error does not present an evident minimum (no apparent overfitting), we have chosen an optimal number of units that corresponded to a decrease of the $NMSE_V$ that was smaller than the 1% when compared to the one obtained for the immediately smaller map.
Even if, the number codebooks is clearly larger than the number of samples available for calibration by one order of magnitude, the model appears not of suffer of overfitting and is still able to generalize the acquired knowledge.
This unexpected result was interpreted as an indicator of the relevance of the adaptation performed by the SOM when employed in function approximation. From an accurate analysis of the model, it appeared that also those model vectors that do not have any data in the corresponding Voronoi zone ("lost", with respect to the training process) actively participate to improve the estimation accuracy with the validation/test data. Similar results could not be achieved using simple vector quantization (see, [3]).
This nice property of the SOM is illustrated using a simple example. Using the methodology described in Subsection 2.2, a monodimensional SOM is trained into a 2D space (one input, and one output) in order to approximate the function $y = x^2$. For a new data x^*, the winner model is identified and the estimation of y^* is calculated. From Figure 6, it is clear how the interpolating model contributes to the quality of the approximation.

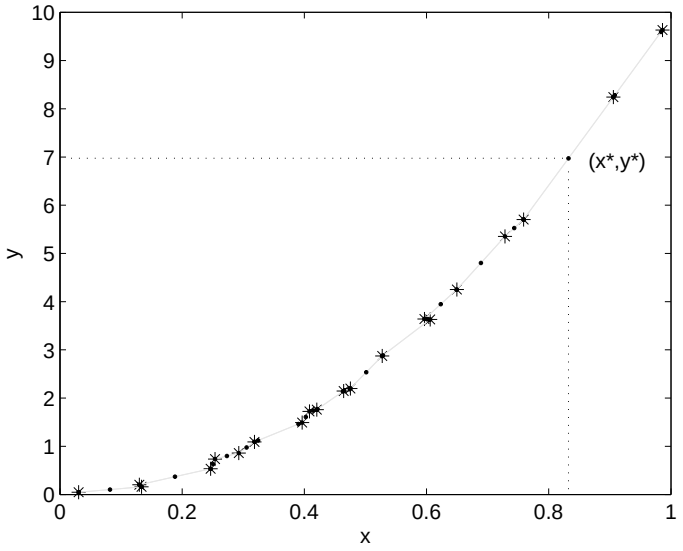

Figure 6: Monodimensional function approximation with the SOM: Data ($\star$) and the SOM(-$\bullet$-)

4 Conclusions and further works

In this paper, an original methodology that combines input selection and regression using the self-organizing map is presented. From the discussed results two major considerations can be drawn.

Firstly, the SOM can be effectively used in the selection of the relevant inputs for a regression model. Secondly, the accuracy of the results demonstrate the potentialities of the self-organization in the regression context.
For the described application, the sparsity of the obtained models and the good quality of the predictions is, indeed, an advantage because of the interpretability of the results.
The methodology will be further investigated and validated. It is our goal to assess its potentiality with other problems of interest: e.g., in pharmacology and chemometrics, where the large number of variables and the reduced number of samples represent a criticality.

References

[1] T. Kohonen (1995), *Self-Organizing Maps*, Berlin, Heidelberg, Springer.

[2] J. Lampinen and T. Kostiainen (2000), Self-Organizating Map in Data-Analysis - Notes on Overfitting and Overinterpretation, in *Proc. ESANN'00 European Symposium on Artificial Neural Networks*, p. 239-244.

[3] A. Lendasse, D. François, V. Wertz, and M. Verleysen (2005), Vector Quantization: a Weighted Version for Time-Series Forecasting, to appear in *Future Generation Computer Systems*.

[4] A. J. Miller (1990), *Subset selection in Regression*, London, Chapman & Hall.

[5] F. Rossi, N. Delanny, B. Conan-Guez and M. Verleysen (2005), Representation of Fuctional Data in Neural Networks, *Neurocomputing*, **vol. 64C** p. 183-210.

[6] F. Rossi, A. Lendasse, D. François, V. Wertz and M. Verleysen, Mutual Information for the Selection of Relevant Variables in Spectrometric Nonlinear Modelling, to appear in *Chemometrics and Intelligent Laboratory Systems*.

[7] O. Simula, J. Vesanto, E. Alhoniemi, and J. Hollmèn (1999), Analysis and Modeling of Complex Systems using the Self-Organizing Map, Chapter in *Neuro-Fuzzy Techniques for Intelligent Information Systems*, Physica Verlag. Springer Verlag.

[8] Tecator Meat Sample Dataset, *http://lib.stat.cmu.edu/datasets/tecator*, Statlib.

[9] J. Vesanto (1999), SOM-based Data Visualizations Methods, *Intelligent Data Analysis*, **vol. 3(2)** p. 111-126.

[10] J. Vesanto and J. Ahola(1999), Hunting for Correlations in Data using the Self-Organizing Map, In *Proc. CIMA'99 Computational Intelligence Methods and Applications*, p. 279-285.

[11] J. Vesanto, J. Himberg, E. Alhoniemi and J, Parhankangas (2000), SOM Toolbox for Matlab5, *http://www.cis.hut.fi/projects/somtoolbox*.