# Measures of Topological Relevance based on the Self-Organizing Map: Applications to Process Monitoring from Spectroscopic Measurements

Francesco Corona*
Helsinki University of Technology
Lab. of Computer and Information Science
P.O. Box 5400 - FI-02015 HUT, Finland
fcorona@cis.hut.fi

Elia Liitiäinen
Helsinki University of Technology
Lab. of Computer and Information Science
P.O. Box 5400 - FI-02015 HUT, Finland
elia@cis.hut.fi

Amaury Lendasse
Helsinki University of Technology
Lab. of Computer and Information Science
P.O. Box 5400 - FI-02015 HUT, Finland
lendasse@cis.hut.fi

Roberto Baratti
University of Cagliari
Dept. of Chemical Engineering and Materials
P.zza D'Armi 1 - I-09123 CAGLIARI, Italy
baratti@dicm.unica.it

## Abstract

*In this work, the problem of real-time monitoring of products' properties from spectrophotoscopic measurements is presented. Light absorbance spectra are used as inputs to soft sensors that estimate outputs otherwise difficult to measure on-line. To overcome the issues associated to calibrating the estimation models from very high-dimensional inputs and a reduced number of observations, we propose to select only a subset of relevant inputs emerging from the topological structure of the data. The topologically preserving representation is performed using the Self-Organizing Map (SOM) and the relevance measured from the U-matrices. Being based on a selection of original spectral variables, the resulting models retain the chemical interpretability of the underlying system. Moreover, the approach is independent on the regression model to be embedded in the soft sensors. In this paper, the utility of the Measures of Topological Relevance (MTR) over the SOM is discussed on two full-scale problems from refining and pharmaceutical industry.*

## 1. Introduction

Real-time monitoring has become an essential component of modern process industry for optimizing the production toward high-quality products while reducing operating costs. The tools of on-line analytical chemistry and chemometrics fulfill the necessary requirements for real-time analysis of key chemical and physical properties for a broad variety of materials. This paper focuses on monitoring products' properties from non-invasive and non-destructive measurements obtained by light spectroscopy analysis.

The principle underlying process monitoring from infrared (IR), near- and medium-infrared (NIR and MIR) spectroscopic measurements is the existence of a relationship between the light absorbance spectrum of a given product and the property of interest. In fact, the spectrum is conditioned by the composition of the product and, in turns, the composition determines the property of interest. This relationship is rarely known *a priori* and it is usually reconstructed by calibrating specific data-derived models, without an explicit regard to first-principle criteria. The resulting spectrophotoscopic models are used to generate interesting insights on the underlying chemistry. Moreover, the wide availability of continuous-flow spectrophotometers makes the modeling approach suitable for the design of soft sensing devices that monitor the key properties of the products starting from the measured spectra [32].

However, the problem of estimating the property (the output) is defined from very high-dimensional and intrinsically redundant inputs (the spectrum). Redundancy is observed as the inherent collinearity existing between the spectral inputs. Furthermore, it is not unusual to calibrate models on a number of observations (the product's samples) that is radically smaller than the number of input candidates. To address these problems, two approaches are commonly used. One standard solution is to rely on full-spectrum

*On leave from the Department of Chemical Engineering and Materials at the University of Cagliari, Italy.

methods for dimension reduction coupled with regression: Principal Components Regression (PCR) and Partial Least-Squares Regression (PLSR) are reference models [11]. The natural refinement of such an approach is to perform a preliminary selection of relevant spectral ranges [20]. However, PCR and PLSR models are intrinsically limited by their linear structure and, because based on combinations of the original variables, are not trivial to interpret. When "kernelized" [25] or other nonlinear [21, 3] generalizations of methods are considered, the insight can be further reduced [2]. Analogous considerations apply to the functional extensions of the methods [22, 9]. The alternative solution consists of selecting, among all spectral candidates, only those inputs that truly contribute to a correct estimation of the output and, that are as much as possible not collinear. Thus, variable selection is understood as the limit extension of range selection where the chemical interpretability of the system is explicitly retained. Some recent advances in spectroscopic modeling are based on such an idea. In the absence of a chemical model, the approach is either based on model properties [1] or on relevance indexes [24]. In either cases, however, the computational burden associated to variable selection can be demanding and the approach unpractical because of the large number of candidates.

In this study, variable selection is approached by exploiting the metric structure of the spectral data, leading to a method that identifies only the spectral inputs with a topology that best matches the output's. The topology preserving modeling of the data is carried out with the Self-Organizing Map (SOM) over which the Measures of Topological Relevance (MTR) between the inputs and the output are estimated from the Unified-distance matrices (U-matrices). Because designed on the original spectral inputs, the resulting models retain a useful understandability of the underlying chemical system. Being the selection performed before building the regression models (i.e., according to a filtering approach [13]), the method is also model-independent; in the sense that, once the input variables are selected, any estimation technique can be used to reconstruct their relationship with the output to be estimated.

The presentation is organized as follows. Section 2 introduces the monitoring problem and briefly overviews the suggested approach to variables selection using the MTR over the SOM. In Section 3, the applications to two real-world problems in process monitoring from the refining and pharmaceutical industry are presented and discussed.

## 2. Methods and algorithms

The problem of monitoring product properties from light absorbance spectra can be reformulated within the context of variable selection and associated function estimation. That is, given observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ - where $\mathbf{x}_i = [x_{i1}, \ldots, x_{id}]^\top$ and $y_i$ are the inputs (on-line spectrum) and output (off-line analysis) variables for the $i-$th observation, respectively - the task consists of modeling the underlying functionality $y = f(\mathbf{x})$ that is assumed to exist between the observations. Because of the very high dimensionality $d$ of $\mathbf{x}$ (several hundreds, up to thousands) and the small number $N$ of observations (several tens, up to few hundreds), it is appropriate to operate in a reduced data space whose dimensionality is circumscribed by the intrinsic complexity of the system. Formally, being $\mathbf{x} \in \mathbb{R}^d$ the given set of candidate input variables, it is necessary to select a subset $\check{\mathbf{x}} \in \mathbb{R}^s$, with $s \ll d$, that builds the best model for $f$, according to some predefined criterion [19].

Here, a three-stage methodology stemming from [4, 5] is adopted. The methodology summarizes as follows:

1. the first stage models the input and output observations onto a Self-Organizing Map where the topological structure of the data is preserved;

2. the second stage investigates how the output's topology is related to the topology of the input;

3. only the inputs with a topology that best matches the topology of the output are selected as relevant.

Once the subset $\check{\mathbf{x}}$ of inputs is selected, any regression model can be used to reconstruct $f$ and predict the output $y$. The technique preferred in our applications is the Least-Squares formulation of the Support-Vector Machine (LS-SVM, [26]). For completeness, we also considered classical linear models for Ordinary Least Squares (OLS) and Ridge regression [14]. The meta-parameters of the Ridge and LS-SVM regression model are validated with resampling methods that estimate the prediction accuracy; the Leave-One-Out Cross-Validation (LOO-CV) is here adopted [14].

### 2.1. Topology preserving mappings with the SOM

The Self-Organizing Map, SOM [17], is an adaptive algorithm to formulate the vector-quantization paradigm [12]. In the following, the basic formulation and essential properties of the SOM algorithm are briefly reported.

The SOM consists of a low-dimensional (typically, 2D) regular array of $K$ nodes where a prototype vector $\mathbf{m}_k \in \mathbb{R}^p$ is associated with every node $k$. Each prototype acts as an adaptive model vector for the observations $\mathbf{v}_i \in \mathbb{R}^p$. In the addressed context of spectroscopy, both the inputs and the output are considered; i.e., $\mathbf{v}_i = [\mathbf{x}_i; y_i]$ and $p = d + 1$. During the computation of the map, the observations are mapped onto the SOM's array and the prototyping model vectors adapted according to the learning rule:

$$\mathbf{m}_k(t+1) = \mathbf{m}_k(t) + \alpha(t) h_{k,c(\mathbf{v}_i)} \Big( \mathbf{m}_k(t) - \mathbf{v}_i(t) \Big), \quad (1)$$

where $t$ is the discrete-time coordinate of the mapping steps, and $\alpha(t) \in (0,1)$ the monotonically decreasing learning rate. The scalar multiplier $h_{k,c(\mathbf{v}_i)}$ denotes a neighborhood kernel function centered at the Best Matching Unit (BMU), the model vector $\mathbf{m}_c$ that best matches with the observation vector $\mathbf{v}_i$. The matching is determined according to a competitive criterion based on the Euclidean metric $||\cdot||$ and, at each step $t$, the BMU $\mathbf{m}_c(t)$ is the prototype $\mathbf{m}_k(t)$ that is the closest to the observation $\mathbf{v}_i(t)$:

$$||\mathbf{m}_c(t) - \mathbf{v}_i(t)|| \leq ||\mathbf{m}_k(t) - \mathbf{v}_i(t)||, \quad \forall k = 1, \ldots, K. \tag{2}$$

The neighborhood kernel $h_{k,c(\mathbf{v}_i)}$ centered at $\mathbf{m}_c(t)$ is usually chosen in the Gaussian form:

$$h_{k,c(\mathbf{v_i})} = \exp\Big(-\frac{||\mathbf{r}_k - \mathbf{r}_c||^2}{2\sigma^2(t)}\Big), \tag{3}$$

where the vectors $\mathbf{r}_k$ and $\mathbf{r}_c$ (in $\mathbb{R}^2$, for a 2D map) represent the geometric location of the nodes on the array, and $\sigma(t)$ denotes the monotonically decreasing width of the kernel that allows for a regular smoothing of the prototypes. On the array, the effect of the kernel decreases with the distance between the BMU and the other prototypes.

The map is computed recursively for each observation. As $\alpha(t)h_{k,c(\mathbf{v}_i)}$ tends to zero with $t$, the set of prototype model vectors $\{\mathbf{m}_k\}_{k=1}^K$ is updated to represent similar observations in $\{\mathbf{v}_i\}_{i=1}^N$ and the prototypes converge toward their asymptotic limits [23, 8]. The resulting model vectors form a submanifold in the original data space where the relevant topological and metric properties of the observations are preserved. Thus, the SOM is to be understood as an ordered image of the original high-dimensional data manifold modelled with a low-dimensional array of prototypes. On the SOM's array, the complex nonlinear structures existing between the data are represented with simple geometric relationships.

**The MTR based on the U-matrix of the SOM**

The Self-Organizing Map is widely employed to getting a visual insight of the data and to starting a preliminary investigation of potential relationships between the component variables. From the SOM, dependencies can be either searched by looking for similar patterns in identical positions in component plane and distance-based representations of the map [28] or estimating the correlation coefficients between such displays, as proposed in [29].

We propose to identify the relevant inputs by exploiting the topology preserving properties of the SOM of the input and output data according to a relevance measure derived from the assumed continuity of the unknown functionality $y = f(\mathbf{x})$. Under this hypothesis, if two points $\mathbf{x}_i$ and $\mathbf{x}'_i$ are close together in the input space, it is expectable that $f(\mathbf{x}_i)$ and $f(\mathbf{x}'_i)$ are also close together in the output space. Therefore, the continuity of $f$ is also represented in the local topology of the data and, thus, recoverable from nearest neighbors graphs. If the neighborhood continuity is not satisfied (i.e., the points $y_i$ and $y'_i$ are not close together in the output space) it can be either due to the presence of noise or because the inputs are not related to the output. In order to benefit from the noise-filtering properties of the SOM, this general principle can be directly explored from the set of model vectors $\{\mathbf{m}_l\}_{l=1}^M$ and the U-matrix of the SOM, as proposed in [5].

The standard approach to recover the topological structure of the data from the SOM is to compute the Unified-distance matrix, or U-matrix [27]. The U-matrix $\mathbf{U}$ is built from local distances for each SOM node and, thus, defines a nearest neighbor graph based on the model vectors of the map. To represent the local topology of the component variables, the corresponding U-matrices are calculated independently along each direction of the data space; that is, $\mathbf{U}_{x_j}$ (with $j = 1, \ldots, d$) for the input variables, and $\mathbf{U}_y$ for the output. The relevance of the input $x_j$ to the output $y$ is calculated from the distance between the topologies, that is:

$$\mathcal{D}(x_j, y) = ||\mathbf{U}_{x_j} - \mathbf{U}_y||_F, \tag{4}$$

where the matrix Frobenius metric $||\cdot||_F$ measures the closeness between the U-matrices; the closer to 0 is the measure, the more relevant is the input for reconstructing the output. In order to clearly represent relevance, the measure $\mathcal{D}(x_j, y) \geq 0$ is preferably inverted and rescaled so that, larger values indicate stronger relevances.

In principles, variable selection is simply performed by ranking the inputs according to their relevance to the output, and selecting a reduced but still representative subset $\check{\mathbf{x}} \in \mathbb{R}^s$. However, this basic selection procedure applied to spectroscopy data is intrinsically limited by the continuous nature of the light's wavelengths domain, regardless the employed relevance index as long as it is continuous. In fact, it is intuitive that absorbances measured at neighboring wavelengths are characterized by a relevance to the output that is very similar. Therefore, the selection of an input $x_j$ that is found to be relevant to predicting $y$ is naturally accompanied by the selection of a broad range of contiguous inputs also characterized by high relevance, but redundant because embedding a near-identical informative content.

## 2.2. An input selection strategy for spectroscopy

In such context, the selection scheme proposed in [4] can be easily adapted to the topological measures of relevance defined in Equation 4. The procedure was originally defined for a standard measure of dependence, the Pearson's

Correlation Coefficient (CC):

$$\mathcal{R}(x_j, y) = \frac{E[x_j y] - E[x_j]E[y]}{\sqrt{E[x_j^2] - E[x_j]^2}\sqrt{E[y^2] - E[y]^2}}, \quad (5)$$

where, in practice, the expectations are approximated based on a finite number of observations. However, the CC is only able to capture dependencies that manifest themselves in the covariance. This motivates the use of alternative measures of relevance. In the case of MTR over the SOM, the selection procedure summarizes as:

1. calculate the full set $\mathcal{D} = \{\mathcal{D}(x_j, y)\}_{j=1}^{d}$ of pairwise relevances between each input-output pair;
2. select the subset of inputs $\check{\mathbf{x}}$ with a topology that best matches the output's: i.e.,

$$\check{\mathbf{x}} = \{\check{x}_{j^*} \subset \mathbf{x} : j^* = \underset{j}{\text{argmax}}\ \mathcal{D}(x_j, y)\}_{j^*=1}^{s}.$$

The procedure identifies only the inputs that are associated to the local maxima of $\mathcal{D}$, thus, relevant to predict the output. In that sense, the selection is optimal with respect to the problem of predicting the output: in fact, among similar inputs, only the maximally relevant ones are retained and the neighboring redundancies are discarded. Being relevance to the output the only supervising criterion for selection, the procedure is still suboptimal with respect to problem of selecting inputs that are also minimally redundant. Nevertheless, the selected variables are implicitely as much as possible dissimilar, because each prototypes different subsets of inputs separated by the local minima of $\mathcal{D}$.

Because the selection scheme is general and valid for any measure of relevance, as long as it is defines a continuous function in the operating domain of wavelengths of the spectrophotometer, in this study we also considered other measures: namely, i) Mutual Information (MI, [6]), and; ii) Noise Variance Estimates (NVE, [10]). For the sake of comparison, also CC, MI and NVE results are included in the experiments performed on the case studies in Section 3.

Mutual information measures the distance between the joint density $p(x_j, y)$ and the product density $p(x_j)p(y)$ in the sense of Kullback-Leibler divergence. The analytic form of the MI is given by:

$$I(x_j, y) = \int p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)} dx_j dy. \quad (6)$$

It can be shown that $I(x_j, y) \geq 0$ and $I(x_j, y) = 0$ if and only if the variables $x_j$ and $y$ are independent. The integral can be viewed as a measure of distance between the actual joint distribution and the joint distribution under the assumption of independence of the variables. To estimate MI we used the estimator introduced in [18].

Noise Variance Estimation is a technique that, under the assumption that there is a functional relationship between $x_j$ and $y$, estimates the part of the output that cannot be modelled with the given inputs (i.e., the noise). As such noise variance estimates can be also understood as the best possible Mean Squared Error (MSE) obtainable by any model. This task can be done in various ways of which we chose the well-known estimator proposed by Gasser [10].

## 3. Applications

The development and the application of the studied soft-sensors is illustrated with two actual monitoring tasks from the refining and the pharmaceutical industry. The selected applications are referenced full-scale problems for variable selection and interpretation, as well as prediction purposes.

### 3.1. Study case 1: Predicting the octane number in finished gasolines

The first application consists of estimating the octane number in gasolines. The American Society for Testing and Materials (ASTM) standard for obtaining such a property is based on an internal combustion engine in which the octane number is measured [15]. The procedure is time consuming, involves expensive and maintenance-intensive equipment and requires skilled labor and, therefore, is not well suited for on-line monitoring. Nevertheless, real-time measurements of such a property are of fundamental importance for both the production and the blending process of finished gasolines. The application of the methodology is discussed on a set of spectral measurements and associated evaluations of the octane number provided by Camo A/S (Trondheim, Norway), which is gratefully acknowledged.

The absorbance spectra are acquired by means of a spectrophotometer operating in the $1100-1550$*nm* wavelengths' range, in Figure 1(a). The absorbance is measured on the basis of the NIR transmission principle with a $2$*nm* resolution. The measurements of the octane number (in the $86 - 92$ range) are evaluated in laboratory by the reference ASTM motor tests. Therefore, each sample consists of the $226-$channel spectrum of absorbances and the corresponding octane number; that is, $\mathbf{x} \in \mathbb{R}^d$ with $d = 226$, and $y \in \mathbb{R}$. The dataset consists of $24$ observations for model calibration and validation and $9$ observations for testing the final model. The data were preliminary preprocessed by removing the outliers and mean-centering. Although in reduced amount, the data are collected over a sufficient period of time considered to span all the important variations in the production of the finished product. Being the relationship between the octane and the spectrum distributed among different inputs, the application is also interesting because

variable selection cannot be easily performed through first-principle interpretation of the spectra [31, 30].

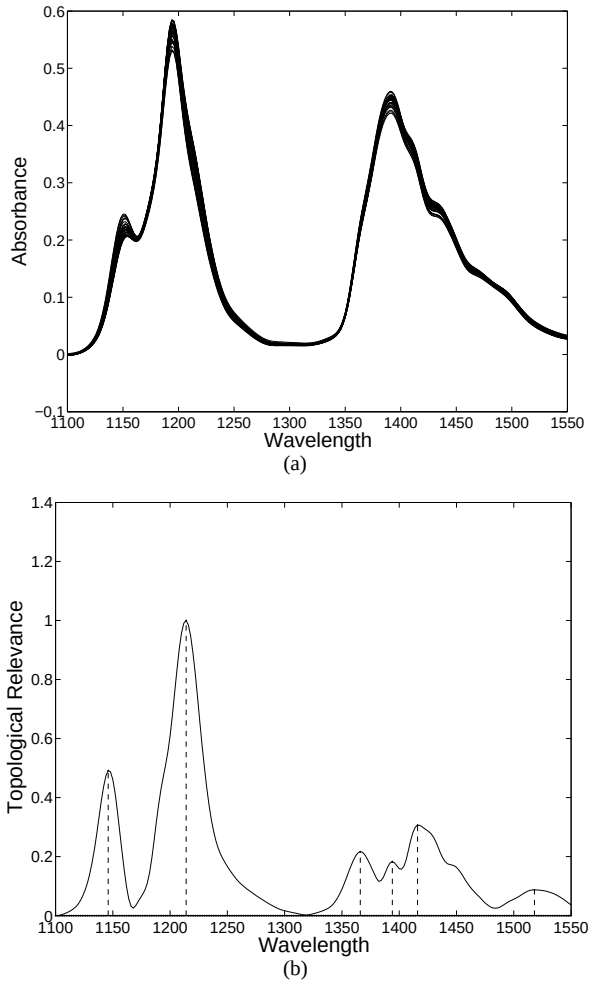


**Figure 1. Case Study 1: The spectral observations (a) and the Measure of Topological Relevance (MTR) on the Self-Organizing Map (SOM) between the inputs and the output (b).**

### Variable selection and chemical interpretability

According to the method discussed in Section 2, the 2D SOM of the input and output observations in the calibration set was computed. The map consists of a hexagonal array of nodes initialized in the space spanned by the eigenvectors corresponding to the 2 largest eigenvalues of the covariance matrix of the data. As usual, the ratio between these eigenvalues was also used to calculate the size ($5 \times 5$ nodes) of the SOM. On the map, the set of topological relevances $\mathcal{D} = \{\mathcal{D}(x_j, y)\}_{j=1}^{d}$ between each input-output pair was calculated and the subset $\check{\mathbf{x}} = \{\check{x}_{j^*}\}_{j^*=1}^{s}$ of relevant inputs was selected, $s = 6$. Being the 6 inputs maximally relevant, they are identified by the local maxima of $\mathcal{D}$, in Figure1(b).

The set of selected inputs (Table 1) is in agreement with the chemical model explaining the influence for the chemical groups on the octane number [16]. The analyzed spectra show the typical overlapped absorbance bands arising from different hydrocarbon functional groups and reflect the samples' composition. The major absorbance features in the experimental region are usually assigned to the 2nd overtone ($1100 - 1300$*nm*) and to the combination bands ($1300 - 1550$*nm*) of the C-H vibrations. In detail:

- the aromatic bonds at $\sim 1150$*nm* ($\check{x}_1$) are related to an increase in octane number. Conversely, the methylene bonds at $\sim 1220$*nm* ($\check{x}_2$) indicate the presence of linear hydrocarbons which are responsible for a reduction in the gasoline quality. The methyl bonds at $\sim 1200$*nm* indicate a larger amount of branched hydrocarbon although the absorbance is also influenced by the amount of linear paraffin: in fact, its effect on octane is not readily explained and the contribution, usually, varies with the gasoline type. Actually, this occurs with the present spectra in which, even if the relevance $\mathcal{D}^s$ shows an inflection at $1200$*nm*, the absorbance does not correspond to a local maximum and, thus, the associated input is not selected;

- by the same token, the effect of the combination bands for methylene ($\sim 1395/1416$*nm*), and methyl ($\sim 1360/1345$*nm*) on octane mimics what observed in the short-wavelength range. With this respect, the methylene absorbance wavelengths are correctly identified ($\check{x}_4$ and $\check{x}_5$), while $\check{x}_3$ accounts for the 1st methyl band. As already noticed above, again the 2nd methyl band is only partially recovered by an inflection in $\mathcal{D}^s$.

As for variable $\check{x}_6$, no spectral features are readily assignable. Its selection can be ascribed to baseline effects.

| | $\check{x}_1$ | $\check{x}_2$ | $\check{x}_3$ | $\check{x}_4$ | $\check{x}_5$ | $\check{x}_6$ |
|---|---|---|---|---|---|---|
| [nm] | 1146 | 1214 | 1366 | 1394 | 1416 | 1518 |

**Table 1. Case Study 1: The set of selected inputs and associated wavelengths.**

In Figure 2, the results obtained with the absolute Pearson's Correlation Coefficients (Figure 2(a)), Mutual Information (Figure 2(b)) and Gasser's Noise Variance estimates are presented (Figure 2(c)). Notice that, in the case of NVE, the local minima reflect the highest relevance. Based on the depicted results, all the measures are capable of identifying either part or all the relevant variables and their behavior

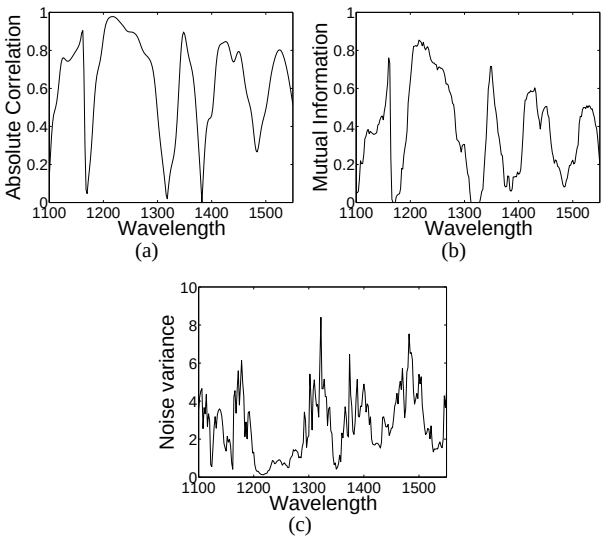


**Figure 2. Case Study 1: Other input-output measures of dependence - Correlation Coefficient (a) and Mutual Information (b) - and the Gasser's noise variance estimate (c).**

resembles the relevance estimated by the MTR. Nevertheless, only the CC is able to represent the smooth nature of the observations and, thus, allow a direct selection of the local maxima in the relevance function. As for the MI and NVE, such property of the data is only partially recovered preventing an automatic variable selection procedure.

#### Regression models and prediction results

Finally, both linear (OLS and Ridge Regression) and nonlinear (LS-SVM) models were calibrated to represent $f$ from the $6$ selected inputs $\check{\mathbf{x}}$. When needed, the metaparameters of the models (the penalty term in Ridge regression and the kernel width and regularization term in LS-SVM) were validated by LOO-CV. The prediction accuracy of the models was evaluated in terms of Root Mean Squared Error ($\text{RMSE}_{\mathcal{T}}$) on the independent set of testing data.

In Table 2, the prediction results are compared to the two standard calibration methods used in spectroscopy, the full-spectrum PLSR and PCR. The number of latent variables in the PLSR and PCR model were also selected by LOO-CV. From the table, it is possible to notice that all the regression models achieve accuracies that are comparable to the ASTM standard of reference. In detail, the LS-SVM gives prediction results that are analogous to the standard PLSR model, whereas the PCR model outperforms all the other methods. Interestingly, also the linear models produce accurate results confirming the quality of the selected variables. This is also demonstrated by an almost negligible value of the penalty term selected for the Ridge regression, indicating a near-absolute absence of shrinkage for the regression coefficients. In fact, the method proved capable to select only those inputs carrying important information, thus, leading to parsimonious models based on only $6$ original variables with a clear chemical understandability. Together with the high accuracy, such properties suggest an efficient implementation of the models for the on-line rating of octane in gasolines.

| | Number of Variables | $\text{RMSE}_{\mathcal{T}}$ |
|---|---|---|
| PCR | 3 (latent) | 0.21 |
| PLSR | 4 (latent) | 0.28 |
| OLS | 6 (original) | 0.34 |
| Ridge | 6 (original) | 0.31 |
| LS-SVM | 6 (original) | 0.24 |

**Table 2. Case Study 1: A comparison between prediction results.**

### 3.2. Study case 2: Predicting the composition of active substance in tablets

The second application consists of estimating the active substance content in pharmaceutical tablets. The problem is discussed in detail for Escitalopram® tablets produced by H. Lundbeck A/S (Valby, Denmark) using the measurements provided by the Spectroscopy and Chemometrics Group at the Faculty of Life Science, University of Copenhagen (Denmark) which is kindly acknowledged for sharing the data. The case is interesting because the identification of the inputs associated to the active substance can be prevented by the superposition of interfering artifacts due to the presence of the excipients and the production processes. Moreover, the Good Manufacturing Practice (GMP) requires pharmaceutical industries to perform frequent Content Uniformity (CU) controls on the finished products; a requirement that is usually fullfilled by time and solvent consuming chromatographic analysis operated by specifically trained laboratory personnel. Therefore, the production of such drug would greatly benefit from the availability of fast and reliable methods alternative to conventional tests.

Four different dosages ($5$, $10$, $15$ and $20$*mg*) of the drug are used (Table 3). The $10$, $15$ and $20$*mg* tablets have the same concentration of active substance (i.e., they are dose proportional with a nominal content equal to $8.0\%$*w/w*) and have a slot and a print on one side. The $5$*mg* tablets have a nominal content of active substance equal to $5.6\%$*w/w*. The tablets have different total weights and, therefore, also dif-

| Nominal active substance weight (*mg*) | Nominal tablet weight (*mg*) | Nominal active substance (*% w/w*) | Number of batches |
|---|---|---|---|
| 5.0 | 90 | 5.6 | 1 full-scale + 3 pilot-scale |
| 10.0 | 125 | 8.0 | 2 full-scale + 3 pilot-scale |
| 15.0 | 188 | 8.0 | 2 full-scale + 3 pilot-scale |
| 20.0 | 250 | 8.0 | 2 full-scale + 3 pilot-scale |
| 4.3-5.7 | 90 | 4.8-6.3 | 3 laboratory-scale |
| 8.3-11.4 | 125 | 6.9-9.1 | 3 laboratory-scale |
| 12.9-17.1 | 188 | 6.9-9.1 | 3 laboratory-scale |
| 17.3-22.8 | 250 | 6.9-9.1 | 3 laboratory-scale |

**Table 3. Case Study 2: Tablets specifications.**

ferent shapes and sizes. Seven full-scale production batches and twelve batches from pilot plant production are available. Furthermore, three specially prepared batches were produced to extend the calibration range to $85 - 115\%$ of the nominal content for each dosage form, giving twelve additional laboratory-scale batches. In total $31$ batches are used, each batch consisting of $10$ tablets that were individually analyzed by the spectroscopic method as well as the reference method. The pilot plant batches are film-coated, while the full- and laboratory-scale batches are uncoated. The tablets contain several excipients, the dominating one being microcrystalline cellulose and, for the coated tablets, the coating material contains titanium dioxide. In addition, it is worthwhile noticing that all the laboratory-scale tablets were stamped with a press using only one punch, whereas the pilot- and full-scale tablets are produced after a total of forty different punches.

The spectra were acquired in the $4000 - 14000 cm^{-1}$ wavenumbers' range (corresponding to the $700 - 2500 nm$ wavelengths' range) with a resolution of $16 cm^{-1}$. The measurements were recorded with an ABB Bomem FT-NIR model MB-160 performing $128$ transmittance scans per sample. The main advantage of the transmission mode, when compared to the reflectance mode, is that the resulting spectra contain also information on the inside of the tablets; thus, making the the method less sensitive to samples heterogeneity and use of coating materials. The transmittance mode is, however, more sensitive to the pressing process used in producing the tablets. The absorbances are available only for the $7400 - 10500 cm^{-1}$ interval (Figure 3) because the $4000 - 7400 cm^{-1}$ range was very noisy, while in $10500 - 14000 cm^{-1}$ very little information is present. The content of active substance in each tablet was evaluated by the reference High Performance Liquid Chromatography (HPLC) method performed in laboratory. Dyrby *et al.* in [7] provide a detailed description of the experimental setting.

Each observation consists of a $404-$channel spectrum (i.e., $\mathbf{x} \in \mathbb{R}^d$ with $d = 404$) and the content of active substance (i.e., $y \in \mathbb{R}$). The available measurements, thus, summarize to $120$ laboratory-scale observations, $120$ pilot-scale observations and $70$ full-scale observations. Our objective consists in developing an estimation model that, although calibrated using only the laboratory-scale and the pilot-scale measurements, is directly usable in monitoring the full-scale production.

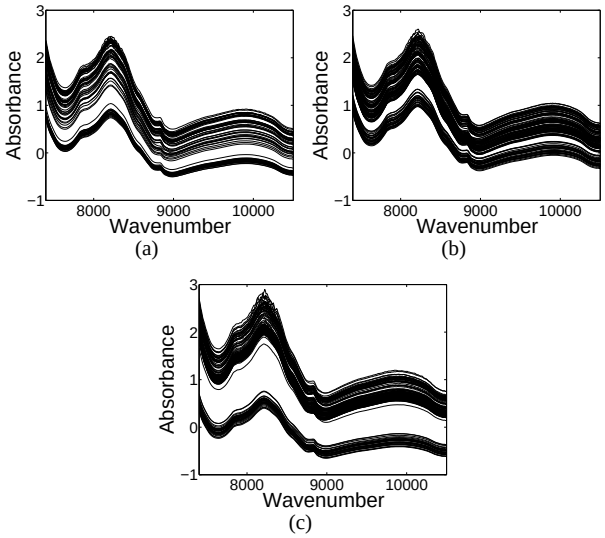


**Figure 3. Case Study 2: A selection of spectral observations from laboratory (a), pilot-scale (b) and full-scale (c) production.**

### Laboratory- and pilot-scale modeling and Full-scale predictions

For the purpose, a preliminary analysis was performed considering the three datasets independently and analyzing, for each, the inputs relevances to the corresponding output. The results are presented in Figure 4. The NIR spectrum of the active substance is highly overlapped with the excipients' in the tablets, leaving just a single working region (around $8800 cm^{-1}$) relatively free of interference, see Figure 3.

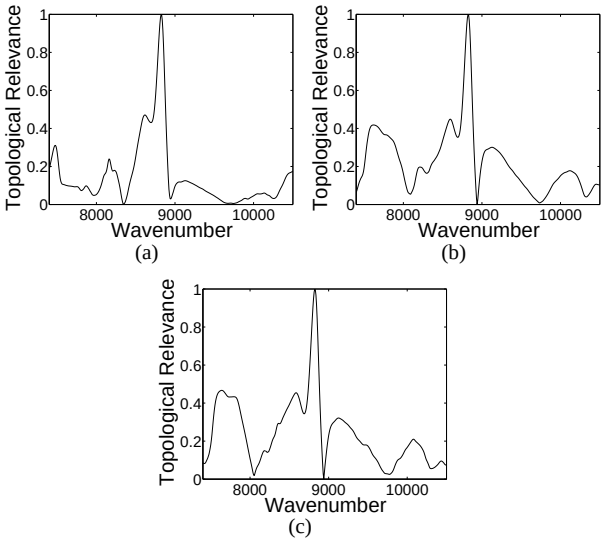


**Figure 4. Case Study 2: The Measures of Topological relevance (MTR) on the Self-Organizing map (SOM) for laboratory (a), pilot-scale (b) and full-scale (c) production.**

In this region, the peak corresponding to the active substance (assigned to the C-H aromatic bond at $\sim 8830 cm^{-1}$), is visible as the shoulder of the broadband of the primary excipient ($\sim 8200 cm^{-1}$). As expected, the proposed method correctly identifies the matching input as the global maximum of $\mathcal{D}$ for all the production scales, in Figure 4. In addition to that, other accompanying inputs, whose assignment to specific vibrational bands is beyond the scope of this work, are also selected in correspondence to the local maxima. However, it is worthwhile noticing that the procedure is able to find a match with specific features in the active substance's spectrum (for instance, $\sim 7500 cm^{-1}$ and $\sim 8600 cm^{-1}$) while assigning a reduced relevance to secondary inputs that are known to be less informative.

Given the analogy between the results obtained with the different production scales, we considered only the laboratory-scale and pilot-scale measurements and reapplied the methodology.

| | $\check{x}_1$ | $\check{x}_2$ | $\check{x}_3$ | $\check{x}_4$ | $\check{x}_5$ | $\check{x}_6$ |
|---|---|---|---|---|---|---|
| $[\text{cm}]^{-1}$ | 7539 | 8200 | 8631 | 8831 | 9101 | 10116 |

**Table 4. Case Study 2: The set of selected inputs and associated wavenumbers.**

The subset of selected variables is summarized in Table 4 and the corresponding relevances to the output depicted in Figure 5. The results obtained with the other indexes of relevance are depicted in Figure 6.

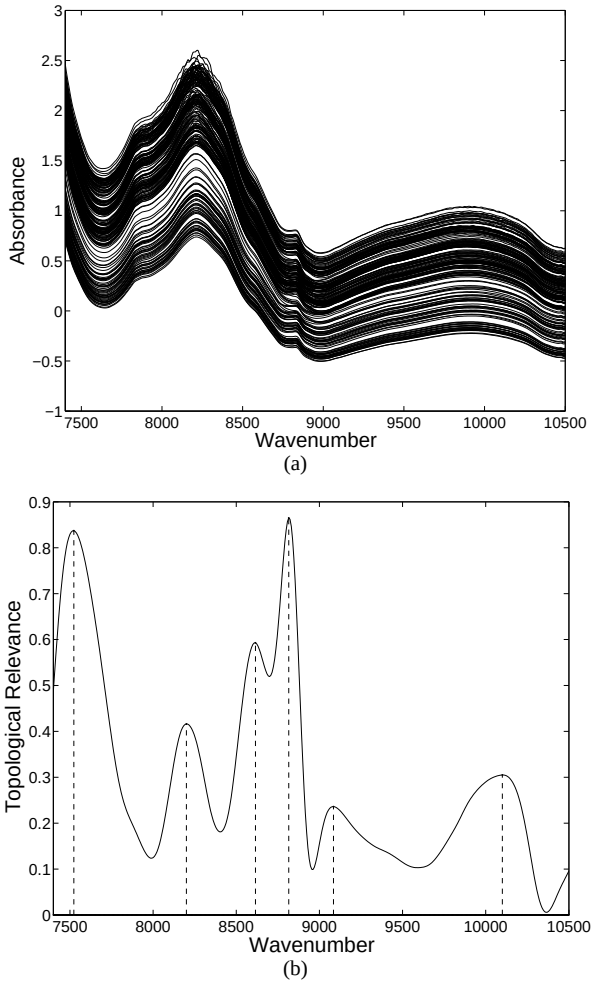


**Figure 5. Case Study 2: The laboratory and pilot-scale spectral observations (a) and the corresponding MTR over the SOM between the inputs and the output (b).**

The prediction accuracy of the regression models used to reconstruct $f$ from the $6$ selected inputs $\check{\mathbf{x}}$ is reported in Table 5 together with the full-spectrum PCR and PLSR results. The results refer to the testing observations consisting of only the entire set of full-scale measurements. Again, the proposed method is not only capable to select the relevant inputs but shows that the associated LS-SVM model gives a prediction accuracy that outperforms the standard PCR and PLSR models. Interestingly, from the noise variance estimates reported in Figure 6(c), it possible to notice that the accuracy of the regression models can be further improved.

Being based on $6$ original variables, the resulting mod-

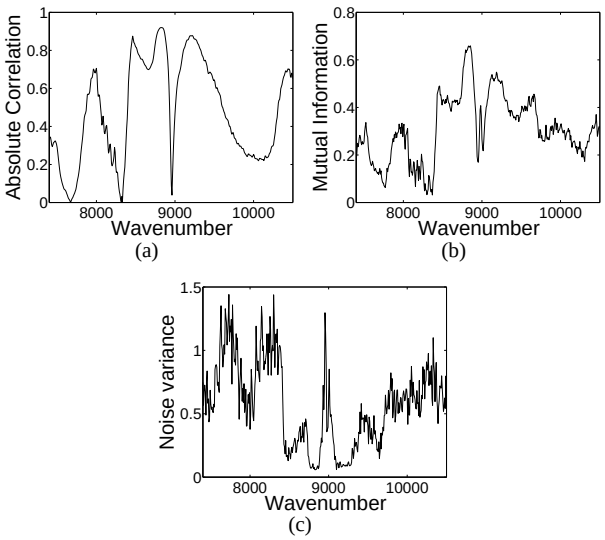


**Figure 6. Case Study 2: Other input-output measures of dependence - Correlation Coefficient (a) and Mutual Information (b) - and the Gasser's noise variance estimate (c).**

| | Number of Variables | $\text{RMSE}_{\mathcal{T}}$ |
|---|---|---|
| PCR | 6 (latent) | 0.44 |
| PLSR | 5 (latent) | 0.42 |
| OLS | 6 (original) | 0.38 |
| Ridge | 6 (original) | 0.38 |
| LS-SVM | 6 (original) | 0.22 |

**Table 5. Case Study 2: A comparison between the results in full-scale production.**

els could be successfully embedded in a soft sensing device capable of obtaining very accurate results and robust to the different properties tablets deriving from interfering artifacts and different production operations.

## 4. Conclusions

In this paper, a methodology for variable selection based on the Measures of Topological Relevance over the Self-Organizing Map was presented and discussed within the context of spectroscopic modeling. The selection methods was applied to monitoring problems in process industry.

From the obtained results a major consideration can be drawn. The sparsity of the obtained models and the good quality of the predictions is, indeed, an advantage because of the interpretability of the results. Moreover, the reduced number of selected variables leaded to simple and robust estimation models that can be readily implemented on-line.

The methodology will be further investigated and validated. It is our goal to assess its potentiality with other problems of industrial interest.

## References


[1] N. Benoudjit, E. Cools, M. Meurens, and M. Verleysen. Chemometric calibration of infrared spectrometers: Selection and validation of variables by non-linear model. *Chemometrics and Intelligent Laboratory Systems*, 70:47–53, 2004.

[2] R. Bolton, D. Hand, and A. Webb. Projection techniques for nonlinear principal component analysis. *Statistics and Computing*, pages 267–276, 2003.

[3] A. Cichocki and R. Unbehauen. *Neural Networks for Optimization and Signal Processing*. Wiley, New York, 1993.

[4] F. Corona and A. Lendasse. Input selection and function approximation using the som: An application to spectrometric modeling. Workshop on Self-Organizing Maps, pages 653–660, 2005.

[5] F. Corona, L. Sassu, S. Melis, and R. Baratti. Measures of topological relevance for soft sensing product properties. IFAC Internationational Symposium on Dynamics and Control of Process Systems, pages 175–180, 2007.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[7] M. Dyrby, S. B. Engelsen, L. Nørgaard, M. Bruhn, and L. Lundsberg-Nielsen. Chemometric quantification of the active substance (containing C≡N) in a pharmaceutical near-infrared (NIR) transmittance tablet using NIR FT-raman spectra. *Applied Spectroscopy*, 56:579–585, 2002.

[8] E. Erwin, K. Obermayer, and K. Schulten. Self-organizing maps: Stationary states, metastability and convergenge rate. *Biological Cybernetics*, 67:35–45, 1992.

[9] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis*. Springer, New York, 2006.

[10] P. A. Gasser, H. G. Müller, M. Köhler, L. Molinari, and A. Prader. Nonparametric regression analysis of growth curves. *Annals of Statistics*, 12:210–229, 1984.

[11] P. Geladi. Recent trends in calibration literature. *Chemometrics and Intelligent Laboratory Systems*, 60:211–224, 2002.

[12] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer, Boston, 1992.

[13] P. A. Guyon and A. Elisseeff. Introduction to variable selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[14] T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2001.

[15] A. S. T. M. International. *Annual Book of ASTM Standards - Petroleum Products and Lubricants*, volume 05. 2007.

[16] J. J. Kelly and B. Callis. Nondestructive procedure for simultaneous estimation of the major classes of hydrocarbon constituents of finished gasolines. *Analytical Chemistry*, 62:1444–1451, 1990.

[17] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, third edition, 2001.

[18] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review, Series E*, 69:066138, 2004.

[19] A. J. Miller. *Subset Selection in Regression*. Chapman & Hall, London, 1990.

[20] B. Nadler and R. R. Coifman. Prediction error in CLS and PLS: The importance of feature selection prior multivariate calibration. *Journal of Chemometrics*, 19:107–118, 2005.

[21] E. Oja. Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, 1:61–68, 1989.

[22] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, second edition, 2005.

[23] H. Ritter and K. Schulten. Convergence properties of Kohonen's topology conserving maps: Fluctations, stability and dimension selection. *Biological Cybernetics*, 60:59–71, 1988.

[24] F. Rossi, A. Lendasse, D. François, W. Wertz, and M. Verleysen. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics and Intelligent Laboratory Systems*, 80:215–226, 2006.

[25] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

[26] J. A. K. Suykens, T. V. Gestel, J. de Brabanter, B. de Moor, and J. Vanderwalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[27] A. Ultsch. Self-organizing neural networks for visualization and classification. In *Information and Classification*, pages 307–313. Springer, Berlin, 1993.

[28] J. Vesanto. SOM-based data visualization methods. *Intelligent Data Analysis*, 3:111–126, 1999.

[29] J. Vesanto and J. Ahola. Hunting for correlations in data using the self- organizing map. Computational Intelligence Methods and Applications, pages 279–285, 1999.

[30] L. G. Weyer. Near infrared spectroscopy of organic compounds. *Applied Spectroscopy Reviews*, 21:1–43, 1985.

[31] O. H. Wheeler. Near-infrared spectra of organic compounds. *Chemical Reviews*, 59:629–666, 1959.

[32] J. J. J. Workman. Review of process and non-invasive near-infrared and infrared spectra. *Applied Spectroscopy Reviews*, 34:1–89, 1999.