8th International IFAC Symposium on
Dynamics and Control of Process Systems
Preprints Vol.3, June 6-8, 2007, Cancún, Mexico

MEASURES OF TOPOLOGICAL RELEVANCE FOR SOFT SENSING PRODUCT PROPERTIES

Francesco Corona * **Lorenzo Sassu** **
Stefano Melis ** **Roberto Baratti** *

* *Università degli Studi di Cagliari, Dipartimento di Ingegneria Chimica e Materiali, Cagliari, Italy*
** *SARTEC - Saras Ricerche e Tecnologie S.p.A., Divisione Ricerche, Assemini (CA), Italy*

Abstract: We address the problem of real-time monitoring of products' properties from spectroscopic measurements. Spectra are used as inputs for soft-sensors that estimate outputs difficult to measure on-line. To overcome the issues associated to calibrating such models from high-dimensional inputs, we propose to select the relevant inputs emerging from the topological structure of the data. The approach is independent on the estimation model to be embedded in the sensor. Being based on the original spectral features, the models retain the interpretability of the underlying system. The application of the method is illustrated on two cases from refining and pharmaceutical industry.

Keywords: Observers, Spectroscopy, Non-parametric regression.

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 specifically spectroscopy, fulfill the necessary requirements for real-time analysis of key chemical and physical properties in a wide range of materials.

The principle underlying process monitoring by Near-Infrared (NIR) spectroscopy is the existence of a relationship between the light's absorbance spectrum of a given specimen and the property of interest. The relationship can be reconstructed by calibrating specific data-driven models, without an explicit regard to first-principle criteria. The resulting models are efficiently used as soft-sensors to estimate the key property from the measured spectrum (Workman, 1999). However, the problem of estimating the property (the output variable) is defined from high-dimensional and inherently redundant inputs (the spectrum). Furthermore, it is not unusual to calibrate models on a number of samples that is radically smaller than the number of input candidates. Operating in such a condition (Donoho, 2000), seldom encountered in other contexts, may lead to ill-posed estimation settings.

To address this problem, two approaches are commonly used. One standard solution is to rely on full-spectrum methods for dimension reduction coupled with regression: Principal Components Regression (PCR) and Partial Least-Squares (PLS) are reference models (Geladi, 2002). The natural refinement of such an approach is to perform a preliminary selection of relevant spectral ranges (Nadler and Coifman, 2005). The alternative solution consists of selecting, among all spectral variables, only those inputs that truly contribute to a correct estimation of the output. The approach is either based on model properties

(Benoudjit *et al.*, 2004), or on relevance indexes (Rossi *et al.*, 2006). Thus, variable selection is the limit extension of range selection where the interpretability of the system can be maintained.

In this study, variable selection is approached by exploiting the metric structure of the spectral data, leading to a method that identifies only the 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) where, the relevance of the inputs is measured from Unified-distance matrices (U-matrix). Because designed on the original spectral inputs, the resulting soft-sensors retain a usefull interpretability of the underlying system. Moreover, the approach is model-independent; in fact, once the variables are selected any estimation technique can be used; the Least-Squares Support-Vector Machine (LS-SVM) is here preferred. The application of the method is discussed on two cases from the refining and pharmaceutical industry.

2. METHOD AND ALGORITHMS

The problem of monitoring product properties from NIR 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 generated the observations. Because of the high dimensionality of $\mathbf{x}$ and the small number N of samples, it is appropriate to operate in a reduced data space whose dimensionality is circumscribed by the intrinsic complexity of the observed system. Formally, being $\mathbf{x} \in \mathbb{R}^d$ the given set of 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 (Guyon and Elisseeff, 2003).

Here, a two-stage method is proposed:

(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 inputs. The inputs that best match the topology of the output are selected as relevant.

Once the subset $\check{\mathbf{x}}$ of inputs is selected, any model of the functionality f can be calibrated and used to predict the output y.

2.1 Topology-preserving mappings using the SOM

The Self-Organizing Map, SOM (Kohonen, 2001), is an adaptive algorithm to formulate the vector-quantization paradigm. In the following, the essential properties of the SOM algorithm are briefly reported.

The basic SOM consists of a low-dimensional (typically, 2D) regular array of M nodes where a parameter vector $\mathbf{m}_l \in \mathbb{R}^n$ is associated with every node l. Each parameter acts as an adaptive model vector for the observation $\mathbf{z}_i \in \mathbb{R}^n$ (in the addressed context of spectroscopy, $\mathbf{z}_i = [\mathbf{x}_i; y_i]$ and $n = d + 1$). During the computation of the SOM, the observations are mapped into the array of nodes and the parameters of the model vectors adapted according to the rule

$$\mathbf{m}_l(t+i) = \mathbf{m}_l(t) + h_{l,c(\mathbf{z_i})}[\mathbf{z}_i(t) - \mathbf{m}_l(t)], \quad (1)$$

where t is the discrete-time coordinate of the mapping step. The map is computed recursively for each observation. The scalar multiplier $h_{l,c(\mathbf{z_i})}$ in Equation 1 is a neighborhood kernel that, if chosen in its Gaussian form, acts as a smoothing function centered around the Best Matching Unit (i.e., the model vector $\mathbf{m}_c$ that best matches with the observation vector $\mathbf{z}_i$, BMU); that is,

$$h_{l,c(\mathbf{z_i})} = \alpha(t) \exp\Big(- \frac{||\mathbf{r}_l - \mathbf{r}_c||^2}{2\sigma^2(t)} \Big), \quad (2)$$

with $\sigma(t)$ denoting the monotonically decreasing width. The learning rate $\alpha(t) \in (0, 1)$ also decreases monotonically, so that: as $h_{l,c(\mathbf{z_i})} \to 0$, the models $\mathbf{m}_l$ are adaptively updated toward their asymptotic limits. The vectors $\mathbf{r}_l$ and $\mathbf{r}_c$ (both in $\mathbb{R}^2$, for the 2D map) represent the geometric location of the nodes on the array. After performing a parallel comparison algorithm, the subscript c (in Equation 1 and 2) was assigned to the BMU. The usual criterion for comparison is the Euclidean metric $||\cdot||$; hence, $\mathbf{m}_c(t)$ followed from

$$||\mathbf{z}_i(t) - \mathbf{m}_c(t)|| \leq ||\mathbf{z}_i(t) - \mathbf{m}_l(t)||, \quad \forall l. \quad (3)$$

The resulting model vectors $\{\mathbf{m}_l\}_{l=1}^M$ form a nonlinear submanifold in the data space where the relevant topological and metric properties of the observations are locally preserved. The SOM is to be understood as an image of the original high-dimensional data manifold as modelled onto a low-dimensional space that displays complex data structures with simple geometric relationships.

2.2 SOM-based measures of topological relevance

The Self-Organizing Map is 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 representations of the map (Vesanto, 1999) or estimating the correlation coefficients proposed by Vesanto and Ahola (1999).

We propose to identify the relevant inputs by exploiting the topology preserving properties of the SOM of the input and output data, and we suggest a relevance measure derived from the assumed continuity of the unknown functionality $y = f(\mathbf{x})$. In this hypothesis, if two points $\mathbf{x}_i$ and $\mathbf{x}'_i$ are close 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: i.e., f can be reconstructed pointwise from the locally linear topology of the neighborhoods. If the neighborhood continuity is not satisfied, it can be either due to noise or because the inputs are not relevant to predict the output. In order to benefit from the noise-filtering properties of the SOM, this general principle can be explored from the models $\{\mathbf{m}_l\}_{l=1}^{M}$ of the map.

The standard approach to recover the topological structure of the data from the SOM is to compute the Unified-distance Matrix, U-matrix (Ultsch, 1993). The U-matrix $\mathbf{U}$ is built from distances between each node and its neighbors. For the sake of brevity, we only recall that: letting $\mathbf{m}_l$ the model associated to node l and $\mathcal{N}(l)$ its neighborhood of K adjacent nodes k ($K = 6$, for the usual hexagonal array), the entries of $\mathbf{U}$ are calculated from:

- local pairwise distances, for all k:
$$d(l,k) = ||\mathbf{m}_l - \mathbf{m}_k||, \tag{4}$$
- locally averaged distances in $\mathcal{N}(l)$:
$$d(l) = \frac{1}{K} \sum_{k \in \mathcal{N}(l)} ||\mathbf{m}_l - \mathbf{m}_k||. \tag{5}$$

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
$$\mathcal{D}(x_j; y) = ||\mathbf{U}_{x_j} - \mathbf{U}_y||_F, \tag{6}$$
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 $\mathcal{D}(x_j; y) \in [0, +\infty)$ is preferably inverted and scaled into $\mathcal{D}^s(x_j; y) \in [0, 1]$; so that, the higher is the relevance, the closer to 1 is the measure. In principles, variable selection is then simply performed by ranking the inputs according to their relevance to the output, and selecting a reduced but still representative subset $\hat{\mathbf{x}} \in \mathbb{R}^s$.

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.

In such context, the selection scheme proposed by Corona and Lendasse (2005) can be easily adapted to the topological measures of relevance defined in Equation 6. Here, the modification of the procedure summarizes as:

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

The procedure identifies only the inputs that are associated to the local maxima of $\mathcal{D}^s$, 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 redundances 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 $\mathcal{D}^s$.

From the set $\check{\mathbf{x}}$ of selected variables, any model that estimates the functionality f can be calibrated and used to predict the output y. The technique preferred in our applications is a *de facto* standard in nonparametric function estimation: the Least-Squares formulation of the Support-Vector Machine, LS-SVM (Suykens *et al.*, 2002). The meta-paramaters of the LS-SVM model are validated with standard resampling methods that estimate the prediction accuracy (Hastie *et al.*, 2001); the Leave-One-Out Cross-Validation (LOO-CV) is here adopted.

3. APPLICATIONS

The development and the application of the studied soft-sensors is illustrated with two actual mon-

itoring tasks from the refining and the pharmaceutical industry. The selected applications are referenced as benchmarks for variable selection and interpretation, as well as prediction purposes.

3.1 Case Study I

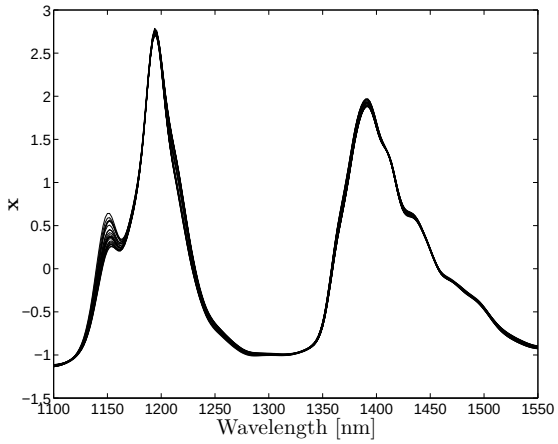
The first application consists of estimating the octane number in gasoline fuels. Real-time monitoring of such property is of fundamental importance for both the production and the blending process of finished gasolines. The design of the soft-sensor is discussed on a dataset provided by Camo Inc., which is gratefully acknowledged. Although in reduced amount, the data are collected over a sufficient period of time considered to span all the important variations in the production. Being the relationship between the octane and the spectrum distributed among different inputs, the application is interesting because variable selection cannot be easily performed through first-principle interpretation of the spectra.

The absorbance spectra are acquired by means of a spectrophotometer operating in the $1100 - 1550nm$ wavelengths' range, in Figure 1(a). The absorbance is measured on the basis of the NIR transmission principle with a $2nm$ resolution. The measurements of the octane number (in the $86-92$ range) are evaluated in laboratory by reference 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 preprocessed by removing the outliers and with mean-centering.

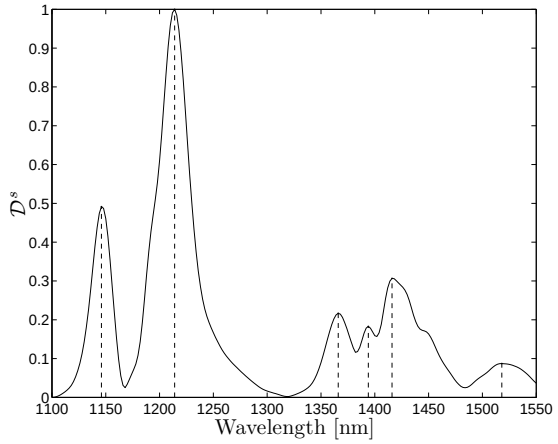
According to the method discussed in Section 2, the 2D SOM of the input and output observations in the calibration set was computed. The map consists of a hexagonal array of nodes initialized in the space spanned by the eigenvectors corresponding to the 2 largest eigenvalues of the covariance matrix of the data. As usual, the ratio between these eigenvalues was also used to calculate the size (5×5 nodes) of the SOM. On the map, the set of topological relevances $\mathcal{D}^s = \mathcal{D}^s(x_j; y)\}_{j=1}^d$ 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 $\mathcal{D}^s$, in Figure1(b).

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

Table 1. Case Study I: the selected inputs and associated wavelengths.



(a)



(b)

Fig. 1. Case Study I: (a) a selection of input spectra; (b) the topological relevances between the inputs and the output and the selected inputs (dashed vertical lines).

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 (Kelly and Callis, 1990). The analysed 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 2$^{\text{nd}}$ overtone ($1100-1300nm$) and to the combination bands ($1300 - 1550nm$) of the C-H vibrations. In details, the aromatic bonds at $\sim 1150nm$ ($\check{x}_1$) are related to an increase in octane number. Conversely, the methylene bonds at $\sim 1220nm$ ($\check{x}_2$) indicate the presence of linear hydrocarbons which are responsible for a reduction in the gasoline quality. The methyl bonds at $\sim 1200nm$ 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 $1200nm$, 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 ($\sim 1395/1416nm$), and methyl ($\sim 1360/1345nm$) 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 1$^{\text{st}}$ methyl band. As already noticed above, again the 2$^{\text{nd}}$ 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 and its selection can be ascribed to baseline effects.

Finally, the LS-SVM was calibrated to model the functionality $y = f(\mathbf{x})$ from $\check{\mathbf{x}}$ and its meta-parameters validated with LOO-CV. The prediction accuracy of the model is evaluated in terms of Root Mean Squared Error on the independent set of testing observations (RMSE_T); in Table 2, the result is compared to the standard calibration method used in spectroscopy, the full-spectrum PLS. The number of latent variables in the PLS model was also selected with LOO-CV.

	Number of Variables	RMSE_T
LS-SVM	6 (Original)	0.2642
PLS	4 (Latent)	0.2760

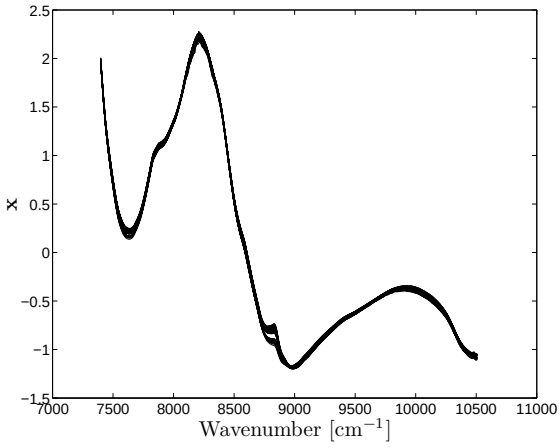
Table 2. Case Study I: a comparison between prediction results.

From Table 2, it is possible to notice that the LS-SVM gives prediction results that are comparable to the standard PLS model. More importantly, the method is capable to select only those inputs carrying important information, thus, leading to a parsimonious and yet accurate soft-sensor.

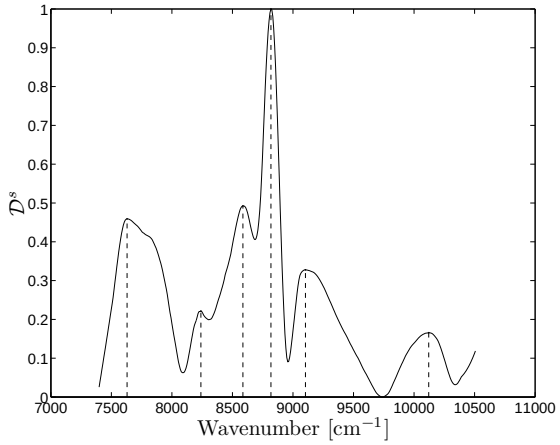
3.2 Case Study II

The second application consists of estimating the active substance content in pharmaceutical tablets. The problem is discussed in details by Dyrby *et al.* (2002), which are gratefully aknowledged for providing 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.

The spectra are acquired in the $4000-14000cm^{-1}$ wavenumbers' range ($700-2500nm$) with a resolution of $16cm^{-1}$; however, the absorbances are available only for the $7400-10500cm^{-1}$ interval, in Figure 2(a). The content of active substance (in the $5.6-8.0\%w/w$ range) is evaluated by the reference High Performance Liquid Chromatography



(a)



(b)

Fig. 2. Case Study II: (a) a selection of input spectra; (b) the topological relevances between the inputs and the output and the selected inputs (dashed vertical lines).

method. Each sample consists of a 404−channel spectrum ($\mathbf{x} \in \mathbb{R}^d$, with $d = 404$) and the content of active substance ($y \in \mathbb{R}$). The dataset contains 120 observations divided in calibration/validation and testing sets, with 60 samples each.

The calibration samples are mapped onto the SOM and the relevances $\mathcal{D}^s$ between the input-output pairs are calculated, in Figure2(b); only 6 inputs are identified as relevant to the output (Table 3).

	$\check{x}_1$	$\check{x}_2$	$\check{x}_3$	$\check{x}_4$	$\check{x}_5$	$\check{x}_6$
[cm^{-1}]	7630	8216	8602	8818	9103	10137

Table 3. Case Study II: the selected inputs and associated wavenumbers.

As reported by Dyrby *et al.* (2002), 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. In this region, the peak corresponding to the active substance (C-H aromatic bond at $\sim 8830 cm^{-1}$), visible as the shoulder of the broad-band of the primary excipient ($\sim 8200 cm^{-1}$), yields the highest correlation with the active substance's concentration. As expected, the proposed method correctly identifies the matching input ($\check{x}_4$) as the global maximum of $\mathcal{D}^s$. In addition to that, the 5 accompanying inputs, whose assignment to specific vibrational bands is beyond the scope of this work, are also selected in correspondence to the local maxima. Anyways, it is worthwhile noting that the proposed procedure is able to find a match with specific features in the active substance's spectrum while assigning a reduced relevance to secondary inputs that are less informative.

In Table 4, the prediction accuracy of the LS-SVM model used to reconstruct f from the 6 selected inputs $\check{\mathbf{x}}$ is reported for comparison with a full-spectrum PLS. The results refer to the testing observations.

	Number of Variables	RMSE_T
LS-SVM	6 (Original)	0.2373
PLS	4 (Latent)	0.2352

Table 4. Case Study II: a comparison between prediction results.

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 comparable to the standard PLS model.

4. CONCLUSION

In this paper, a method to address the problem of variable selection to estimate quality indexes in products from NIR spectra is proposed. The selection method is based on input-output topological relevances. The reduced number of selected variables leads to simple and robust estimation models which are reliable and accurate process analysis and monitoring tools.

REFERENCES

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

Corona, F. and A. Lendasse (2005). Input selection and function approximation using the SOM: An application to spectrometric modelling. WSOM'05, Workshop on Self-Organizing Maps. pp. 253–260.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. AMS'00 American Mathematical Society Conference, "Mathematical Challenges of the 21st Century".

Dyrby, M., S. B. Engelsen, L. Nørgaard, M. Bruhn and L. Lundsberg-Nielsen (2002). 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.

Geladi, P. (2002). Recent trends in calibration literature. *Chemometrics and Intelligent Laboratory Systems* **60**, 211–224.

Guyon, I. and A. Elisseeff (2003). Introduction to variable selection. *Journal of Machine Learning Research* **3**, 1157–1182.

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

Kelly, J. J. and J. B. Callis (1990). Nondistructive analytical procedure for simultaneous estimation of the major classes of hydrocarbon constituents of finished gasolines. *Analytical Chemistry* **62**, 1444–1451.

Kohonen, T. (2001). *Self-Organizing Maps*. 3rd extended ed.. Springer. Berlin.

Nadler, B. and R. R. Coifman (2005). The prediction error in CLS and PLS: The importance of feature selection prior to multivariate calibration. *Journal of Chemometrics* **19**, 107–118.

Rossi, F., A. Lendasse, D. François, W. Wertz and M. Verleysen (2006). Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics and Intelligent Laboratory Systems* **80**, 216–226.

Suykens, J. A. K., T. Van Gestel, J. de Brabanter, B. de Moor and J. Vanderwalle (2002). *Least Squares Support Vector Machines*. World Scientific. Singapore.

Ultsch, A. (1993). Self-organizing neural networks for visualization and classification. In: *Information and Classification*. pp. 307–313. Springer. Berlin.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis* **3**, 111–126.

Vesanto, J. and J. Ahola (1999). Hunting for correlations in data using the self-organizing map. CIMA'99, Computational Intelligence Methods and Applications. pp. 279–285.

Workman, J. J. Jr. (1999). Review of process and non-invasive near-infrared and infrared spectroscopy. *Applied Spectroscopy Reviews* **34**, 1–89.