ELSEVIER

Journal of Pharmaceutical and Biomedical Analysis
30 (2002) 453–466

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

www.elsevier.com/locate/jpba

# An evaluation of direct orthogonal signal correction and other preprocessing methods for the classification of clinical study lots of a dermatological cream

J. Luypaert [a], S. Heuerding [b], S. de Jong [c], D.L. Massart [a,*]

[a] *ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *Novartis Pharma AG, Pharmaceutical and Analytical Development, CH-4002 Basel, Switzerland*
[c] *Unilever Research Vlaardingen, P.O. Box 114, 3130 AC Vlaardingen, The Netherlands*

Received 4 April 2001; received in revised form 18 February 2002; accepted 22 February 2002

**Abstract**

Different preprocessing methods (direct orthogonal signal correction (DOSC), standard normal variate (SNV), multiple scatter correction (MSC), first and second derivation, offset correction and detrend correction) are applied to two sets of NIR spectra of a dermatological cream with different concentrations of an active compound. The influence of these preprocessing methods on the classification of the samples into the right concentration class is evaluated using 1 and 3 nearest neighbour method (with Euclidean distance and correlation coefficient as distance parameters) as classification method. PLS and PCR modelling are also used to make a prediction of the concentration of the active compound. The direct orthogonal signal correction gives best results in most of the classification methods. 

*Keywords:* k Nearest neighbours; Pharmaceutical cream; NIR spectroscopy; Pre-processing; Orthogonal signal correction

## 1. Introduction

Clinical study lots are designed to compare the effect of new drugs with a placebo formulation or with existing drugs from the same class. Identification of such study lots is usually carried out with chromatographic methods, which are time consuming and elaborate, since samples need a lot of sample preparation. Industry is looking for faster methods and NIR combined with pattern recognition methods seems to be a promising technique [1–6].

In this work, NIR is applied to identify creams with different concentrations of an active compound. After the NIR spectra are acquired, a preprocessing method can be applied to remove certain physical light effects, thus enhancing the chemical information in the spectra. In this work, the effect of seven different preprocessing methods on the classification of the samples is analysed. The results from the relatively new preprocessing method called direct orthogonal signal correction

* Corresponding author. Tel.: +32-2-477-4734; fax: +32-2-477-4735
*E-mail address:* fabi@vub.vub.ac.be (D.L. Massart).

0731-7085/02/$ - see front matter 
PII: S0731-7085(02)00401-6

(DOSC) are studied in more detail and compared with the results from other preprocessing methods. The different classes are known and a new sample has to be assigned to the most similar class. The k-nearest neighbour method in two variants (1 nearest neighbour and 3 nearest neighbours) is used as a classification method. This is a non-parametric method in which no assumptions of the data distribution are required [7]. The Euclidean distance (ED) and the correlation coefficient (CC) are used as (dis-)similarity measures.

Two sample presentations are used. First an optical fibre attached to the instrument is used to perform the NIR measurements and a second way of measuring is the internal measuring mode of the instrument in which the sample cup is put into the sample drawer of the instrument. To evaluate the results, the correct classification rates (CCR) of the test set are compared.

## 2. Theory

### 2.1. Preprocessing methods

#### 2.1.1. Direct orthogonal signal correction (DOSC)

Orthogonal signal correction [8] is applied to NIR spectra ($\boldsymbol{X}$) to remove from the spectra as much as possible the variation that is unrelated (i.e. orthogonal) to $y$, the vector of the parameter which has to be modelled (e.g. the concentration). After the DOSC correction, a new PLS or PCR model can be built and this model will be less complex than the model built with the original, uncorrected data. Algorithms to perform an orthogonal signal correction have been proposed by Wold [9], Sjöblom [10], Wise and Gallagher [11] and Fearn [12]. A similar method called direct orthogonalisation has been developed by Andersson [13]. All of these methods find an approximate solution to the problem set out, i.e. finding a subspace of $\boldsymbol{X}$ that is orthogonal to $y$ and accounts for the largest possible proportion of $X$-variance. Recently, the exact solution was found independently by Westerhuis et al. [8]. The corresponding algorithm was coined direct orthogonal signal correction (DOSC). The first step in this algorithm is a decomposition of $\boldsymbol{X}$ (the spectral matrix) into two orthogonal parts, one part related to $y$ and another part that is orthogonal to it. This is carried out by projecting (or regressing) $\boldsymbol{y}$ onto $\boldsymbol{X}$. In this way, one decomposes $y$ into $\hat{\boldsymbol{y}}$, the part of $\boldsymbol{y}$ lying in $X$-space and $\boldsymbol{f}$, the residual that is unrelated to $\boldsymbol{X}$, i.e.

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{f} \tag{1}$$

For spectral data, $\boldsymbol{X}$ is generally of less than full column rank giving $\boldsymbol{y} = \boldsymbol{y}$. The column rank of a matrix is the number of independent columns (wavelengths). Since in (NIR) spectra, neighboring wavelengths are not independent, but highly correlated, the rank of a spectral matrix is less than its number of columns.

Next, $\boldsymbol{X}$ is projected onto $\hat{\boldsymbol{y}}$ giving $\hat{\boldsymbol{X}}$ and $\boldsymbol{E}$, the residual part of $\boldsymbol{X}$ that is orthogonal both to $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$, i.e.

$$\boldsymbol{X} = \tilde{\boldsymbol{X}} + \boldsymbol{E} \tag{2}$$

Principal component analysis (PCA) or singular value decomposition (SVD) is applied to $\boldsymbol{E}$ in order to find a small number of principal components $\boldsymbol{T}$ corresponding to the largest singular values. This $\boldsymbol{T}$ is a basis for the low-dimensional subspace that accounts for the maximum of variance of $\boldsymbol{E}$, the part of $\boldsymbol{X}$ that is unrelated to $y$.

The DOSC-corrected spectra of the calibration data can now be written as:

$$\boldsymbol{X}_{\mathrm{DOSC}} = \boldsymbol{X} - \boldsymbol{T}\boldsymbol{P}' \tag{3}$$

where $\boldsymbol{T}\boldsymbol{P}'$ is the 'orthogonal' part removed from the original spectra with $\boldsymbol{P}$ the loading matrix:

$$\boldsymbol{P} = \boldsymbol{X}\boldsymbol{T}(\boldsymbol{T}'\boldsymbol{T})^{-1} \tag{4}$$

The directions $\boldsymbol{T}$ can be expressed as linear combinations of $\boldsymbol{X}$:

$$\boldsymbol{T} = \boldsymbol{X}\boldsymbol{R} \tag{5}$$

Here, $\boldsymbol{R}$ is the matrix of weights of the original variables in the principal orthogonal directions, which can be obtained via $\boldsymbol{X}^{+}$, the Moore–Penrose generalised inverse of the original data $\boldsymbol{X}$:

$$\boldsymbol{R} = \boldsymbol{X}^{+}\boldsymbol{T} \tag{6}$$

Given weights $\boldsymbol{R}$ and loadings $\boldsymbol{P}$, one can directly obtain corrected spectra for new data:

$$X_{\mathrm{DOSC}} = X - XRP' \quad (7)$$

### 2.1.2. Other preprocessing methods

*2.1.2.1. Standard normal variate transformation (SNV).* SNV [14–17] is a mathematical transformation method of the log(1/$R$) spectra used to remove slope variation and to correct for scatter effects. Each spectrum $\boldsymbol{x}_i$ is corrected individually by first centring the spectral values (i.e. subtracting the mean log(1/$R$) value $\bar{x}_i$ of the individual spectrum from each value $x_{ij}$). Then the centred spectra are scaled by the S.D. ($s_i$) calculated from the individual spectral values.

The SNV transformed spectrum ($\boldsymbol{x}_{i,\mathrm{cor}}$) is:

$$\boldsymbol{x}_{i,\mathrm{cor}} = (\boldsymbol{x}_i - \bar{x}_i)/s_i \quad (8)$$

*2.1.2.2. Multiplicative scatter correction (MSC).* MSC [14–16,18] corrects for difference in light scatter between samples before calibration. The average spectrum $\bar{\boldsymbol{x}}$ of the calibration set is chosen as a reference spectrum. Then each spectrum $\boldsymbol{x}_i$ is first modelled as follows:

$$\boldsymbol{x}_i = a_i + b_i\bar{\boldsymbol{x}} + \boldsymbol{c}_i \quad (9)$$

where $\boldsymbol{e}_i$ is the vector of residuals representing the difference between $\bar{\boldsymbol{x}}$ and $\boldsymbol{x}_i$ ($\boldsymbol{e}_i$ represents the chemical information in $\boldsymbol{x}_i$), $a_i$ is the intercept and $b_i$ the slope. $a_i$, $b_i$ and $\boldsymbol{e}_i$ can be estimated by linear regression of $\boldsymbol{x}_i$ versus $\bar{\boldsymbol{x}}$. The corrected spectrum $\boldsymbol{s}_i$ is written as:

$$\boldsymbol{s}_i = (\boldsymbol{x}_i - a_i)/b_i \quad (10)$$

*2.1.2.3. First and second derivation after smoothing.* By calculating first and second derivatives, baseline drifts are eliminated and also small spectral differences are enhanced. To avoid enhancing the noise, which is a consequence of derivation, spectra are first smoothed [19,20]. This smoothing is performed by using the Savitzky–Golay algorithm, which is a moving window averaging method: a window is selected where the data are fitted by a polynomial of a certain degree. The central point in the window is replaced by the value of the polynomial.

*2.1.2.4. Offset correction.* Offset correction [21] is used to correct for a parallel baseline shift. The absorbance of a chosen wavelength $c$ (usually the first) is subtracted from each spectrum independently.

$$\boldsymbol{x}_{i,o} = \boldsymbol{x}_i - x_{ic} \quad (11)$$

with $\boldsymbol{x}_{i,o}$ the offset transformed spectrum, $\boldsymbol{x}_i$ the original spectrum and $x_{ic}$ the absorbance of the chosen wavelength of that spectrum.

*2.1.2.5. Detrending.* Detrending [17] is applied on spectra to remove the effects of baseline shift and curvi-linearity. It is characteristic for NIR spectra that the log(1/$R$) values increase from 1100 to 2500 nm. This effect is generally linear but for densely packed samples, it becomes curvilinear.

The method consists of modelling the baseline as a function of wavelength with a second-degree polynomial and this function is then subtracted from each spectrum independently.

$$\boldsymbol{x}_{i,d} = \boldsymbol{x}_i - \boldsymbol{b}_i \quad (12)$$

where $b_i$ is the baseline 'spectrum' at wavelength $i$ computed with the second-degree model.

## 2.2. Feature selection method

### 2.2.1. Fisher criterion (FC)

The Fisher criterion [21] represents the ratio of the between-group variance over the within group variance. Variables with a high between group variance and a small within group variance have an important value for the discrimination. The Fisher criterion (FC) is calculated as follows:

$$\mathrm{FC} = \sum_{j=1}^{k} n_j(\bar{x}_{ji} - \bar{x}_i)^2 / \sum_{j=1}^{k} (n_j - 1)s_{ji}^2 \quad (13)$$

with $k$, the number of groups, $n_j$ the number of objects in group $j$, $x_{ji}$ the mean absorbance of the objects belonging to group $j$ at the $i$th wavelength, $x_i$ the mean absorbance of the objects belonging to all groups at the $i$th wavelength and $s_{ji}$ is the S.D. of the absorbance of the objects belonging to

group $j$ at the $i$th wavelength. The original variables (wavelengths) are ordered from the highest to the lowest FC-value and then selected top-down, so that only important variables are retained and less important ones are rejected.

### 2.3. Classification method

#### 2.3.1. k-Nearest neighbours method

kNN [7] is a simple non-parametric classification method, where the distances between an unknown object and all objects of the training set are calculated. The new object is attributed to the class to which the distance is the smallest. The variants used here are the 1 and 3 NN method. In the 3 NN variant, the distance of the three nearest neighbours is calculated and the new sample is classified into the group to which the majority of the three samples belong. The condition to apply this method is that the classes consist of similar amounts of objects. Otherwise, more complex variants are required. Distance parameters used in this study are the Euclidean distance (ED) and the correlation coefficient.

*2.3.1.1. Euclidian distance.* The Euclidian distance (ED) [22] is the geometric distance between two objects. The distance $D_{kl}$ between two objects $k$ and $l$ can be mathematically described as:

$$D_{kl} = \sqrt{\sum_{j=1}^{m}(x_{kj} - x_{lj})^2}$$

with $m$ the number of variables. The closer the distance is to zero, the more similar the objects are.

*2.3.1.2. Correlation coefficient.* The correlation coefficient $r(n, m)$ [22] between two objects $n$ and $m$ can be defined as:

$$r(n,m) = \frac{\sum_{i=1}^{I}(x_{in} - \bar{x}_n)(x_{im} - \bar{x}_m)}{\sqrt{\sum_{i=1}^{I}(x_{in} - \bar{x}_n)^2 \sum_{i=1}^{I}(x_{im} - \bar{x}_m)^2}}$$

with $I$ the number of variables and $x_{in}$ the value of the $i$th variable for object $n$. The closer this coefficient is to 1 the more similar the two objects are.

#### 2.3.2. PLS and PCR modelling

Since the aim in this study is to make a classification, no precise prediction of the concentrations is needed. The optimal complexity of the models is not chosen according to the lowest root mean square error of cross validation (rmsecv), but a cut off value of 0.50 for the rmsecv is chosen to select the complexity of the models. The spectra are first column centred.

*2.3.2.1. Principal component regression (PCR).* PCR [22,23] is a multivariate calibration method that consists of two steps. The first step is a principal component analysis (PCA) of the data matrix $X$. In this step, the original variables (e.g. absorbances at different wavelengths) are converted into a smaller number of latent variables (scores on principal components (PC's)) in order to reduce the dimensionality of the original data set. The PC's are constructed so that the first few retain most of the variation contained in all the original variables. The second step is a multiple linear regression (MLR) between the scores obtained in the PCA and the characteristic that has to be modeled. The order of the PC's used in the MLR step is chosen according to the correlation coefficient with respect to the $y$ characteristic. In what follows, the $y$ characteristic is the concentration of active compound in the cream but other properties can also be modeled.

*2.3.2.2. Partial least squares regression (PLS).* PLS [22] is a similar method to PCR. However, in the first step of PCR only $\boldsymbol{X}$ is used to compute the PC's. In PLS, the equivalent of the PC's are obtained such that they not only represent $\boldsymbol{X}$ well, but at the same time are correlated well with the $\boldsymbol{y}$.

## 3. Experimental

### 3.1. Samples

The creams used in this work contain 1 and 3% of an active compound. In addition, placebo

creams are measured. The placebo cream consists of nine ingredients of which water is an important part.

### 3.2. Instrumentation

The spectra are measured with a Bran+Luebbe InfraAlyser 500 spectrometer. For the measurements with the optical fibre, a Bran+Luebbe EDAPT-1 Fibre Optic Reflectance Probe is attached to the InfraAlyser.

### 3.3. Sample presentation

Two different sample presentations are investigated. The first measurements were carried out with the optical fibre. The spectra are taken between 1100 and 2200 nm (with a measuring point every 2 nm) because the optical fibre does not allow performing accurate measurements above 2200 nm. The creams are put into a glass container with a filling height of 0.5 and 0.25 cm and placed onto the optical fibre, so that the light has to pass through the glass material. A second way of sample presentation is the so-called internal measuring mode. The cream is put into a special measuring cup that can be introduced into the sample drawer of the spectrometer. These cups are covered with a microscope glass slide.

### 3.4. Computer program

For all computations, a Matlab (version 4.0) Toolbox, designed in our laboratory, was used.

## 4. Results and discussion

### 4.1. Measurements with the optical fibre

For each concentration level (0, 1 and 3%) of the active compound, four cups are filled with the cream: two cups are filled with a cream layer of 0.5 cm and two cups are filled with a layer of $\approx 0.25$ cm to evaluate the influence of the layer thickness. Unfortunately, the measurements of one of the 0.25 cm cups for the 3% level were not carried out due to a measurement error. The measurements are performed at $20 \pm 1$ °C. Each sample is measured three times a day during 3 subsequent days, so that 99 spectra are obtained. The data set is divided randomly into two subsets: one calibration or training set with 57 spectra and a test set with 42 spectra. On the plots in Fig. 1 the test spectra are shown without pre-treatment and after DOSC.

A clear distinction is observed between the three different concentrations for the DOSC pre-treated data, but not for the original data. The DOSC method works as a kind of background subtraction: for the placebo creams (0%), the spectrum is filtered away. The scores of the test spectra on the first two PCs are also plotted to have an idea how well the classes can be separated from each other after a particular pre-treatment (Fig. 2).

The differentiation obtained for the different pre-treatment methods can be made on the PC1–PC2 score plot. This differentiation is most clear after DOSC although also without pre-treatment, after first and second derivation, with respectively, a smoothing window of 15 and 21 measuring points and after detrending a distinction between the three concentrations can be made. On the PC score plots it can be seen that sometimes along PC 2, there are two subgroups in each concentration group. The most compact subgroup represents the spectra from the samples with a filling height of 0.5 cm, the more disperse subgroup consists of samples with a filling height of 0.25 cm. This can be explained by the fact that filling a cup with 0.25 cm of cream so that the height is constant, is more difficult than filling it with 0.5 cm. It also shows that filling height has an effect on the spectrum and should therefore be controlled.

#### 4.1.1. k Nearest neighbour method

The first step in modelling is selecting the optimal number of features which has to be used in the model. The features to include in the model (wavelengths or principal components (PC's)) are chosen according to the highest FC. First, a classification model is build using the absorbances at the wavelength (or the scores on the PC) with the highest FC, i.e. with the highest discrimination capacity between the different groups. A leave one out cross validation (LOOCV) is performed on the

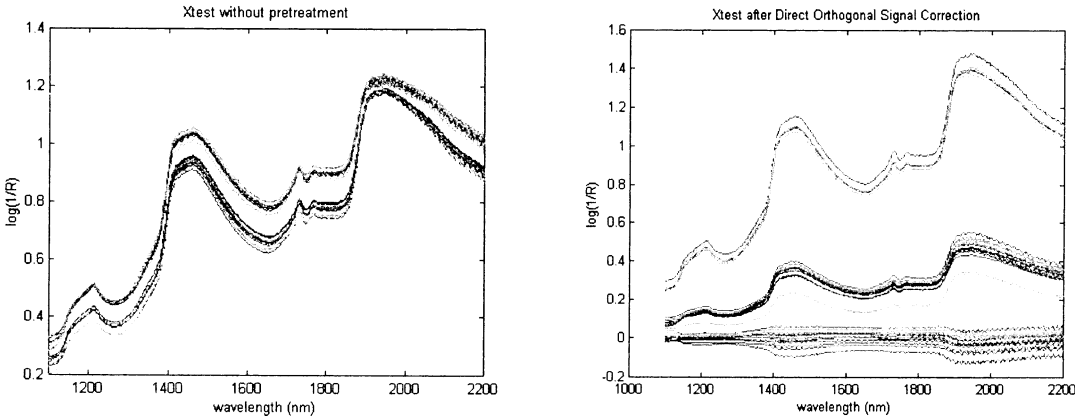


Fig. 1. Spectra of the test set (external measurements): (a) without pre-treatment and (b) after DOSC.

calibration set and the correct classification rate (CCR) is retained. Then the feature with the second highest FC is added to the model and a LOOCV is performed again. This is repeated for the first 25 variables (wavelengths or PCs), ordered according to the highest FC. The model with the highest CCR and the smallest number of features is used as classification model. Table 1 represents the optimal number of features, this means the number of features (wavelengths or PC's) with which the highest correct classification rate (CCR) is reached for the calibration set and the corresponding CCR values for the test set.

Using the Euclidean distance as distance parameter for the k nearest neighbour method gives better results for the CCR than using the correlation coefficient. For the original spectra (without pre-treatment), for the first derived spectra, for the detrend corrected and for the DOSC corrected spectra for the four studied methods using the Euclidean distance, it is possible to make classifications with a single feature and with a CCR of 1. In this case, DOSC pre-treated data give no better results than the original spectra or than some other pre-treatments. This can be explained because of the small amount of variance in the data due to other sources than differences in concentration (same temperature, good filling of the cups,...).

For the SNV pre-treated data, the CCR also has a value of 1 but the classification model has a complexity of two to four features depending on the chosen method (1 NN, 3 NN and wavelengths or PCs as features). In this case, the classification model is of lower complexity using wavelengths as features in the model instead of using PCs. Using the MSC pre-treated data for the classification, two to three features are needed to make the classification. For this data set, the offset correction does not seem to be the best way to pre-treat the spectra. For the 1 NN method, the optimal complexity of the classification model is the same as for the SNV and MSC pre-treated data, but the CCR is worse. For the 3 NN method, using the Euclidean distance, seven PCs are needed to make the classification and the correct classification rate is $<1$.

The results of the kNN method using correlation coefficients as distance parameter, show that the optimal complexity of the classification models is higher than using the ED as distance parameter. Using wavelengths as features, a CCR $=1$ is obtained after SNV (1 NN) and after first and second derivation (with smoothing) but the classification models are complex. In this case, DOSC gives bad results using wavelengths as features, but for the classification using PCs as features, DOSC gives the best CCR results. Using PCs for the classification, generally the optimal complexity of the classification models is better but a CCR equal to 1 for the test set is not obtained.

Comparing the 1 and 3 NN method, generally one can conclude that the 1 NN method gives better CCR values for the same pre-treatment using the CC as distance parameter. Looking to

Table 1
Optimal number of features (o.n.f.) and correct classification rate (CCR) values of the corresponding models for the measurements with the optical fibre after different preprocessing methods

| Selected features | Euclidean distance | | | | Correlation coefficient | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 NN | | 3 NN | | 1 NN | | 3 NN | |
| | wl | PC | wl | PC | wl | PC | wl | PC |
| *No pt.* | | | | | | | | |
| o.n.f. | 1 | 1 | 1 | 1 | 25 | 6 | 24 | 5 |
| CCR | 1 | 1 | 1 | 1 | 0.6190 | 0.9524 | 0.4762 | 0.8333 |
| *SNV* | | | | | | | | |
| o.n.f. | 2 | 3 | 2 | 4 | 19 | 4 | 24 | 4 |
| CCR | 1 | 1 | 1 | 1 | 1 | 0.8810 | 0.9762 | 0.8571 |
| *MSC* | | | | | | | | |
| o.n.f. | 2 | 3 | 2 | 2 | 11 | 6 | 18 | 4 |
| CCR | 1 | 1 | 1 | 0.9762 | 0.9524 | 0.8333 | 0.9524 | 0.8333 |
| *1st der.* | | | | | | | | |
| o.n.f. | 1 | 1 | 1 | 1 | 25 | 9 | 23 | 4 |
| CCR | 1 | 1 | 1 | 1 | 1 | 0.9286 | 1 | 0.8810 |
| *2nd der.* | | | | | | | | |
| o.n.f. | 1 | 1 | 1 | 1 | 13 | 13 | 13 | 13 |
| CCR | 0.9762 | 1 | 0.9762 | 1 | 1 | 0.8571 | 1 | 0.7381 |
| *Offset* | | | | | | | | |
| o.n.f. | 2 | 3 | 2 | 7 | 23 | 4 | 23 | 4 |
| CCR | 0.7619 | 0.9524 | 0.8095 | 0.9524 | 0.7143 | 0.9286 | 0.5476 | 0.7619 |
| *Detrend* | | | | | | | | |
| o.n.f. | 1 | 1 | 1 | 1 | 10 | 6 | 8 | 7 |
| CCR | 1 | 1 | 1 | 1 | 0.8810 | 0.7857 | 0.8095 | 0.7857 |
| *DOSC* | | | | | | | | |
| o.n.f. | 1 | 1 | 1 | 1 | 22 | 4 | 7 | 5 |
| CCR | 1 | 1 | 1 | 1 | 0.7619 | 0.9762 | 0.6905 | 0.9762 |

wl, Wavelength; PC, principal component.

the optimal complexity when the CC is used, no general conclusion can be made about the fact that 1 or 3 NN gives better results.

For this data set it is clear that Euclidean distance is preferable above the correlation coefficient as distance parameter in the kNN method. It is even possible to classify all objects of the test set into the right class using only one feature (wavelength or PC) of the original spectra. Also for the first and second (only with PCs) derived spectra, the detrend corrected and the DOSC corrected spectra, one feature is enough to obtain optimal correct classification rates.

#### 4.1.2. PLS and PCR modelling

A second method to make classifications is using PLS and PCR modelling. The first step in this method is calculating the rmsecv values of the calibration set with models using one to 25 components. The spectra are first column centered. For the PCR models, the order of the PCs, used in the models is chosen according to the correlation coefficient to the concentration. In this study, it is not the aim to predict the exact concentration of the new samples, but a classification of new samples has to be made. Therefore, it is not necessary to look for the lowest rmsecv. A

Table 2
Number of components (# comp.), root mean square error of cross validation values (rmsecv) and correct classification rates (CCR) of the corresponding models for the measurements with the optical fibre after different preprocessing methods

| | No pt. | SNV | MSC | 1st Der. | 2nd Der. | Offset | Detrend | DOSC |
|---|---|---|---|---|---|---|---|---|
| PLS | | | | | | | | |
| # comp | 2 | 3 | 3 | 2 | 2 | 2 | 2 | 1 |
| rmsecv | 0.3720 | 0.2028 | 0.2003 | 0.4264 | 0.4142 | 0.4042 | 0.4540 | 0.0016 |
| *CCR* | 1 | 1 | 1 | 1 | 1 | 1 | 0.9762 | 1 |
| PCR | | | | | | | | |
| # comp | 2 | 3 | 3 | 2 | 2 | 2 | 2 | 1 |
| rmsecv | 0.4187 | 0.3968 | 0.4091 | 0.4306 | 0.4679 | 0.4160 | 0.3978 | 0.0230 |
| *CCR* | 1 | 0.9524 | 0.9524 | 0.9524 | 1 | 1 | 1 | 1 |

cut off value of 0.50 for the rmsecv is used to select the complexity of the model. The least complex model with a rmsecv smaller than 0.50 is used to make a 'rough' prediction of the concentration of the new samples. Then the samples are classified using as rules: a sample with a predicted concentration smaller than 0.5% is classified into the 0% group, a sample with a predicted concentration between 0.5 and 2% belongs to the 1% class and all samples with a predicted concentration higher than 2% are classified into the 3% group. To evaluate the performance of the models, the CCR is calculated for the test set. Table 2 shows the complexities of the models, after the different preprocessing methods, the rmsecv and the CCR values for both PLS and PCR models.

PLS and PCR give the same number of components. For most preprocessing methods, the CCR values of the PLS models are equal or better than the PCR models. The best results for model complexity and for CCR are obtained with the DOSC pre-treated data. A CCR of 1 is obtained with a model using one PLS-component or one PC. Comparing the rmsecv values for the calibration set with the other pre-treatments, DOSC performs very well: with only one component, a rmsecv of 0.0016 is achieved for the PLS model and using the PCR model a rmsecv of 0.0230 is achieved. For all PLS models, except for the detrended spectra, a CCR of 1 is obtained. For the original (untreated), the first and second derived and the offset corrected data, only two PLS components are required. The SNV and the MSC pre-treated data need three PLS components to achieve a CCR of 1.

### 4.2. Internal measuring mode

In this measuring mode, ten cups of each concentration are filled and covered with a microscope glass slide. Five are filled in a smooth way, with a constant filling height. The other five are filled in a bad way, i.e. with a thin layer of cream, with a raw surface or in a non-flat way. Also in this measuring mode, the samples are measured three times a day during 3 subsequent days. The measurements are performed at $20 \pm 1$ °C and also at $30 \pm 1$ °C. These sources of variance (different temperature, bad filling) are introduced to see to what extent a certain preprocessing method is able to filter the spectral variances due to them and to what extent these factors are critical in the subsequent classification procedure. The spectra are taken between 1100 and 2500 nm. After having examined the spectra visually, four spectra are deleted because an error occurred during the measurements. The 536 remaining spectra are divided randomly into a calibration set of 337

Fig. 2. PC 1–PC 2 score plots of external measurements after (a) no pre-treatment; (b) SNV; (c) MSC; (d) first derivation; (e) second derivation; (f) offset correction; (g) detrend correction; (h) direct orthogonal signal correction. * = 0.25 cm filling height; · = 0.50 cm filling height.

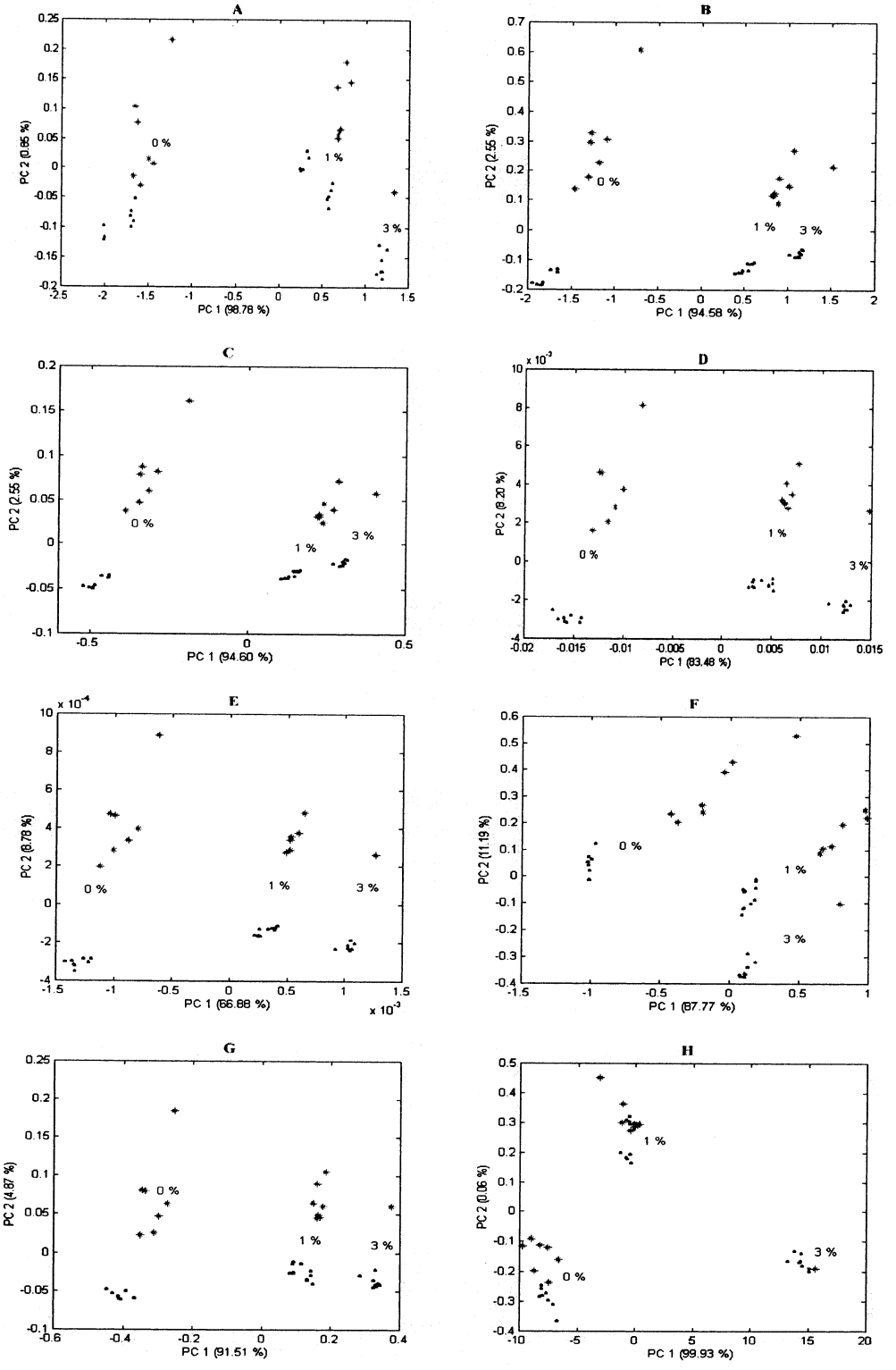
A
PC 2 (0.85 %)
PC 1 (98.78 %)
0 %
1 %
3 %
B
PC 2 (2.55 %)
PC 1 (94.58 %)
0 %
1 %
3 %
C
PC 2 (2.55 %)
PC 1 (94.60 %)
0 %
1 %
3 %
D
x 10^-3
PC 2 (8.20 %)
PC 1 (83.48 %)
0 %
1 %
3 %
E
x 10^-4
PC 2 (6.78 %)
PC 1 (66.88 %)
x 10^-3
0 %
1 %
3 %
F
PC 2 (11.19 %)
PC 1 (87.77 %)
0 %
1 %
3 %
G
PC 2 (4.87 %)
PC 1 (91.51 %)
0 %
1 %
3 %
H
PC 2 (0.06 %)
PC 1 (99.93 %)
0 %
1 %
3 %

Fig. 2

Table 3
Optimal number of features (o.n.f.) and correct classification rate (CCR) values of the corresponding models for the internal measurements after different preprocessing methods

| Selected feature | Euclidean distance | | | | Correlation coefficient | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 NN | | 3 NN | | 1 NN | | 3 NN | |
| | wl | PC | wl | PC | wl | PC | wl | PC |
| *No pt.* | | | | | | | | |
| o.n.f. | 9 | 12 | 7 | 12 | 25 | 14 | 25 | 20 |
| CCR | 0.9146 | 0.9698 | 0.9397 | 0.9447 | 0.6784 | 0.9296 | 0.6432 | 0.8995 |
| *SNV* | | | | | | | | |
| o.n.f. | 9 | 6 | 22 | 7 | 7 | 9 | 24 | 11 |
| CCR | 0.8291 | 0.9849 | 0.9045 | 0.9899 | 0.4271 | 0.9698 | 0.3568 | 0.9497 |
| *MSC* | | | | | | | | |
| o.n.f. | 25 | 6 | 13 | 7 | 7 | 9 | 25 | 10 |
| CCR | 0.8844 | 0.9899 | 0.9095 | 0.9899 | 0.4221 | 0.9698 | 0.3568 | 0.9548 |
| *1st Der.* | | | | | | | | |
| o.n.f. | 15 | 5 | 16 | 5 | 16 | 9 | 16 | 24 |
| CCR | 0.9849 | 0.9950 | 0.9950 | 0.9899 | 0.9950 | 0.9598 | 0.9950 | 0.9497 |
| *2nd Der.* | | | | | | | | |
| o.n.f. | 1 | 4 | 1 | 4 | 5 | 17 | 5 | 11 |
| CCR | 1 | 0.9950 | 1 | 0.9849 | 1 | 0.9598 | 1 | 0.9397 |
| *Offset* | | | | | | | | |
| o.n.f. | 25 | 8 | 20 | 5 | 22 | 8 | 18 | 12 |
| CCR | 0.8693 | 0.9749 | 0.7889 | 0.9799 | 0.9246 | 0.9296 | 0.9397 | 0.8945 |
| *Detrend* | | | | | | | | |
| o.n.f. | 10 | 4 | 6 | 4 | 11 | 11 | 10 | 8 |
| CCR | 0.9598 | 0.9698 | 0.9397 | 0.9296 | 0.8593 | 0.9598 | 0.8040 | 0.9347 |
| *DOSC* | | | | | | | | |
| o.n.f. | 1 | 1 | 1 | 1 | 24 | 25 | 25 | 20 |
| CCR | 1 | 1 | 1 | 1 | 0.8844 | 0.9950 | 0.8442 | 0.9899 |

Table 4
Number of components (# comp.), root mean square error of cross validation values (rmsecv) and correct classification rates (CCR) of the corresponding models for the internal measurements after different preprocessing methods

| | No pt. | SNV | MSC | 1st Der. | 2nd Der. | Offset | Detrend | DOSC |
|---|---|---|---|---|---|---|---|---|
| PLS | | | | | | | | |
| # comp | 5 | 4 | 4 | 3 | 3 | 5 | 5 | 1 |
| rmsecv | 0.4869 | 0.4758 | 0.4800 | 0.3378 | 0.2199 | 0.4553 | 0.1994 | 0.0024 |
| *CCR* | 0.9095 | 0.9347 | 0.9347 | 0.9950 | 1 | 0.9598 | 1 | 1 |
| PCR | | | | | | | | |
| # comp | 3 | 4 | 4 | 3 | 3 | 4 | 3 | 1 |
| rmsecv | 0.4932 | 0.4835 | 0.4815 | 0.4400 | 0.2601 | 0.4520 | 0.4712 | 0.0024 |
| *CCR* | 0.9548 | 0.9045 | 0.9095 | 0.9447 | 0.9950 | 0.9698 | 0.9146 | 1 |

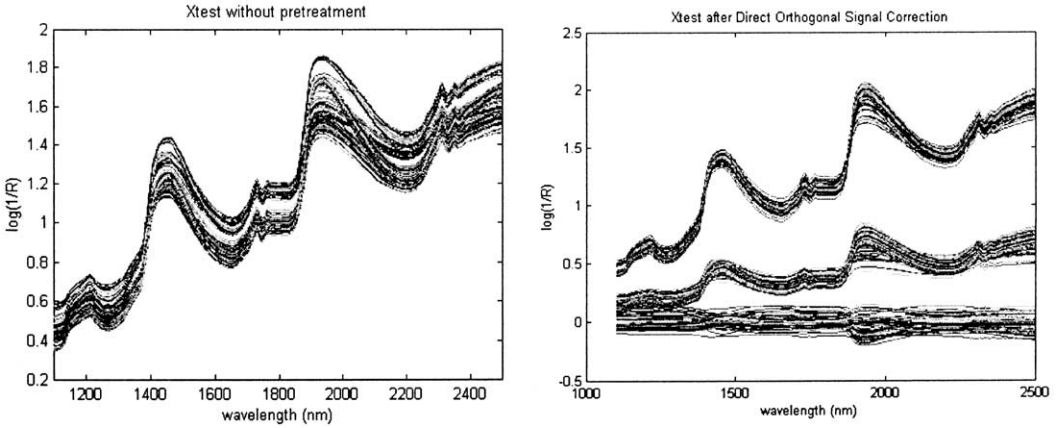


Fig. 3. Spectra of the test set (internal measurements): (a) without pre-treatment and (b) after DOSC.

spectra and a test set of 199 spectra. The spectra of the test set are shown in Fig. 3. This figure also shows the DOSC pre-treated spectra.

The PC1–PC2 scores of the test spectra, before and after every pre-treatment are plotted to have an idea how well the classes can be separated from each other after a particular pre-treatment (Fig. 4). After some pre-treatments, two groups are present for each concentration on the PC1–PC2 score plot, for instance for the DOSC corrected data. The smallest group represents the measurements of the good filled cups while the variation on the spectra of the bad filled cups is bigger. A clear separation according to the temperature cannot be made on this PC1–PC2 plot because the variation due to the bad filling of the cups is much higher than the variation due to the temperature differences.

#### 4.2.1. k Nearest neighbour method

Table 3 shows the optimal number of factors (o.n.f.) and the CCR results for each pre-treatment. Compared with the first data set, the differences between the results from the Euclidean distance and the correlation coefficient as distance parameter, are not so clear although it can be seen that for almost every pre-treatment, the performance is better for the ED. With the ED a CCR = 1 is obtained with the DOSC corrected spectra (for the four methods) and with the second derived spectra (only when wavelengths are used as features). For these pre-treated data, one feature ($\lambda = 2284$ nm or the first PC for the DOSC corrected data, $\lambda = 1686$ nm for the second derived spectra) is enough to classify all new samples into the right class.

A CCR = 0.9950 is obtained for the first derived spectra using five principal components and 1 nearest neighbour and also using 16 wavelengths and the 3 NN method. The second derived spectra also give a CCR = 0.9950 when the 1 NN method with four PCs used. Other methods giving CCR results higher than 0.98 are: SNV and MSC pre-treated data using PCs as features, both for the 1 and 3 NN method; first derived spectra using wavelengths for the 1 NN and using PCs for the 3 NN method and the second derived spectra using the 3 NN method with PCs as features. CCR results $< 0.90$ are obtained with the SNV and MSC corrected data using the 1 NN method combined with wavelengths as features. Offset correction using wavelengths as features is the classification method that gives the worst results because a high number of features is required and

Fig. 4. PC1–PC2 score plots of internal measurements after different pre-treatments: (a) no pre-treatment; (b) SNV; (c) MSC; (d) first derivation; (e) second derivation; (f) offset correction; (g) detrend correction; (h) direct orthogonal signal correction. * = 0%; + = 1%; · = 3%.

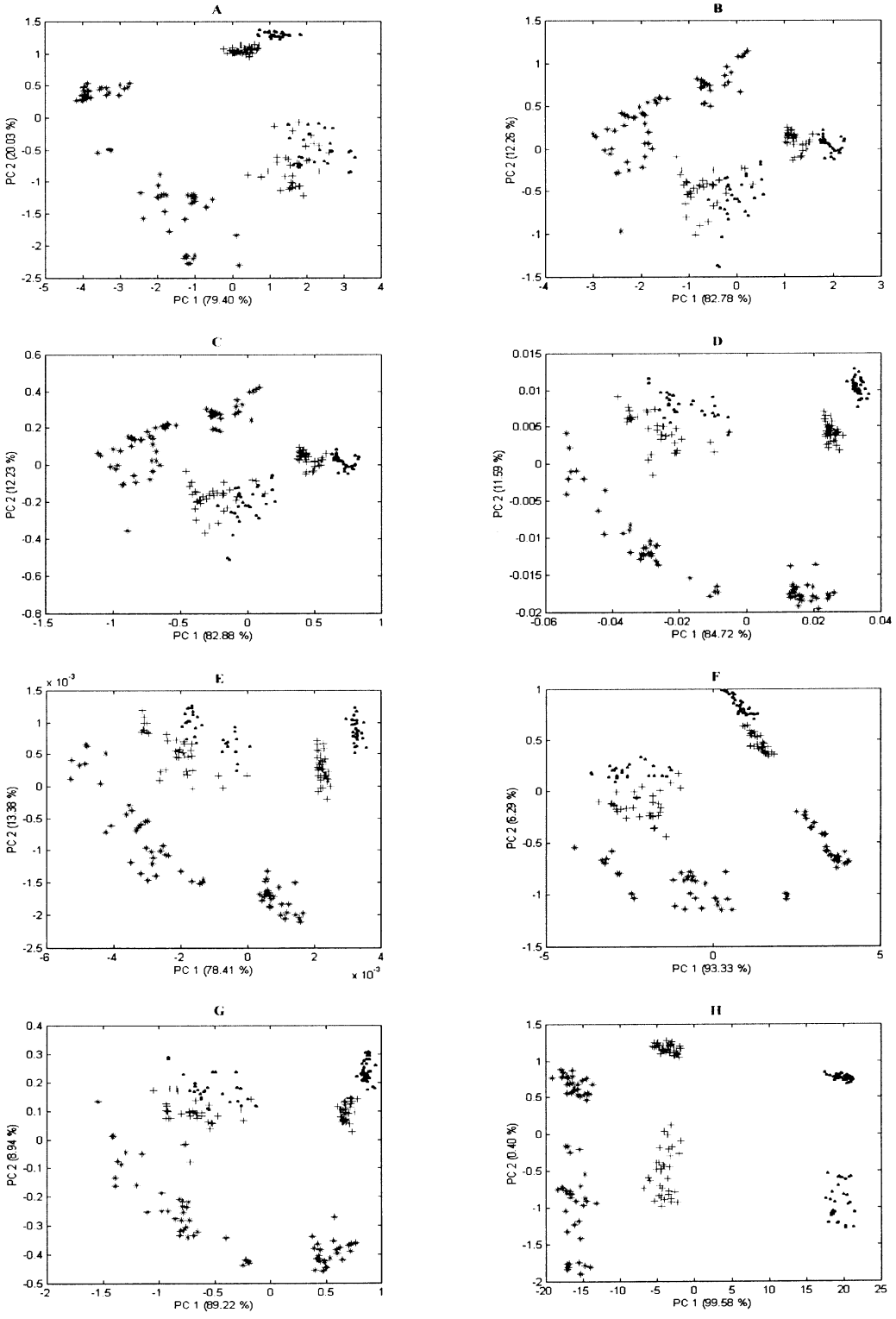


Fig. 4 (*Continued*)

the CCR results are even $<0.80$ for the 3 NN method.

With the correlation coefficient as distance parameter, poor CCR results are obtained for the untreated spectra and for the SNV and MSC corrected spectra when wavelengths are used. Also the offset corrected data using PCs (3 NN) and detrended and DOSC corrected data, when wavelengths are used as features give no good CCR results. Good results are obtained with the first derived spectra, using wavelengths as features and also with the DOSC corrected data using PCs as features, although a high number of features are required. A CCR = 1 is obtained with the second derived spectra, using wavelengths as features.

#### 4.2.2. PLS and PCR modelling

Before the modelling, all spectra are column centered. Evaluating the results of the PLS and the PCR modelling, Table 4 shows that the number of components used to obtain a rmsecv value $<0.50$, is smaller for the untreated data, for the offset corrected data and for the detrended data using PCR models. For the other pre-treated data, the same number of components is required for PLS as for PCR. For the untreated spectra and for the offset corrected spectra, the CCR values are better (higher) using PCR as modelling technique. For all other pre-treated spectra, PLS modelling gives better CCR values. With PLS a CCR = 1 is obtained after second derivation, detrending and DOSC. The number of PLS components required is 3, 5 and 1, respectively. When using PLS as modelling technique, the original spectra and SNV and MSC corrected spectra give no good CCR results. Here again, DOSC gives the best results, because with only one PLS component, an optimal classification of the test samples is obtained.

Using PCR as modelling technique, the best result is also obtained after DOSC. The one PC model gives a CCR = 1. Second derived spectra also give good CCR results (0.9950), but for this pre-treatment, three principal components are required in the model. SNV and MSC pre-treated spectra give the lowest CCR.

In conclusion, it can be said that PLS gives better results as modelling technique for these data and that DOSC is the best preprocessing technique because it gives CCR = 1 with the simplest model (only one component).

## 5. Conclusions

The second data set is more heterogeneous (bad filling, two different temperatures), so that the effect of the pre-treatments is clearest on this data set. It can be observed that DOSC pre-treated data show the clearest separation between spectra from the different classes. For all methods using the Euclidean distance as distance parameter, a perfect classification is obtained after DOSC correction, with only one feature. Also when a PLS or a PCR model is used to predict the concentration, only one feature is required. Second derivation also gives good CCR results both for the kNN method and for the PLS or PCR models. In this classification application, DOSC is shown to be an effective technique to remove information from the spectra not related to the concentration of active compound of the cream. These pre-treated spectra, combined with kNN method, using the Euclidean distance as distance parameter, always give a correct classification using only one feature (PC or wavelength). Also, PLS or PCR modelling after the DOSC correction yield classification results that are excellent with a model using only one component.

## References

[1] A. Candolfi, W. Wu, D.L. Massart, S. Heuerding, J. Pharm. Biomed. Anal. 16 (1998) 1329–1347.
[2] M.A. Dempster, J.A. Jones, I.R. Last, B.F. MacDonald, K.A. Prebble, J. Pharm. Biomed. Anal. 11 (1993) 1087–1092.
[3] B.F. MacDonald, K.A. Prebble, J. Pharm. Biomed. Anal. 11 (1993) 1077–1085.
[4] P. Dubois, J.-R. Martinez, P. Levillain, Analyst 112 (1987) 1675–1679.
[5] E.W. Ciurczak, T.A. Maldacker, Spectroscopy 1 (1986) 36–39.
[6] P. Corti, E. Dreassi, G. Corbini, L. Montecchi, J. Paggi, Analysis 18 (1990) 117–121.
[7] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of

Chemometrics and Qualimetrics: Part B, Elsevier Science, Amsterdam, 1998, pp. 223–225.
[8] J.A. Westerhuis, S. de Jong, A.K. Smilde, Chemom. Intell. Lab. Sys. 56 (2001) 13–25.
[9] S. Wold, H. Antti, F. Lindgren, J. Ohman, Chemom. Intell. Lab. Sys. 44 (1998) 175–185.
[10] J. Sjöblom, O. Svensson, M. Josefson, H. Kullberg, S. Wold, Chemom. Intell. Lab. Sys. 44 (1998) 229–244.
[11] B.M. Wise, N.B. Gallagher, 1999. http://www.eigenvector.com/MATLAB/OSC.html
[12] T. Fearn, Chemom. Intell. Lab. Sys. 50 (2000) 47–52.
[13] C.A. Andersson, Chemom. Intell. Lab. Sys. 47 (1999) 51–63.
[14] M.S. Dhanoa, S.J. Lister, R. Sanderson, R.J. Barnes, J. Near Infrared Spectrosc. 2 (1994) 43–47.
[15] G. Sinnaeve, P. Dardenne, R. Agneessens, J. Near Infrared Spectrosc. 2 (1994) 163–175.
[16] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spectrosc. 43 (1989) 772–777.
[17] T. Næs, T. Isaksson, NIR News 5 (1994) 4–5.
[18] B.G. Osborne, T. Fearn, P.H. Hindle, Practical NIR Spectroscopy, 2nd ed, Longman Scientific and Technical, UK, 1993, pp. 42–43.
[19] P.A. Gorry, Anal. Chem. 62 (1990) 570–573.
[20] A. Candolfi, R. De Maesschalck, D. Jouan-Rimbaud, P.A. Hailey, D.L. Massart, J. Pharm. Biomed. Anal. 21 (1999) 115–132.
[21] W. Wu, B. Walczak, D.L. Massart, K.A. Prebble, I.R. Last, Anal. Chim. Acta 315 (1995) 243–255.
[22] R.G. Brereton, Chemometrics, Applications of Mathematics and Statistics to Laboratory Systems, Ellis Horwood Ltd, Chichester, 1990.
[23] R. De Maesschalk et al., PCR tutorial. http://vub.vub.ac.be/ ~ fabi/calibration/multi/pcr/pages/framemenu.html