ELSEVIER

Journal of Pharmaceutical and Biomedical Analysis
16 (1998) 1329–1347

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

Comparison of classification approaches applied to NIR-spectra of clinical study lots

A. Candolfi [a], W. Wu [a], D.L. Massart [a,*], S. Heuerding [b]

[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*

Received 14 April 1997; received in revised form 29 May 1997; accepted 29 May 1997

Abstract

NIR-spectroscopy combined with pattern recognition approaches is applied to classify samples of clinical study lots in the pharmaceutical industry. The performance of linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and K-nearest neighbour (KNN) method is evaluated on a tablet data set and a capsule data set. To establish a classification model a strategy is followed, which is described in this work. Frequently, in the pharmaceutical industry, several batches of the same clinical study lot are produced. We tested whether it is possible to merge several batches in one class for modelling or, instead, whether it is necessary to model each batch individually.

Keywords: NIR-spectroscopy; Clinical study lots; Batches; Linear discriminant analysis; Quadratic discriminant analysis; K-nearest neighbour method

1. Introduction

The identification of clinical study lots in the pharmaceutical industry is a time-consuming procedure, which is usually done by high performance liquid chromatography (HPLC) or thin-layer chromatography (TLC). Such clinical study lots are designed to evaluate the effect of a new drug compared to the effect of a placebo formulation or to a clinical comparator, which is already on the market. To make the results of a clinical study objective, all samples distributed to the participants of a study look the same, regardless of their chemical nature. It is crucial that a severe control is performed to verify that the drugs are correctly identified and thus guarantee a successful clinical study. One requires a fast and reliable method to do so. A satisfactory analytical method, which is less time-consuming than HPLC or TLC, seems to be near infrared (NIR) spectroscopy combined with pattern recognition methods [1–5].

The identification of clinical study lots can be performed by a method with discriminating character or by a class-modelling technique [6]. If all possible classes are known and one would like to discriminate between those given classes, discriminating methods can be used. New objects will be

* Corresponding author.

0731-7085/98/$19.00
PII S0731-7085(97)00154-4

assigned to the most similar class. The task then is to find classification rules which define optimal boundaries to separate the classes. Among techniques with discriminating character, parametric methods (e.g. linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA)), which take into account information about the data distribution and non-parametric methods (e.g. K-nearest neighbour (KNN)) which have no assumptions about the data distribution, can be distinguished. Such discriminating methods are suitable for the above described situation, i.e. all new objects must belong to one of the known classes. A shortcoming of such methods, however, is that the detection of a sample not belonging to any category becomes impossible. Appropriate methods for this purpose are class-modelling techniques, for instance, soft independent modelling of class analogy (SIMCA). In SIMCA, one checks whether new objects fall into any of the modelled classes. This is done with an F-test on a certain level of significance for each class. Usually the significance level is 0.05, which means that for each class 5% of the objects will be rejected. In practice, when clinical study lots are identified, one takes randomly a few samples and analyses them. However, since several samples have to be identified, the probability of rejecting objects will then be increased and this is often not acceptable in routine analysis.

In the presented work we focus on discriminating methods, where each object has to be allocated to the most similar class. We compare LDA and QDA on the one hand as parametric methods and KNN on the other hand as a non-parametric method. The performance of these methods will be discussed on two NIR data sets, one tablet data set, containing four concentrations, but in total nine classes and one capsule data set, containing six concentrations and total eleven classes. There are more classes present than concentrations in each data set, because several batches exist of the same tablet or capsule concentration. The behaviour of the different discriminating approaches are compared when each batch is modelled, or instead, all batches with the same concentration are considered as forming one class.

2. Theory

2.1. Strategy for the development of a classification method

A classification method development starts with a data investigation part, where one wants to study the structure and quality of the data. If necessary, transformation of the data with a suitable transformation or signal processing method can then be performed. Diagnostics are helpful to reveal inhomogeneities in the data. If the amount of data is large enough, the data can be divided into training and test sets. The training set is used to build the model. Within the training set, leave-one-out cross-validation (LOO-CV) is carried out to determine the number of features to build the model. The prediction of the samples of the independent test set is considered as validation.

The different steps of this strategy and the applied methods are described in the following sections.

2.2. Data investigation

Data investigation is performed on each class separately and on all data together. Plots of the spectra, of the Fisher criterion (FC) (see Section 2.6.1.1) and PC score plots are studied to obtain an idea about the quality of the data and its problems. These simple tools reveal, for instance, if pre-treatment is necessary and show the existence of inhomogeneities.

2.3. Pre-treatment methods

Raw NIR-spectra are often affected by noise or by uncontrolled variations of the baseline, due to instrument properties and influences of the measurement conditions (particle size of powdered samples, temperature, humidity). To correct for this, different pre-processing methods can be applied. Prior to any data pre-treatment a trial with raw data should be carried out. This helps to estimate the quality of the data and gives an idea of what kind of problems can occur [7].

2.3.1. Standard normal variate (SNV)

SNV removes the multiplicative interferences of scatter and particle size. To eliminate slope variations on individual spectrum sample basis, each spectrum is transformed independently using the following equation:

$$X_{ij\mathrm{SNV}} = (x_{ij} - m_i) / \sqrt{\frac{\sum (x_{ij} - m_i)^2}{p - 1}} \quad (1)$$

where $X_{ij\mathrm{SNV}}$ is the SNV transformed x value of the ith object at the jth wavelength, x is the log $(1/R)$ value for the wavelengths, m_i is the row mean of the log (1/R) values and p is the number of variables in the spectrum. [8].

2.3.2. First derivative

The goal of derivatives is to separate overlapping peaks and to remove baseline shift and slope changes. The particle size effect is partly reduced. A smoothing procedure is usually carried out beforehand, to avoid increase of noise. First derivative spectral transformation, with data smoothing, was carried out here as described by Gorry [9]. A window width of 17 variables is used.

2.4. Diagnostics

2.4.1. Single Grubb's test on Rao's statistics

Outlying samples in a population can be detected by the mean of the single outlier Grubb's test proposed by Grubbs and Beck [10,11], where one computes the largest absolute value of the normalised deviation z and compares it to a tabulated critical z value.

$$z = (x_i - \bar{x})/s \quad (2)$$

where x_i is the measured value of a suspected outlier, $\bar{x}$ the class mean and s the class S.D.

Since the Grubb's test is a univariate test, it was modified to apply it to latent variables (PCs) obtained by singular value decomposition (SVD) on centered data. The modification consists of using the sum of squared scores as input data. The statistics was described by Mertens et al. [12] and is named Rao's statistics (D^2).

$$D_i^2 = \sum_{j=1}^{p} t_{ij}^2 \quad (3)$$

The sum of the squared scores (D^2) is calculated for each object i. This includes the scores (t) from the jth to the pth PC, where j ranges from 1 to p, with p equal to the number of PCs. Here for the outlier detection on each class separately, the test is applied several times, including all PCs ($j = 1$), including all PCs without the first ($j = 2$), the first and second ($j = 3$), the first, second and third PC ($j = 4$). One is actually computing the residuals towards a model based on the first PCs, the ones which are left out for the computation of D^2. A high value means that the object is extreme since it does not fit the PC-model valid for most other objects. An additional modification of the Grubb's test on Rao's statistics was proposed by Centner et al. [13], namely to compare x_i to 0 instead of comparing it to $\bar{x}$, which leads to the following formula:

$$z = x_i / \left(\sum_{i=1}^{n} x_i^2/(n-1) \right)^{1/2} \quad \text{for } i = 1,\ldots, n. \quad (4)$$

where x_i is the suspected outlier and n the number of objects in the class.

2.5. Dataset division

The Kennard and Stone algorithm [14,15] is applied for dividing the data into training and test sets. The training set is used for building and optimising the model, the test set enables an external validation of the model. The aim of the Kennard and Stone algorithm is to select training set samples uniformly spaced over the whole object space. It is a two step procedure. In the first step the two objects, which are the farthest apart from each other, are selected. In the second step the object, being the farthest away from the first two, is selected. The Euclidean distance is used as distance criterion. Step two is repeated until the decided number of objects is selected.

2.6. Feature selection and modelling

2.6.1. Feature selection

Three different feature selection methods, namely a univariate method using the FC, Fourier

transform (FT) and principal component analysis (PCA), are applied. Each of the methods transforms the original variables to new features, respectively the FC for each variable, Fourier coefficients and latent variables. The optimal amount of original or transformed variables is selected by cross-validation.

2.6.1.1. Univariate method—Fisher criterion (FC) (univar.). The FC describes the ratio of between class variance/within class variance. This criterion is helpful to decide which variables have an important discriminating power. The variables ranked from the highest to the lowest parameter e_i will therefore be selected top-down using the following formula [16]:

$$e_i = \sum_{j=1}^{k} n_j(\bar{y}_{ji} - \bar{y}_{.i})^2 / \sum_{j=1}^{k} (n_j - 1)s_{ji}^2 \quad (5)$$

where $j = 1,2,3,...,k$ is the number of classes, n_j the number of objects in class j, $\bar{y}_{ji}$ the mean absorbance of the objects belonging to class j at the ith wavelength, $\bar{y}_{.i}$ the mean absorbance of the objects belonging to all classes at the ith wavelength and s_{ji} the S.D. of the absorbance of the objects belonging to class j at the ith wavelength. Based on parameter e_i the corresponding features are selected and used as input to the classifier.

2.6.1.2. Fourier transform (FT). FT is mainly a signal processing method [17], where each spectrum is transformed separately from the wavelength domain to the frequency domain. Hereby it is used as feature reduction method, where the Fourier coefficients are selected top-down from the second Fourier coefficient on, since the first one describes the mean of the spectra.

2.6.1.3. Principal component analysis (PCA). The principal components (PCs) are chosen by applying the FC. The latent variables with the highest FC are selected.

2.6.2. Classifiers

2.6.2.1. Quadratic discriminant analysis (QDA), linear discriminant analysis (LDA). QDA and LDA are supervised pattern recognition techniques with discriminating character. The goal is to classify a new object x_i into one of K given classes. The following classification score is applied [18]:

$$cf(x_i) = (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \ln|\Sigma_k| - \ln \pi_k \quad (6)$$

where Σ_k is the class covariance matrix of class k, μ_k the mean vector of class k and π_k the prior probability of class k. The new object will belong to the class for which it has the lowest classification score. This quadratic classification rule leads to QDA. For applying QDA, the data should fulfil the following assumptions: the data must be normally distributed and the number of variables must not be larger than the number of objects in the smallest class.

LDA is a simplification of QDA. In LDA one assumes that the covariance matrices of all classes are equal. This enables to pool them. One single term for the individual covariance matrices is obtained for the classes, which can be inserted to Eq. (6). If the prior class probability of all classes is equal, and constants are ignored, the above equation reduces to the Mahalanobis distance. In LDA the class borders are linear.

2.6.2.2. K-nearest neighbour method (KNN). KNN is a simple non-parametric method, where the distance between an unknown object and all objects of the training set (all classes) is computed. The unknown is classified into the class with the object to which the distance is the smallest. Two variants of KNN are applied, 1NN and 3NN. In 3NN, one computes the three nearest neighbours and classifies the unknown object according to the majority rule to the class with the maximum of nearest neighbours. The simplicity of this method leads to some problems. To make the classification reasonable the classes must consist of similar amounts of objects, otherwise modifications in the method have to be considered [17].

Different distance parameters can be used for KNN. In this case studies of the Euclidean distance and the correlation coefficient are used as classification criteria. The disadvantage of the Euclidean distance is that it does not take into account the variance–covariance of the data. It is

possible to overcome this problem by using the Mahalanobis distance, which is however more difficult to compute. Since the Euclidean distance was originally proposed for KNN we focus on this criterion.

3. Experimental

3.1. Data

3.1.1. Tablet data set

This data set contains spectra of tablets in four dosage strengths. Some of the dosage strengths are available as several batches. In total there are nine classes present, one class consists of 27 samples, one of 28 samples, four classes of 29 samples and three classes of 30 samples. This leads to 261 objects in this data set. The tablets contain an active drug and seven excipients. The main excipients are Crospovidone and lactose. Lactose compensates for the varying amount of the active drug.

Dosage	Concentration of the active in %	Number of objects
1. 2 mg	2.22	27
2. 2 mg	2.22	28
3. 6 mg	4.8	29
4. 6 mg	4.8	29
5. 12 mg	6.67	30
6. Placebo I	—	30
7. Placebo II	—	30
8. Placebo III	—	29
9. Placebo IV	—	29

3.1.2. Capsule data set

This data set contains spectra of hard gelatine capsules in six dosage strengths. Some of the concentrations are again available as several batches. In total there are 11 classes present, two classes consists of 27 samples, two of 28 samples, four classes of 29 samples and three classes of 30 samples. This leads to 316 objects in this data set. The filling of the capsules consist of the active drug and four excipients, the main one being cellulose.

Dosage	Concentration of the active in % with shell	Number of objects
1. Placebo	—	30
2. 0.5 mg	0.225	30
3. 1 mg	0.450	27
4. 1 mg	0.450	27
5. 1.5 mg	0.676	28
6. 1.5 mg	0.676	29
7. 1.5 mg	0.676	29
8. 3 mg	1.351	29
9. 3 mg	1.351	29
10. 3 mg	1.351	28
11. 6 mg	2.703	30

All data were collected with a FT-NIR instrument IFS28/NIR from Bruker connected with an optical fibre. The samples belonging to one class were measured in one run. The measurements of the individual classes of the two data sets were randomly carried out within 1 week.

The NIR-spectra, measured in the reflectance mode, are transformed to absorbance as $A = \log(1/R)$. R stands for the diffuse reflectance measurement of the sample versus that of a Spectralon background. The spectra were obtained over the range of 10 000–4000 cm^{-1} (1000–2500 nm), leading to 778 variables. Per measurement ten scans were collected, the resulting spectrum is the average spectrum of these scans. For convenience the wavenumber is expressed by its index in the data matrix.

3.2. Computer programs

All the computations were carried out with a Classification Toolbox which was designed in our laboratory. This Toolbox is written with Matlab for Windows, Version 4.0 as well as all other procedures which were necessary for carrying out the work.

4. Results and discussion

4.1. Tablet data set

4.1.1. Modelling each class

This data set contains tablets from four different concentrations, but in total from nine classes. This means that for some of the tablet concentrations several batches are present, namely for the 2 mg tablets, two batches, for the 6 mg tablets, two batches, for the 12 mg tablets, one batch and for the placebo tablets, four batches. In the first step of this research each class is modelled separately, i.e. a model is built with nine classes.

To eliminate edge effects 15 variables are discarded respectively at the beginning and the end of the spectra. The resulting spectra have a dimension of 748 variables. Simple display methods are studied to extract prior information from the data. Fig. 1a,b,c shows the corresponding FC per variable for each type of pre-processed data.

For the original data (Fig. 1a) the spectral regions with high discriminating ability between classes are situated between the variables 380 and 490 and the variables 680 and 720. One can observe an additional discriminating spectral region, between the variables 100 and 170, for the SNV pre-treated data (Fig. 1b). The two previous ones remain around at the same variable position. The magnitude of the selective spectral bands is increased by a factor 10–25 for the SNV data, which is due to the decreased within class variance after pre-processing compared to the original data. For the first derivative data (Fig. 1c) the discriminating spectral regions are lying between the variables 100 and 120, 470 and 500, 690 and 700. For this kind of data the selective spectral bands are less broad, their magnitude is comparable to the ones for the SNV data, except for the first spectral band, which is increased. The three main selective spectral regions are representative for the active drug in the tablet. The spectrum of the active drug has high absorbance values at the variables which indicate a high FC. Therefore these spectral regions show the difference of concentration of the active drug in the tablets. PCA, as a display method, is used to estimate the difficulty of the classification problem for this

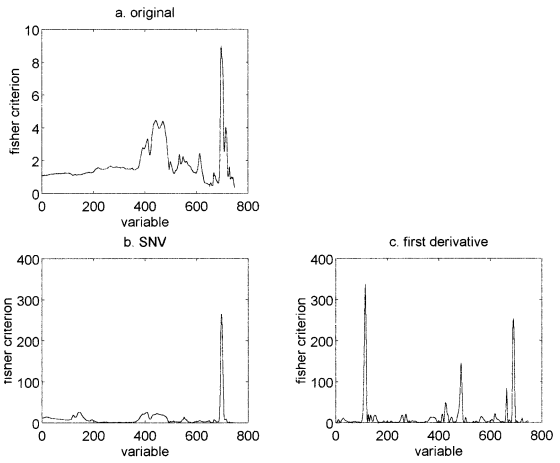

Fig. 1. Fisher criterion for the tablet data (nine classes) obtained from: (a) original; (b) SNV; and (c) first derivative data.

data set. The PC1-PC2 score plots are presented in Fig. 2a for the original data, in Fig. 2b for the SNV data and in Fig. 2c for the first derivative data. Since the amount of data is large, plots containing all scores are unclear, especially because of the overlapped labels. Therefore, only each second object is displayed on the figures.

The score numbers stand for the different classes. The tablets containing 2 mg of the active drug are indicated by the scores 1 and 2, the tablets containing 6 mg of the active drug are indicated by the scores 3 and 4, the tablets containing 12 mg of the active drug are indicated by the scores 5 and the placebo tablets by the scores 6, 7, 8 and 9. One can see on the score plot of the original data, that the 4 concentrations are separated along both PC1 and PC2. After pre-processing the data with SNV, the main variance of the original data which is baseline shift and which was described by PC1 is removed and the separation of the different concentrations is now shown along PC1, whereby PC2 separates somewhat the batches of the same concentration. An even more evident separation of the concentrations along PC1 is obtained on the score plot of the first derivative data.

An outlier detection procedure is carefully applied, investigating each class separately of the

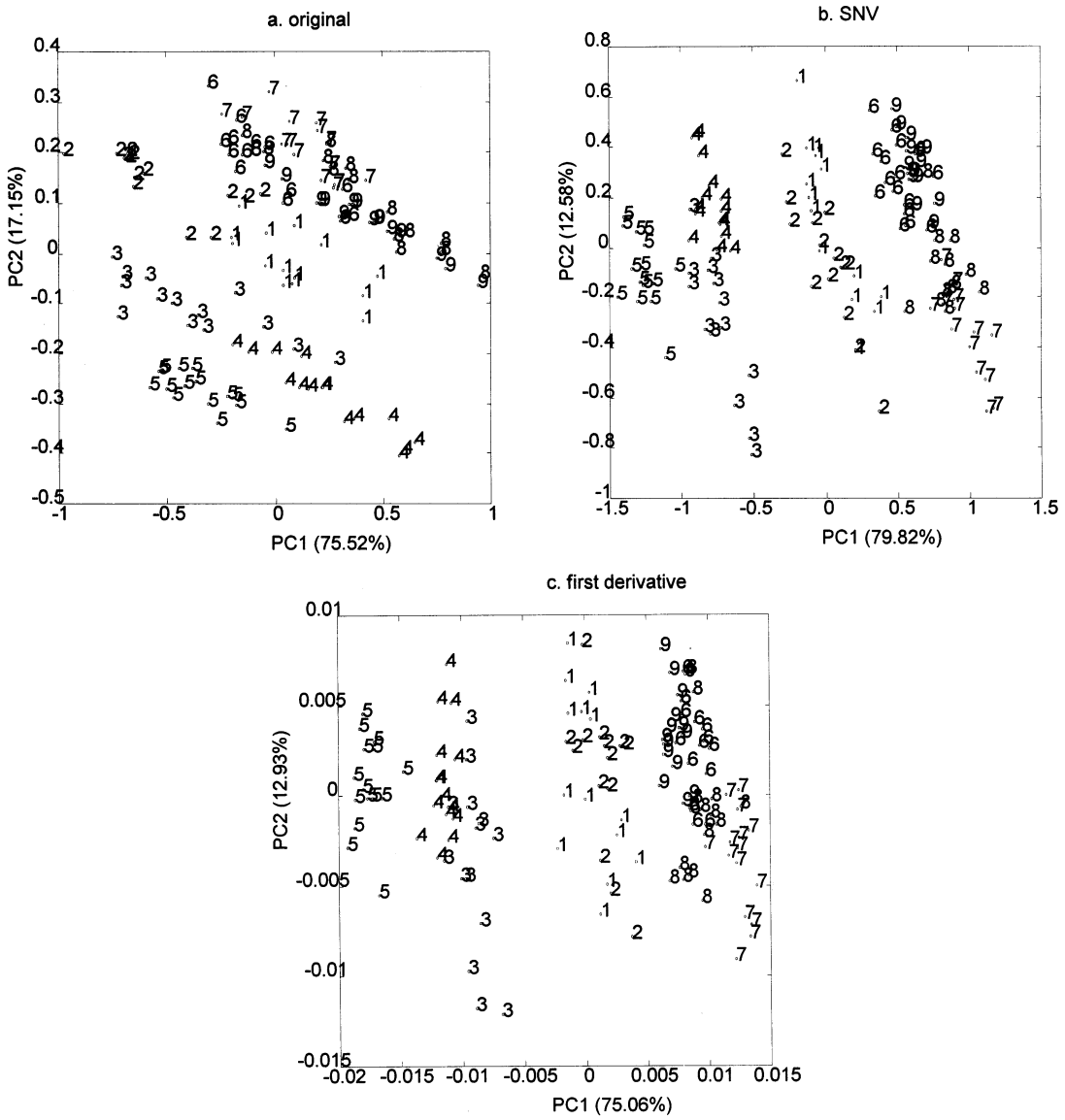

Fig. 2. PC1 vs. PC2 score plot for the tablet data (nine classes) obtained from: (a) original; (b) SNV; and (c) first derivative data.

raw data and after SNV and first derivative transformation of the data. For this purpose we used the single Grubb's test on Rao's statistics at $\alpha = 5\%$. A few objects are detected to be outliers, but since we did not find an assignable cause, we decided to keep all objects, since they may represent the normal variability that we can expect in the different classes.

The data of all classes are separated into training and test sets using the Kennard and Stone algorithm. For each class 20 objects are selected to belong to the training set, the remaining ones are kept as test set. Thus constructed, the training set contains 180 spectra (nine classes, 20 objects per class) and the test set 81 spectra (3×10, 4×9, 1×8 and 1×7

Table 1
Classification of tablets (nine classes)

			Original			SNV			First derivative		
			Nb of variables	Success rate (training)	Success rate (test)	Nb of variables	Success rate (training)	Success rate (test)	Nb of variables	Success rate (training)	Success rate (test)
LDA		Univar.	15	0.9833	0.9753	18	0.9944	0.9877	19	1	1
		FT	21	1	1	19	1	1	25	0.9778	0.9877
		PCA	20	1	1	18	1	1	18	1	1
QDA		Univar.	6	0.8	0.7901	7	0.8111	0.8148	7	0.8833	0.8519
		FT	10	0.9444	0.9383	12	0.9611	0.8889	12	0.9389	0.9506
		PCA	12	0.9167	0.9506	11	0.9556	1	6	0.9278	0.9506
1NN	Euclidean distance	Univar.	17	0.7389	0.7901	24	0.6667	0.8272	7	0.7389	0.7901
		FT	16	0.8389	0.9136	13	0.9222	0.9753	15	0.8278	0.9506
		PCA	25	0.8056	0.9506	11	0.9222	0.9877	14	0.9111	0.9506
1NN	Correlation coeff.	Univar.	19	0.8278	0.8642	25	0.7611	0.8642	25	0.8056	0.8642
		FT	24	0.9389	0.963	25	0.9167	1	7	0.8722	0.9506
		PCA	19	0.5778	0.7778	11	0.6833	0.9136	14	0.6778	0.8519
3NN	Euclidean distance	Univar.	21	0.6944	0.7407	24	0.6389	0.8272	7	0.7778	0.6543
		FT	19	0.7889	0.9383	16	0.9	0.9877	25	0.8333	0.963
		PCA	9	0.7611	0.9012	11	0.9111	0.9877	16	0.8889	0.9753
3NN	Correlation coeff.	Univar.	21	0.8278	0.8889	23	0.7278	0.8765	18	0.7889	0.8765
		FT	18	0.9167	0.963	11	0.9111	0.9383	21	0.8722	0.9753
		PCA	25	0.5222	0.7407	15	0.6944	0.8395	13	0.6778	0.8045

objects). This procedure is repeated for each type of pre-processed data.

In the modelling stage we compare the performance of the different classifiers combined with the three feature selection methods (univar., FT, PCA). Within the training set LOO-CV is carried out to optimise the model, this means to find the number of variables with which one obtains the best classification rate. It was defined beforehand that the maximum amount of selected features is 25. The results are presented in Table 1. In the table the findings for the amount of selected features, the success rates obtained for the training set with cross-validation and the success rates for the classification of the independent test samples with the model are summarised. Success rate 1 indicates that 100% of the objects are correctly classified.

The data analysis is carried out on original data, SNV and first derivative pre-processed data. Already with the original data it is possible to obtain successful classification with one of the methods. The results achieved with first derivative and SNV data are somewhat better than the results achieved with original data. The selected pre-processing methods therefore seem to be suitable for correcting undesirable effects of these tablet spectra.

The success rates for the training set are obtained by LOO-CV for the optimised models. The results for the test set are achieved by predicting the test set objects with the optimal models. In this work we want to find the best classification model for the data set. Therefore it is necessary to include all possible variation of the data in the training set, which is obtained by selecting the objects with the Kennard and Stone algorithm. As a consequence, the classification rates of the test set are slightly better than those of the training set, because for each class most objects of the test set are situated somewhat closer to the class centroid of the respectively class. For validation, this means that the test set results are somewhat too optimistic.

The performance of the classifiers can be compared. The data assumptions for LDA are rather strict (normally distributed data, equal variance–covariance matrix for all classes) but less so for QDA (normally distributed data, variance–covariance matrix can differ for the classes). Non-parametric methods have no assumptions about the data. In general parametric methods are more powerful as long as the assumptions are not severely violated. Clearly, better results are obtained with the parametric methods for this data set. Among the parametric methods LDA performs better than QDA. This is an indication that the assumptions for LDA are sufficiently well fulfilled. The performance of QDA is somewhat worse because more parameters have to be estimated and therefore more objects per class are needed. KNN is applied in the form of 1NN and 3NN. In 3NN the objects are classified according to the majority rule. If an object is classified an equal number of times in several classes, then it will not be classified at all. KNN is a method which does not consider the shape of the classes and defines similarities to a class only according to similarities of individual points [19]. The chance of classifying objects wrongly is rather high with 1NN. The more neighbours included, the more reliable the results become. Some of the failures of 3NN are non-classified objects. Therefore the results of the two variants of KNN cannot be directly compared. KNN is performed with two different classification criteria, the Euclidean distance and the correlation coefficient. The correlation coefficient computes the correlation pairwise between the variables. KNN carried out with the correlation coefficient as classification criterion leads to better results in the case of univariate feature selection and feature selection with FT than with the Euclidean distance. For PCA, used as feature selection method, it is the opposite.

For the method comparison three different methods for feature selection are investigated, a univariate approach (univar.), FT and PCA. The univariate method, selecting the features according to the highest FC, is the simplest one. The spectral bands which are selected according to the FC are presented in the Fig. 1 a,b,c. The amounts of selected features is rather high for most of the methods. Between six and 25 variables are used for modelling. Pre-processing of the data decreases the amount of necessary features and im-

proves the classification results. The reason is that uncontrolled variations in the data require additional variables for modelling. It is possible to remove parts of such variation by applying appropriate transformation methods. With PCA the number of features is smaller compared to the other methods. Still a lot of PCs are necessary to model nine classes together.

For this data set, modelling nine classes, the best results are obtained with LDA combined with PCA as feature reduction method on SNV and first derivative pre-processed data. In both cases a successful classification of all objects of the training set and test set is achieved with 18 latent variables.

4.1.2. Modelling all batches of a concentration as one class

In the second step we want to investigate the method performance when each concentration is modelled. Therefore all batches of the same concentration are associated in one class and as a result a model is built with only four classes. This procedure seems to be more logical. The aim of classification is to discriminate between different concentrations, for instance, to know whether a new object belongs to the 2 mg tablets or to the 6 mg tablets. One is not interested in knowing to exactly which batch the sample belongs. In general, different batches of the same formulation should match each other, which is however, not always the case in real life situations due to new feed stock of the material, different equipment for production, etc. [20].

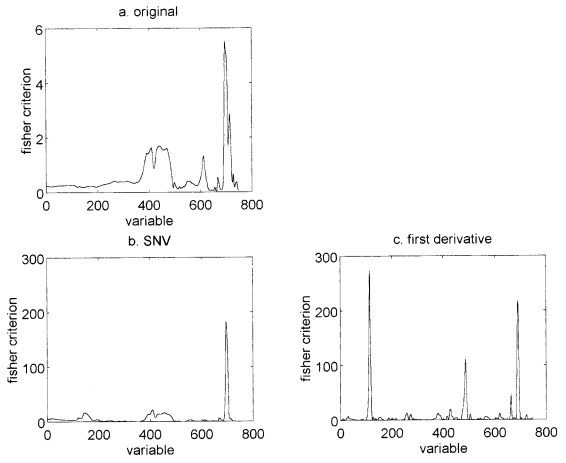

Fig. 3. Fisher criterion for the tablet data (four classes) obtained from: (a) original; (b) SNV; and (c) first derivative data.

The plots of the FC per variable for all types of pre-processed data are similar to the ones obtained in the first data analysis for the nine classes, as can be seen in Fig. 3a,b,c.

The selective spectral bands are found at the same location of the spectra, the magnitude of the peaks is somewhat decreased since the within class variance is increased. Indeed the different batches of one concentration do not exactly overlap (see Fig. 2 a,b,c). Therefore the variance (spread) of the objects within one class is higher compared to before. The Kennard and Stone algorithm is again used to select the training and test sets. For the 2 and 6 mg tablets, 40 objects are selected to belong to the training set, 15 respectively 18 samples remain for the test set. For the class of 12 mg tablets, 20 training set objects are selected, 10 are left over for the test set. For the placebo class, 80 objects are selected for the test set and the other 38 samples are included in the test set.

As in the previous data analysis the performance of LDA, QDA and KNN (1NN and 3NN) is compared, each of the methods combined with PCA, FT and the univariate feature selection method, on original, SNV and first derivative data. The maximum number of selected features is again 25, the optimum determined by LOO-CV. The model validation is carried out by an independent test set. The results are presented in Table 2.

It immediately becomes clear, that almost any method yields optimal results. In addition, it is evident that the number of selected variables is highly reduced compared to before. For the pre-processed data less features are necessary to obtain a success rate 1. LDA, QDA and KNN with Euclidean distance, combined with the univariate feature selection method are very successful, since only one variable is necessary to discriminate between the classes. The selected feature is variable 696 and 114 respectively (depending on the type of pre-processing), which has the highest FC

Table 2
Classification of tablets (four classes)

			Original			SNV			First derivative		
			Nb of variables	Success rate (training)	Success rate (test)	Nb of variables	Success rate (training)	Success rate (test)	Nb of variables	Success rate (training)	Success rate (test)
LDA		Univar.	3	1	0.9877	1	1	1	1	1	1
		FT	5	1	1	4	1	1	3	1	1
		PCA	9	1	1	7	1	1	1	1	1
QDA		Univar.	4	1	1	1	1	1	1	1	1
		FT	5	1	1	4	1	1	3	1	1
		PCA	5	1	1	8	1	1	3	1	1
1NN	Euclidean distance	Univar.	13	0.8944	0.9506	1	1	1	1	1	1
		FT	13	0.9833	1	11	1	1	11	1	1
		PCA	17	0.9833	1	16	1	1	3	1	1
1NN	Correlation coeff.	Univar.	9	1	1	6	1	1	6	1	1
		FT	11	1	1	9	1	1	7	1	1
		PCA	24	0.65	0.8765	24	0.8222	0.9753	7	0.7667	0.6543
3NN	Euclidean distance	Univar.	25	0.8444	0.9383	1	1	1	1	1	1
		FT	13	0.9722	1	11	1	1	11	1	1
		PCA	13	0.9611	1	16	1	1	1	1	1
3NN	Correlation coeff.	Univar.	9	1	1	7	1	1	6	1	1
		FT	11	1	1	11	1	1	11	1	1
		PCA	25	0.6944	0.8395	25	0.8167	0.9383	25	0.7944	0.8519

(see Fig. 3b and 3c). This means that the multivariate approaches are reduced to univariate methods with one discriminating variable. There is however a risk that such models are not stable enough when selecting only one original variable. For instance, small wavelength shifts might lead to bad prediction. More Fourier coefficients and in some cases more latent variables are needed to obtain optimal classification rate. It is important to remember, that PCs are linear combinations of the original variables and include also spectral information which has no discriminating property. KNN is used in its original form for equal number of objects in each class, although the classes of the training set contain different amounts of objects. Consequently, one might obtain even better results with a method with alternative (i.e. not simple majority) rules [17,21].

For this data set, one can summarise that it is advantageous to model each concentration as a class, since the batches of the same concentration are similar. As a result simple classification models for less classes can be established with a few selected features only. The best performance is obtained by LDA and 3NN (first derivative data). For both methods only one PC, namely PC1, is needed.

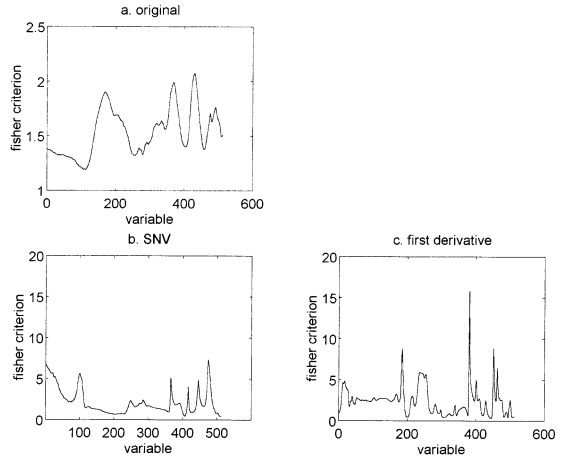

Fig. 4. Fisher criterion for the capsule data (11 classes) obtained from: (a) original; (b) SNV; and (c) first derivative data.

4.2. Capsule data set

4.2.1. Modelling each class

This data set contains spectra from six different concentrations. For some concentrations more than one batch is included, so that the data set consists in total of 11 classes and 316 objects. The collected capsule spectra do not contain relevant information in the spectral region from 10 000 to 8000 cm^{-1}. For this reason 251 variables are discarded at the beginning and to remove edge effects, 15 more variables are discarded at the end of the spectra. The resulting spectra have a dimension of 512 features, representing the spectral region of 8064–4116 cm^{-1}. The data analysis and modelling is repeated following the same procedure as used for the first data set. Fig. 4a–4c represents the FC for each type of pre-processed data.

For the original data (Fig. 4a) no features are observed with a high ratio of between class variance to within class variance. Only after pre-processing (Fig. 4b and c) a few variables with a somewhat higher FC could be obtained, since the within class variance is decreased. The three highest peaks in the FC spectrum for the first derivative data correspond to peaks in the original spectrum of the capsules.

PCA is performed on the pre-processed, centered data. The PC1–PC2 score plots are given in the Fig. 5a,b,c. Since the capsule data set is even larger than the tablet data set, only each third object is displayed in the figures, in order to simplify the plots.

Clearly most of the classes are overlapped. For the original and SNV data class 5 is fairly separated from the other classes and for the first derivative data the classes 3 and 8 from the remaining classes. To obtain more information the LDA displays for the three types of data are shown in Fig. 6a,b,c. The 10 first PCs where used to determine the canonical variables.

Canonical variable 1 (CV1) is plotted against canonical variable 2 (CV2). It seems feasible to achieve a discrimination for the different classes, since already three to four classes are discriminated on the CV1–CV2 plots. Derivatives lead to clearly better class separation than SNV. One

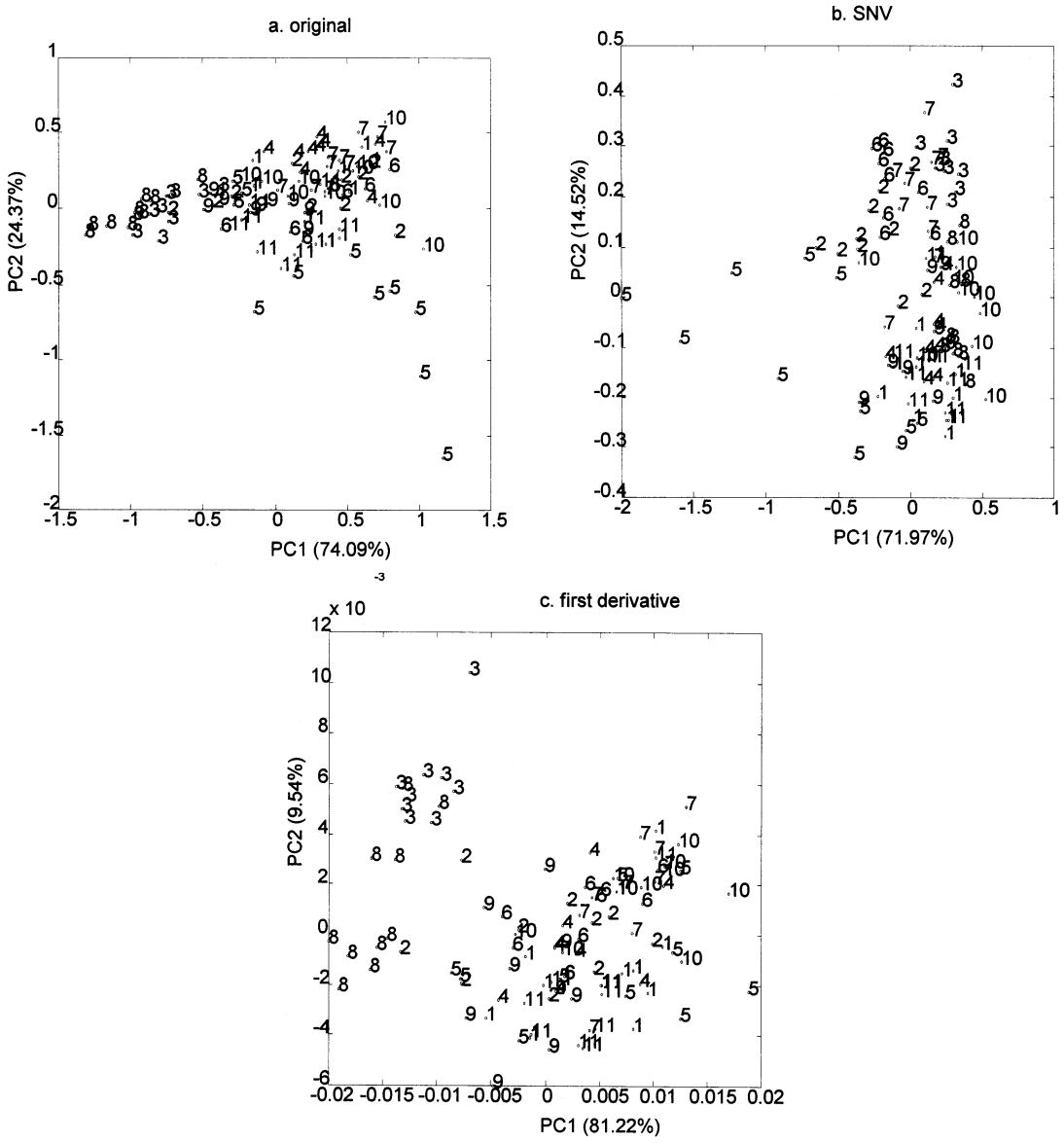

Fig. 5. PC1 vs. PC2 score plot for the capsule data (11 classes) obtained from: (a) original; (b) SNV; and (c) first derivative data.

could suspect from these plots that collecting all batches of the same concentration together in one class might not be reasonable, because the clusters for each concentration are becoming very large and the data non-normally distributed. These displays reveal that the capsule data set is more difficult than the tablet data set.

For further data investigation the single Grubb's test on Rao's statistics is applied on each class separately to check for outliers. Some objects are detected as outliers at an α-level of 5%. After consulting the person responsible for the measurement, we decided to keep all objects, since it is considered that this represents the normal

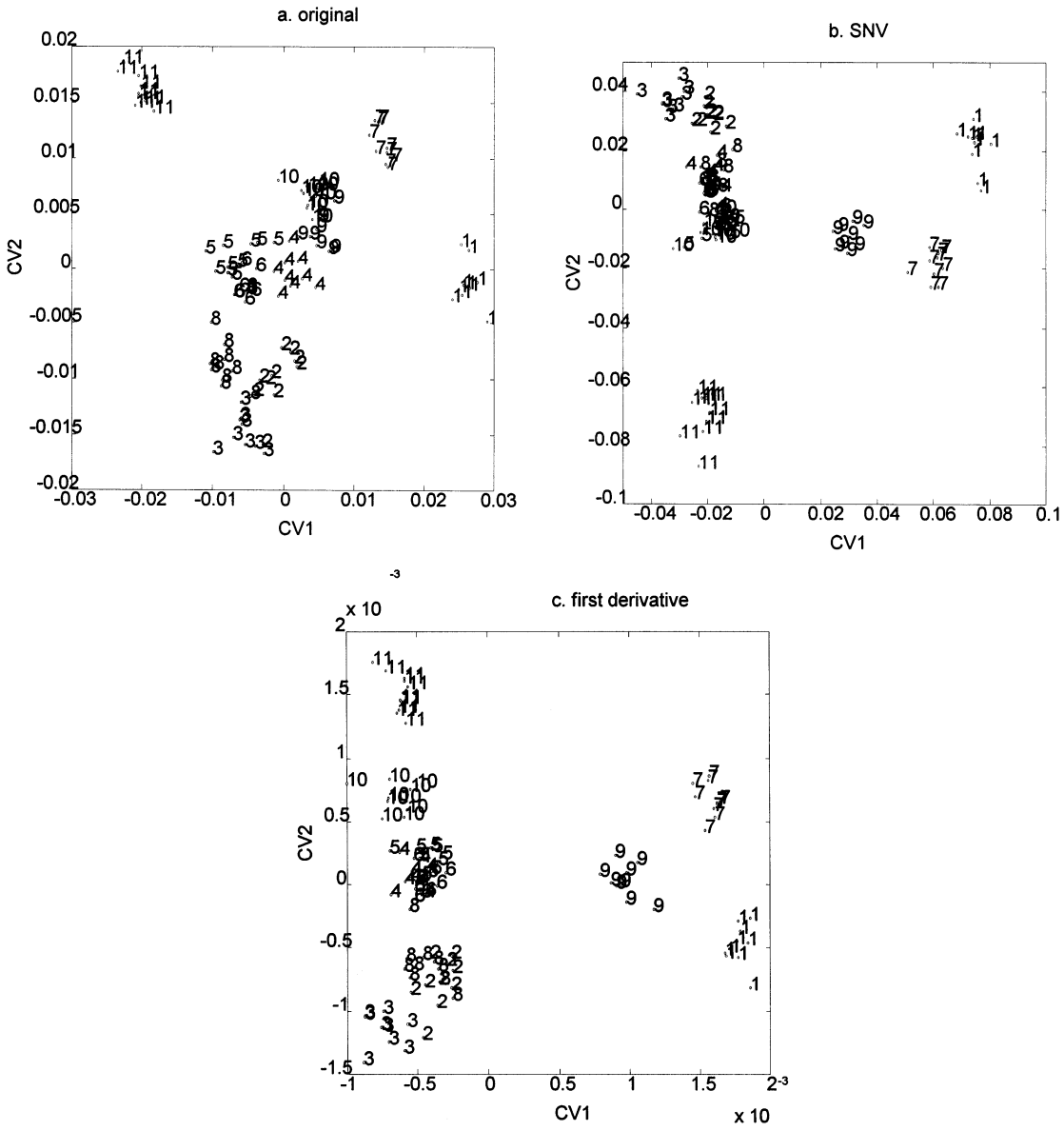

Fig. 6. CV1 vs. CV2 score plot for the capsule data (11 classes) obtained from: (a) original; (b) SNV; and (c) first derivative data.

variability of the samples. The data set is again divided into a training and a test set by applying the Kennard and Stone algorithm to each class. The data set division is repeated for each type of pre-processing. The resulting training set consists of 220 objects (11 classes, 20 objects), the test set of 96 objects (3×10, 4×9, 2×8 and 2×7 objects).

The same methods are applied as to the tablet data set, namely LDA, QDA, 1NN and 3NN (Euclidean distance and correlation coefficient) combined with FT, PCA and the univariate fea-

Table 3
Classification of capsules (11 classes)

			Original			SNV			First derivative		
			Nb of variables	Success rate (training)	Success rate (test)	Nb of variables	Success rate (training)	Success rate (test)	Nb of variables	Success rate (training)	Success rate (test)
LDA		Univar.	13	0.9	0.9167	24	0.9955	1	24	1	0.9896
		FT	10	1	1	14	1	1	9	1	0.9896
		PCA	11	1	1	15	1	1	10	1	1
QDA		Univar.	7	0.6045	0.6667	9	0.7818	0.7708	7	0.9318	0.9375
		FT	10	0.9955	0.9896	10	0.9818	0.9896	9	0.9955	0.9896
		PCA	10	0.9909	0.9896	11	0.9909	1	8	0.9955	1
1NN	Euclidean distance	Univar.	18	0.3136	0.3646	23	0.7	0.7708	24	0.8273	0.8542
		FT	15	0.4591	0.5938	23	0.6682	0.9375	23	0.5636	0.9583
		PCA	3	0.6364	0.7083	5	0.8364	0.9688	8	0.8091	0.9792
1NN	Correlation coeff.	Univar.	24	0.4364	0.5313	25	0.5182	0.5104	25	0.8318	0.8854
		FT	24	0.6864	0.8646	24	0.65	0.9479	24	0.8636	0.8854
		PCA	20	0.4682	0.7708	12	0.7727	0.9479	8	0.8364	0.9479
3NN	Euclidean distance	Univar.	23	0.1955	0.2604	20	0.6727	0.7188	25	0.7591	0.8438
		FT	15	0.3091	0.5313	21	0.5727	0.8125	4	0.5	0.6667
		PCA	3	0.6364	0.6979	5	0.7725	0.9271	8	0.7409	0.9375
3NN	Correlation coeff.	Univar.	21	0.3364	0.375	21	0.4227	0.3958	16	0.8091	0.8542
		FT	22	0.6318	0.7813	25	0.5591	0.8125	24	0.7909	0.9792
		PCA	15	0.3955	0.5938	9	0.6727	0.875	8	0.75	0.9479

ture reduction method, on original data, SNV and first derivative data. LOO-CV is performed in the training set to find the optimal number of variables (maximum 25) and the independent test set is used for the validation of the model. The results obtained for the different methods established for 11 classes are presented in Table 3.

Compared to the original data, improved success rates for the training and test sets are obtained after pre-processing of the data. This is especially clear for the methods which show bad performance, such as 1NN and 3NN. KNN does not include any optimisation of the discrimination between classes in its methodology. By applying a pre-processing method at least the within class variance is somewhat decreased, which automatically increases the between/within class variance ratio. Best results are obtained with first derivative data.

The success rates of the training and test sets are comparable for the parametric methods, where class borders are defined. However for the non-parametric methods the performance obtained with the test set is better than the one of the training set. This can be explained by the way KNN works, namely by comparing objects to its neighbours. Since the training and test selection was carried out with the Kennard and Stone algorithm, the samples from the training set are more likely to be situated at the border of a class and will therefore have a higher chance to lie next to an object from another class than the test set samples, located in the centre of a class.

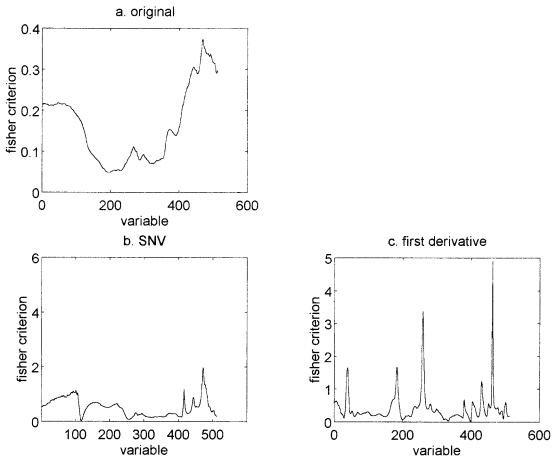

Fig. 7. Fisher criterion for the capsule data (six classes) obtained from: (a) original; (b) SNV; and (c) first derivative data.

The comparison of the different methods clearly shows that LDA and QDA perform best even with original data. Since the assumptions for LDA seem to be sufficiently well fulfilled, the highest performance is obtained with this method. The reason that the results of QDA are not as good can be explained with the same arguments already given for the tablets. The quality of the results obtained with KNN is not acceptable for this data set, especially not in the case of the original data. PCA seems to be the preferable feature reduction method. In most of the cases the best success rate is obtained when working with latent variables and also less features are selected. In some cases pre-processing decreases the amount of selected variables.

The best results for this data set are obtained with LDA, carried out with 10 PCs obtained after feature reduction, on first derivative data and with LDA, performed with 11 PCs on original data. With these models a success rate of 1 is obtained for the training and test sets.

4.2.2. Modelling each concentration

As mentioned in the discussion of Figs. 5 and 6, modelling each concentration as a class might not seem promising when looking at the capsule score plots. However, these plots show only the scores on two latent variables, PC1 and PC2. It might still be possible that in higher dimensions the different batches of the same concentration match each other. For this reason the batches of the same concentration are again associated in one class and the data analysis repeated. The models are built for a data set containing six classes (placebo, 0.5 mg, 1 mg, 1.5 mg, 3 mg and 6 mg active).

The plots of the between class variance/within class variance are presented in Fig. 7a,b,c.

The plot of the FC obtained for the original data has values from 0 to 0.4, which means that the within class variance is bigger than the between class variance. After pre-processing with

Table 4
Classification of capsules (six classes)

			Original			SNV			First derivative		
			Nb of variables	Success rate (training)	Success rate (test)	Nb of variables	Success rate (training)	Success rate (test)	Nb of variables	Success rate (training)	Success rate (test)
LDA		Univar.	18	0.9636	0.9688	24	0.9545	0.9792	25	0.9864	1
		FT	18	1	0.9896	17	1	1	25	1	1
		PCA	20	1	0.9896	18	0.9955	1	19	1	1
QDA		Univar.	9	0.8455	0.8542	13	0.7182	0.7813	10	0.9318	0.9896
		FT	12	0.9682	1	11	0.9864	0.9688	13	0.9818	0.9896
		PCA	12	0.9773	1	10	0.9818	0.9792	7	0.9773	0.9896
1NN	Euclidean distance	Univar.	23	0.2864	0.4583	21	0.75	0.8854	25	0.8818	0.9688
		FT	5	0.4818	0.6042	23	0.7136	0.8958	18	0.5955	0.9479
		PCA	8	0.8818	0.8542	14	0.8364	1	9	0.9091	0.9896
1NN	Correlation coeff.	Univar.	23	0.7273	0.8646	25	0.65	0.7813	23	0.8864	0.9479
		FT	24	0.75	0.8542	20	0.6864	0.9063	25	0.8864	0.9688
		PCA	8	0.7455	0.7396	18	0.7728	0.9896	13	0.85	0.9792
3NN	Euclidean distance	Univar.	17	0.2227	0.2813	24	0.7364	0.8646	21	0.8727	0.9375
		FT	4	0.3318	0.5104	25	0.6545	0.8646	14	0.5227	0.8542
		PCA	8	0.8227	0.8438	18	0.7909	0.9479	8	0.9273	0.9896
3NN	Correlation coeff.	Univar.	23	0.7318	0.8646	17	0.6227	0.7292	21	0.8455	0.9063
		FT	24	0.6955	0.7708	21	0.6318	0.8438	19	0.8273	0.9479
		PCA	8	0.6909	0.6875	7	0.7455	0.8333	13	0.8	0.9688

SNV, two spectral regions with values slightly over one are obtained, after first derivative already five spectral regions with FC values larger than one. The magnitude of the ratio is however still very low.

Each class is divided into training and test sets by applying the Kennard and Stone algorithm. For the placebo, 0.5 and 6 mg capsules, 20 objects are selected for the training set, 40 samples for the 1 mg capsules and 60 objects for the 1.5 and 3 mg capsules. The remaining objects are collected in the test set.

The models and the results obtained with the different feature selection and classification methods for the original, SNV and first derivative data are presented in Table 4.

The results concerning the success rates for the parametric methods are comparable to the ones obtained when modelling 11 classes. Although less classes were modelled together this time, which should consequently lead to models with less factors, more variables are necessary to reach these results, except for QDA. This leads to unstable models in general.

Better results are obtained when each batch is modelled individually. For this data set the method of choice is LDA or QDA combined with PCA.

5. Conclusion

Classification models are elaborated for two NIR data sets coming from pharmaceutical industry. The first data set contains spectra from tablets in four different concentrations and nine classes in total. The second data set consists of spectra from capsules in six concentrations and 11 classes in total. A full data analysis is carried out including diagnostics, feature reduction and modelling and validation of the model.

We focused in the method evaluation part on hard modelling techniques. The drawback of these methods is that they are discriminating between given classes and do not perform positive identification. It might be necessary to propose a two step procedure for the final classification of clinical study lots, first to discriminate between given classes and secondly to apply a method which allows a positive identification. The second part is not discussed in this manuscript.

It is worthy to evaluate during the method development whether one should model each batch individually or each concentration, in order to obtain simple and stable models. Simple tools such as diagnostics (plots of the FC, PC score plots) help to reveal what should be done. However, one has to be careful with drawing too fast conclusions, because they might be misleading in the situation of difficult data sets (here capsules data set). For the table data set it turned out that the modelling of each concentration is preferable, whereby for the capsule data set one should classify the different batches. The different classes for tablets can be correctly classified with LDA, combined with PCA as feature reduction method. Also for the capsules LDA combined with PCA should be the method of choice. First derivative is the most suitable signal processing method for both data sets. This type of pre-processing is however not necessarily very robust, since small changes in spectra due to instrumental reason can lead to bad predictions [22]. For long term application of the models one should therefore prefer SNV or work with the original data. With these types of pre-processing perfect classification is still obtained, but with more variables.

References

[1] P. Corti, E. Dreassi, G. Ceramelli, S. Lonardi, R. Viviani, S. Gravina, Analysis 19 (1991) 198–204.
[2] P. Corti, L. Savini, E. Dreassi, G. Ceramelli, L. Montecchi, S. Lonardi, Pharm. Acta Helv. 67 (1992) 57–61.
[3] B.F. MacDonald, K.A. Prebble, J. Pharm. Biomed. Anal. 11 (1993) 1077–1085.
[4] P.K. Aldridge, R.F. Mushinsky, M.M. Andino, C.L. Evans, Appl. Spectrosc. 48 (1994) 1272–1276.
[5] K.M. Morisseau, C.T. Rhodes, Drug Dev. Ind. Pharm. 21 (1995) 1071–1090.
[6] M.P. Derde, D.L. Massart, Anal. Chim. Acta 191 (1986) 1–16.
[7] D. Bertrand, Data Pretreatment and Signal Analysis in Spectroscopy, Advanced Comett Chemometrics School, Libramont (B), 1993.
[8] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spectrosc. 43 (1989) 772–777.
[9] P.A. Gorry, Anal. Chem. 62 (1990) 570–573.
[10] F.E. Grubbs, G. Beck, Technometrics 14 (1972) 847–854.

[11] P.C. Kelly, J. Assoc. Off. Anal. Chem. 73 (1990) 58–64.
[12] B. Mertens, M. Thompson, T. Fearn, Analyst 119 (1994) 2777–2784.
[13] V. Centner, D.L. Massart, O.E. de Noord, Anal. Chim. Acta 330 (1996) 1–17.
[14] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137–149.
[15] W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I.R. Last, K.A. Prebble, Chemometr. Intell. Lab. Syst. Lab. 33 (1996) 35–46.
[16] W. Wu, B. Walczak, D.L. Massart, K.A. Prebble, I.R. Last, Anal. Chim. Acta 315 (1995) 243–255.
[17] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: a Textbook, Elsevier, Amsterdam, 1988, p. 217.
[18] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D.L. Massart, Anal. Chim. Acta 329 (1996) 257–265.
[19] D.D. Wolff, M.L. Parson, Pattern Recognition Approach to Data Interpretation, Plenum, New York and London, 1983, p. 80.
[20] A. Candolfi, S. Heuerding, D.L. Massart, Anal. Chim. Acta 345 (1997) 185–196.
[21] D. Coomans, D.L. Massart, Anal. Chim. Acta 136 (1982) 15–27.
[22] E. Bouveresse, C. Casolino, D.L. Massart, Assessing the validity of near-IR monochromator calibrations over time, submitted.

.