A Wavenumber selection approach for sample classification in the petroleum sector

Soares, Felipe; Anzanello, Michel.

Abstract

In recent years, spectroscopy techniques such as Near-infrared (NIR) and Fourier Transform Infrared (FTIR) have been adopted as analytical tools in different fields. A spectrum of a sample usually has hundreds of wavenumbers, fact that can jeopardize the accuracy of statistical analysis, being the variable selection an important step in prediction and classification tasks based on spectroscopy data. This paper proposes a novel methodology for wavenumber selection in classification tasks, applied in two data sets from the petroleum sector. The method consists of two main stages: determination of intervals based on the distance between the average spectra of the classes and the selection of the most suitable intervals through cross-validation. An improvement of 11.52% in the misclassification rate was achieved for a NIR spectra data set of diesel, decreasing from 11.71% to 10.36% after the application of the proposed method. For a biodiesel FTIR data set the method proved to be robust, achieving a zero misclassification rate after the selection process, compared to its initial value of 4.71%.

Key Words: Spectroscopy, Fuel classification, Wavenumber selection, Diesel/Biodiesel.

1. Introduction

Spectroscopic techniques have recently gained a great importance in various fields such as food, pharmaceutical and petroleum. It is a powerful analytical tool to assess solids and liquids due to its ability to provide detailed chemical information and require little or no sample pretreatment. Usually a large number of wavenumbers (variables derived from spectroscopic techniques) can be rapidly acquired by modern equipment, however the large number of variables tends to reduce exploratory and predictive performance of several multivariate techniques [1,2].

In multivariate data analysis, one of the major issues is the dimensionality of the data, i.e., the number of wavenumbers (or variables) related to each observation. In the pre-processing phase of high-dimensional data, such as spectral chemometrics, the dimension reduction is an important phase of the pre-processing and can improve the quality or speed of subsequent analyses [1,3]. Literature brings several methods and approaches tailored to wavenumber extraction and selection.

A technique widely applied to wavenumber extraction is Principal Components Analysis (PCA), which aims to replace the original wavenumbers by a new set of uncorrelated wavenumbers (called components) originated as linear combinations of the original ones; each component explains the maximum of variance in the original data as possible [4,5]. Usually a few components are required to explain most of the variance [6].

In wavenumber selection the aim is to find in the original set of *d* wavenumbers a new subset of *n* wavenumbers, with $n<d$, where the results do not change dramatically as the information in the discarded wavenumbers is redundant [7,8]. Many different methods have been proposed in the field of supervised learning, especially in chemical analysis [1,2,9], however, there is no standard wavenumber selection procedure for classification purposes.

In this paper, we propose a novel method for wavenumber selection for classification of chemical samples into quality related classes. For that matter, we first apply a technique for splitting the data set into spectra intervals and transform such data using PCA. Such intervals are then integrated to different classification tools, such as k-nearest neighbors (KNN) and linear discriminant analysis (LDA) to categorize the samples into proper classes; the classification accuracy provided by each classification technique is finally calculated and compared. The method can be summarised as two main stages: the determination of the intervals using the average spectra of each class and the selection of the most suitable intervals using classification tools along with cross-validation. This method results in measures of misclassification rates according the three classification tools and the three selection approaches, giving a broad view of the performance of each one.

The paper is structured as follows: section 2 consists of a brief literature review of variable selection applied to spectral data and classification tasks; in section 3 we present the method and the two data sets from fuels used in this study. In section 4 results and presented and compared; conclusions are depicted in section 5.

2. Background

We now present the fundamentals of the multivariate techniques our method relies on.

2.1 Principal Component Analysis

PCA is a well know tool for dimension reduction and graphical visualisation of data. In PCA, the original values are projected onto a new orthogonal coordinate system that better represents the data, maximising the variance. To get this new basis the PCA finds eigenvalues and eigenvectors of the covariance matrix of the data, aiming to maximise the variance and minimise the redundancy. Consider **X** as an *n x d* mean-centred matrix, where *n* is the number of observations and *d* the features. The covariance matrix **C** must be computed as in Eq. (1).

$$\mathbf{C_X} = \frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\,\mathbf{X} \qquad (1)$$

The goal is to diagonalise the matrix **C** to obtain uncorrelated features, so the Eigen decomposition yields Eq. (2). The **P** *d x d* matrix contains, by columns, the eigenvectors of the matrix **C**, and **D** is a diagonal matrix containing the correspondent eigenvalues.

$$\mathbf{P}^{-1}\mathbf{C_X}\mathbf{P} = \mathbf{D} \qquad (2)$$

Next, it is necessary to sort the eigenvectors according their eigenvalues magnitudes; the greater is the eigenvalue, the greater is the variation along that direction. Consider **V** as the eigenvectors of **P** sorted in a descent order according the values of the diagonal of **D**.

$$\mathbf{V} = [\mathrm{P_a}\;\mathrm{P_b}\;\mathrm{P_c} \ldots]\text{where } \mathrm{D_{aa}} > \mathrm{D_{bb}} > \mathrm{D_{cc}} \ldots \qquad (3)$$

V is a *d x d* matrix where the columns are known as loadings (or weights) and allows the original data to be projected on the new basis. Typically, a few *k* components of **V** are required to express the data properly. The projected data **Y** is expressed by Eq (4), also known as PCA scores. Further details on PCA can be found in [3–5,8,10].

$$\mathbf{Y} = \mathbf{XV} \text{ where } \mathbf{V} \text{ can be truncated to a } d\ x\ k \text{ matrix} \quad (4)$$

2.2 *Linear Discriminant Analysis*

Linear Discriminant Analysis (LDA) is a well known method for data classification, being firstly introduced by Fisher in 1936 [11]. The LDA can be seen as a dimension reduction tool that projects the data on a new subspace where the classes are better separated, then the projected data can be used to construct a classifier. A threshold y_0 for the projected data is chosen, so if $y(x) \geq y_0$ the sample is classified as belonging to class C_1, otherwise it will be assigned to class C_2 [4].

To apply the LDA for a k classes case it's necessary to evaluate the within-class covariance matrix by:

$$\mathbf{S_W} = \sum_{k=1}^{K} \mathbf{S_k} \tag{5}$$

Where

$$\mathbf{S_k} = \sum_{n \in C_k} (X_n - m_k)(X_n - m_k)^T \tag{6}$$

$$m_k = \frac{1}{N_k} \sum_{n \in C_k} (X_n) \tag{7}$$

Also the between-class covariance matrix must also be evaluated through Eq. (8). In case the variance of classes are significantly different, a Non Linear Discriminant Analysis is to be carried out.

$$\mathbf{S_B} = \sum_{k=1}^{K} N_k(m_k - m)(m_k - m)^T \tag{8}$$

$$m = \frac{1}{N} \sum_{n=1}^{N} X_n \tag{9}$$

N_k is the total number of samples in the k-th class, m_k is the mean of the samples in the class, m is the mean of all samples and X_n is an individual sample of the training set that contains N samples.

Since the projection of the data is given by $\mathbf{Y} = \mathbf{W^T X}$, the within-class and between-class covariance matrices become $\mathbf{W^T S_W W}$ and $\mathbf{W^T S_B W}$. Once LDA aims to project the data in a subspace where the classes are better separated, the goal is to find a $\mathbf{W}$ where the ratio of between-class to the within-class covariance matrixes – known as Fisher criterion – is maximised [5,12,13].

$$J_F(\mathbf{W}) = \frac{|\mathbf{W^T S_B W}|}{|\mathbf{W^T S_W W}|} \tag{10}$$

Thus:

$$\mathbf{W}^* = argmax(\frac{|\mathbf{W^T S_B W}|}{|\mathbf{W^T S_W W}|}) \tag{11}$$

The maximisation of the criterion is deeply discussed by Fukunaga in reference [14].

2.3 K-nearest neighbor

The k-nearest neighbor (KNN) is a widely applied as a classification tool. The method was first proposed by Fix and Hodges [15], leading to several variations [16–19]. To classify a new X sample in a C_d class, KNN computes the distances of the new sample to the samples in the training set using a distance metric, e.g., Euclidean or Mahalanobis distances. The matrix of distances is sorted and the new observation is inserted into the class that appears the most in the first k-closest neighbors. The posterior probability of class membership of the new X sample is given by Eq (12).

$$p(C_d|X) = \frac{k_d}{k} \qquad (12)$$

where k is the chosen number of closest neighbors and k_d is the number of neighbors in the training set that belong to the class C_d. Alternatively, the X sample will be assigned to the class with greater posterior probability, on other words, to the class that has more samples among the k-nearest set [4]. One may choose the k number via cross-validation or approximate it by the square root of the number of samples [13,14].

2.4 Probabilistic Neural Networks:

Neural networks have their origins as an attempt to replicate mathematically the processing of information in biological systems by mimicking the brain activity. A neural network usually consists of input, hidden and an output layers. The samples are presented to the units in the input layer, which are linearly combined in the hidden layers. Next, each combination is transformed according a nonlinear activation function, e.g., logistic or hyperbolic tangent functions, giving the output of the hidden layer [4]. The hidden units outputs are linearly combined and transformed again in the output layer units. The choice of the activation function depends on the nature of the output [4,5] . Training samples are presented to the network to determine the weights of the linear combinations.

An adaptation of neural networks for classification, the probabilistic neural network (PNN), was developed by Specht [20]. In PNN the activation function is replaced by a statistically derived one, asymptotically approaching the Bayes optimal decision surface. The PNN has the advantage of being easy to re-train, however PNN can be computer memory consuming [21]. Details about probabilistic neural networks are available in [20–23].

2.5 Spectroscopy and wavenumber selection in chemometrics applications:

Spectroscopy combined with multivariate techniques has been a recurring topic in literature, leading to remarkable results in analytical calibrations and classification. Dyrby *et al.* [24] developed chemometric calibrations based on NIR and Raman spectroscopy using partial least squares regression (PLS) applied to the pharmaceutical industry. A reference method was used to measure the active substance content of pharmaceutical tablets, which had four different dosages, and then PLS applied to build predictive models. The results presented by the authors had relative prediction errors comparable to the estimated error of the reference method, what makes spectroscopy combined with multivariate data analysis a cost-effective alternative to the standard method. Pimentel *et al.* [25] also presented a calibration model to determine the content of biodiesel in diesel fuel blends based on middle (MID) and NIR spectroscopy. PLS models were built and proven to be a practical method to predict biodiesel content aimed at monitoring the quality of biodiesel fuel.

The literature reports several authors [26–29] who applied spectroscopy and multivariate analysis to insert petroleum products into categories, such as gasoline and diesel/biodiesel blends. Kim *et al.* [29] constructed a real-time classifier (RTC) based on PCA and a Bayesian classifier to classify NIR spectra of six different petroleum products. The proposed classifier offered faster and more accurate classification of products on-line than the conventional approaches. Li and Dai and Balabin *et al.* [27,28] applied classification tools to Raman and NIR spectroscopy data to classify gasoline by brand and source, comparing several classifiers that included LDA, Quadratic discriminant analysis (QDA), least squares support vector machine (LSSVM) and KNN. In both works, LDA presented the greatest misclassification ratio, while LSSVM and KNN had better results. Pontes *et al.* [26] used partial least squares discriminant analysis (PLS-DA) and LDA to construct classifiers based on NIR spectra to identify diesel/biodiesel adulteration. Both methods provided similar results, however PLS-DA obtained better classification rates for all cases.

Most of the aforementioned studies employed some wavenumber selection during the building of the models aiming to improve performance. Xiaobo *et al.* [1] brings a deep review on the most applied methods in NIR wavenumber selection, such as successive projection algorithms (SPA) and interval selection methods. The authors emphasize that the algorithms work in different ways and are suitable to different type of data, requiring the user to select the most appropriated one.

More aligned with the propositions of this paper, an interval PLS (iPLS) wavenumber selection procedure proposed by Norgaard *et al.* [30] was used by Dyrby *et al.* [24] to optimize prediction and interpretation of the models. The iPLS is a graphically oriented approach that splits the spectrum in several equidistant intervals that are then modelled by local regressions. The prediction error measure is evaluated for each interval; the interval yielding the smallest error is chosen. In Ferrão *et al.* [31], the synergy interval PLS (siPLS) and iPLS are compared to the full spectra PLS in the prediction of quality parameters of biodiesel/diesel blends. The siPLS is a method similar to iPLS, but all the combinations of two, three or four intervals are evaluated and the one with smaller error measure is chosen. The results show that the siPLS algorithm has a better performance probably due to the combination of intervals that are not necessarily adjacent, as reasoned by the authors.

3. Materials and Method

In this section we present the method and the two data sets assessed in this article.

3.1 Biodiesel/Diesel blends data set

A total of 85 samples of Brazilian Biodiesel/Diesel blends are available in the data set used in Ferrao *et al.* [31]. The samples were prepared with biodiesel constituted of soybean methyl esters and two types of diesel: metropolitan and countryside. The concentrations ranged from 0.2% of biodiesel (v/v) to 30.0% (v/v).

Fourier transform infrared spectroscopy (FTIR) spectra of the samples were obtained by Ferrao *et al.* [31] at room temperature, with spectra from 650 to 4000 cm^{-1} and spectral resolution of 4 cm^{-1}. The resulting data set consists of 1738 wavenumbers (λ) and 85 samples. Since there are two different types of diesel, the samples can be classified according metropolitan (56 samples) or countryside (29 samples) diesel.

3.2 Diesel fuel data set

The Southwest Research Institute, sponsored by the U.S. Army, collected the NIR spectra of samples of diesel fuel along with five properties. One of the properties is the cetane number (analogous to the octane number in gasoline), indicating the ignition quality of the fuel. The full data set is available at Eigenvector Research Inc. website (http://www.eigenvector.com/).

In Brazil, the leading fuel company (Petrobras) uses the cetane number as a measure to classify the diesel in standard or premium classes [32]. A standard diesel fuel has a minimum cetane number of 42, while the premium diesel has a minimum of 51. Based on that information, the samples were classified in standard and premium classes according the cetane number. The resulting data set contains 401 wavenumbers and 222 samples divided in two classes; standard has 172 samples and premium has 50 samples.

3.3 Proposed Method

The proposed method resembles the iPLS proposed by Norgaard *et al.* [30], however it's focused in classification and finds the intervals in a different approach. While iPLS splits the spectra in several equidistant intervals, the proposed method relies on finding peaks in the difference spectrum and relevant neighbourhood regions around such peaks.

Consider a matrix **X** consisting of N fuel samples; it can be split in two matrices $\mathbf{X_{C1}}$ and $\mathbf{X_{C2}}$ according to each class sample. The proposed method for wavenumber selection is based on finding the regions of the spectra where the difference of the average spectra of each class is above a certain threshold. The detailed steps to find the intervals are presented below:

Step 1 – Split the matrix **X** in the matrices $\mathbf{X_{C1}}$ and $\mathbf{X_{C2}}$ according class 1 and class 2;

Step 2 – Compute the average spectra (arithmetic mean) of each class as the vectors $\overline{x_1}$ and $\overline{x_2}$;

Step 3 – Compute the absolute value of the difference of $\overline{x_1}$ to $\overline{x_2}$, as in equation (13):

$$D = |\overline{x_1} - \overline{x_2}| \quad (13)$$

Step 4 – Find the *p* peaks in D that are above a given threshold *t* (e.g. the mean of D);

Step 5 – Find an interval $\mathbf{I_p}$ for each peak identified in Step 4 through their local minimum. The interval starts in the previous local minimum related to the peak and ends in the next local minimum, and must not overlap other intervals. This step yields *p* intervals of wavenumbers. Fig 1 illustrates this step.

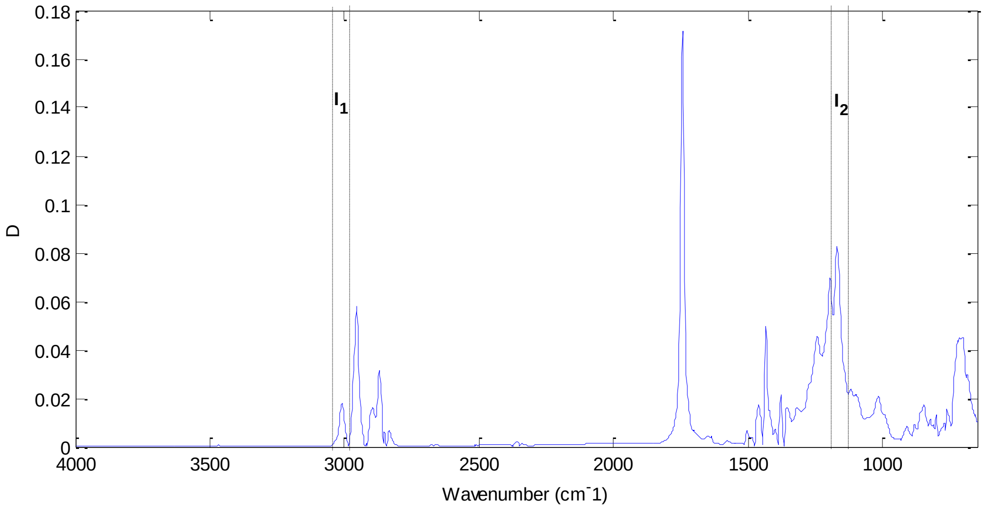

Figure 1: Example of two peaks of D and their correspondent intervals I_1 and I_2

Once the *p* intervals are found, we carry out a process to select the most relevant intervals to be included in the classification technique. In this work we used three different approaches to select the intervals, all based on the misclassification rate in the 10-fold cross validation, namely: Forward Selection (FS), Backward Elimination (BE) and all possible combinations of 1, 2, 3 and 4 intervals.

In the first selection approach, Forward Selection, the misclassification rate (MR) of the cross-validation in evaluated for each interval; the interval with the lowest MR is chosen to be included in the model. The previous subsets of chosen intervals are used in the following iterations. The algorithm stops when there is no improvement in the classification process, or when all intervals are selected.

The Backward Elimination (the second approach tested) is similar to the FS, however it starts with all intervals in the model. For each iteration, one interval is omitted at a time and the MR is evaluated, the interval that causes the highest MR is eliminated. The algorithm stops when no improvement is achieved, or when there is just a single interval remaining in the model. The last approach we test is the combination of all possible, which is based on the work of Norgaard *et al.* [30] regarding interval selection in PLS (synergy PLS). The approach is adapted to the classification context; all combinations of 1, 2, 3 and 4 intervals are cross-validated and the combination with the smallest MR is chosen.

Each of the aforementioned selection approaches are integrated to LDA, kNN and PNN classification tools. Given that LDA does not perform well when the number of samples is smaller than the number of variables, PCA is applied to the data before each iteration as a pre-processing step. All the combinations of selection approaches and classification tools are assessed; the one yielding the maximum accuracy and minimum number of retained intervals is chosen.

4. Results and Discussion

4.1 Parameters definition

To better compare the selection approaches and classification tools, their parameters should be optimized independently. Similar to the procedure followed by Balabin and Smirnov [2] all combinations of selection approaches and classification tools were optimized based on a 10-fold cross validation procedure aiming to minimize the misclassification rate. Optimizing each tool independently allow one to compare them without bias, even though the CPU time for the optimization can be large [2].

The parameters optimized for the kNN tool were the number of principal components and the number of k-neighbors, while for the PNN the spread parameter and the number of PCs were optimized. The number of PCs in the LDA was cross-validated between 1 and 7, as with more than 7 principal components the algorithm tended to fail due to singularity issues.

4.2 Biodiesel/Diesel blends data set results

The proposed wavenumber selection method intended to classify biodiesel/diesel samples in countryside and metropolitan biodiesel. The method yielded a substantial reduction in the misclassification rate, achieving a 0% error in three different combinations of classification tools and selection approaches. The complete results are depicted in Table 1. The best MR without applying the wavenumber selection was 4.71% with LDA as a classification tool, followed by kNN with 5.88% and PNN with 8.24%.

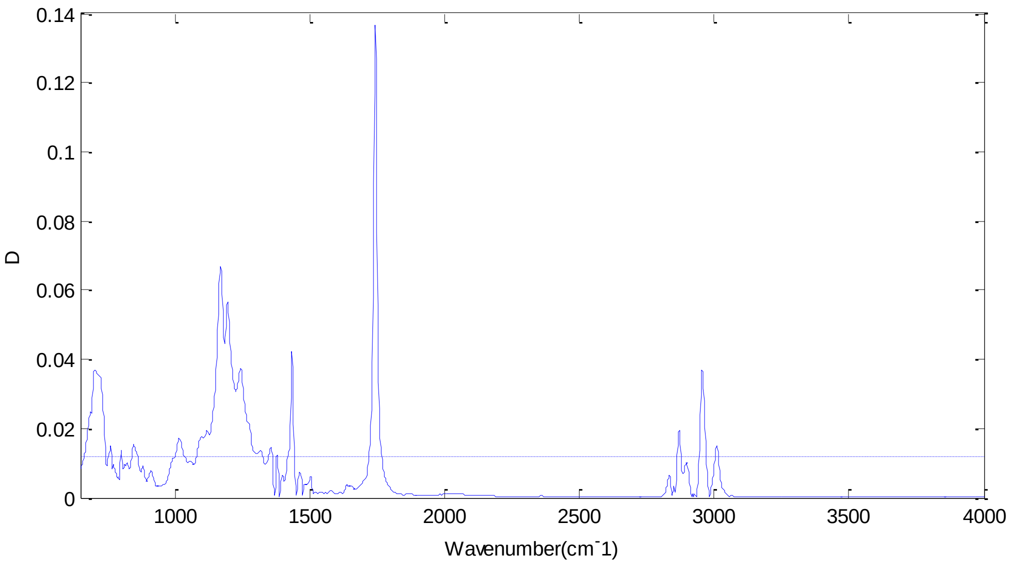

Figure 2: Values of D (solid line) for the biodiesel data set alongside the defined threshold (dashed line)

Fig 2 shows the difference D for the biodiesel data set, computed according equation 13, alongside the threshold defined in step 4 as twice the mean of the values in D.

The selection approach based on all combinations of 1, 2, 3 and 4 intervals presented the best results, achieving the maximum accuracy using kNN and LDA for classification. The optimal k was set to 3 and the number of principal components was also optimal at the value of 3. For the LDA the best number of components found was 4, also achieving 0% of misclassification rate on the forward selection approach. Although PNN didn't achieve zero misclassification, the selection procedure yielded a remarkably improvement in the accuracy, with the misclassification falling from 8.24% to 1.18%.

Table 1: Accuracy (misclassification rate) after wavenumber selection for the biodiesel data set

Misclassification rate (MR)						
	Forward Selection		Backward Elimination		All combinations	
kNN	1.18%	(+- 0.12)	1.18%	(+- 0.12)	0%	(+- 0.00)
LDA	0%	(+- 0.00)	3.53%	(+- 0.31)	0%	(+- 0.00)
PNN	1.18%	(+- 0.12)	1.18%	(+- 0.12)	1.18%	(+- 0.12)

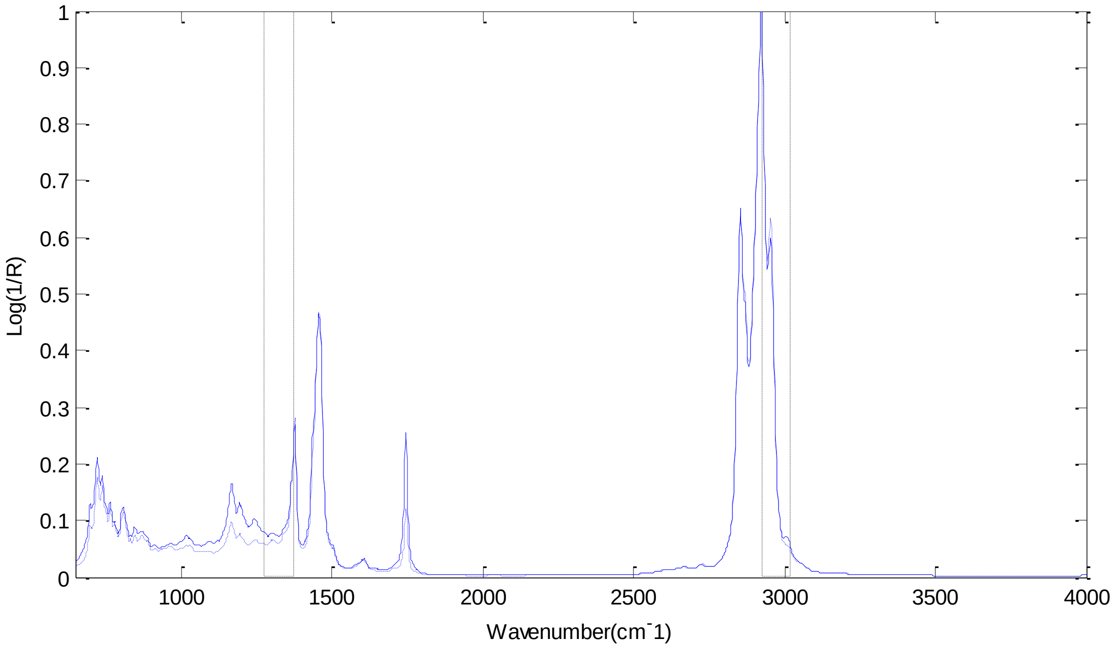

Figure 3: Average spectra of both classes (solid and dashed lines) and the two spectral bands selected by kNN and LDA in the biodiesel data set represented by the dashed vertical lines.

Selected spectral bands by kNN are in the range of 1275 – 1373 cm^{-1}, while the ones selected by LDA are around 2933 -2984 cm^{-1}. Such results that are consistent with Ferrão *et al.* [31] findings for sulphur content for the same data. Fig 3 shows the average spectra of the two classes alongside the spectral bands.

In the case of kNN, the total number of variables after the selection was 48 out of 1738, meaning that the method was able to classify correctly all samples using 2.71% of the whole spectra, proving to be suitable for the purpose of this data set.

4.3 Diesel fuel data set results

The procedure proposed in section 3.3 was performed for the data set and the misclassification rates for the 10-fold CV are presented in Table 2. The profile of the curve D for this data set is shown in Fig 4 with the threshold set as the mean of D.

The best MR found is 10.36% corresponding to all combinations approach with kNN as classification tool, where the optimized number of PCs is 28 and the number of neighbors is 32, leading to a selection of 3 intervals. The intervals are located in two different spectral bands, in the range of 922 – 996 nm and 1240 – 1270 nm, as depicted in Fig 5.

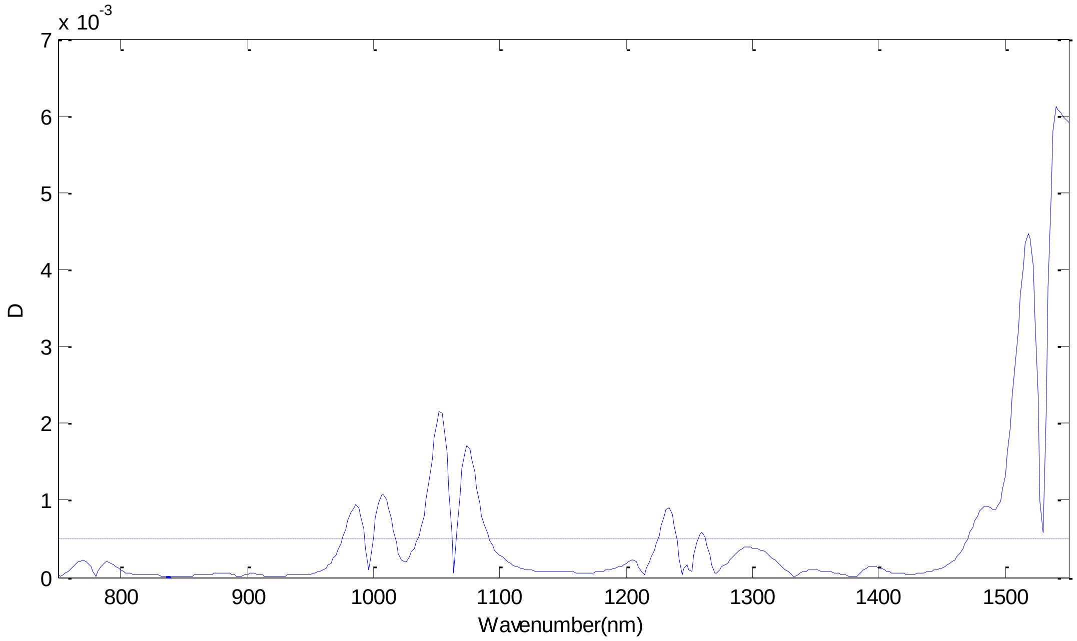

Figure 4: Values of D (solid line) for the diesel data set alongside the defined threshold (dashed line)

Misclassification before selection process was 11.71% for kNN, 13.06% for PNN and 16.67% for LDA, being the method able to achieve an improvement of 11.52% compared to the best result when no selection is applied.

Table 2: Accuracy (misclassification rate) after wavenumber selection for the diesel data set

Misclassification rate (MR)						
	Forward Selection		Backward Elimination		All combinations	
kNN	10.81%	(+- 1.31)	10.81%	(+- 1.31)	10.36%	(+- 1.10)
LDA	15.77%	(+- 1.66)	15.77%	(+- 1.66)	15.77%	(+- 1.66)
PNN	12.61%	(+- 1.54)	12.61%	(+- 1.54)	12.61%	(+- 1.54)

The best accuracy was found using 64 out of 401 of the wavenumbers, representing just 16% of all the variables. Although the misclassification rate did not have a large decrease as in the previous data set, the method still were able to improve it at the same time that eliminated wavenumbers that did not contribute to the model, being able to better classify the samples in standard and premium diesel classes.

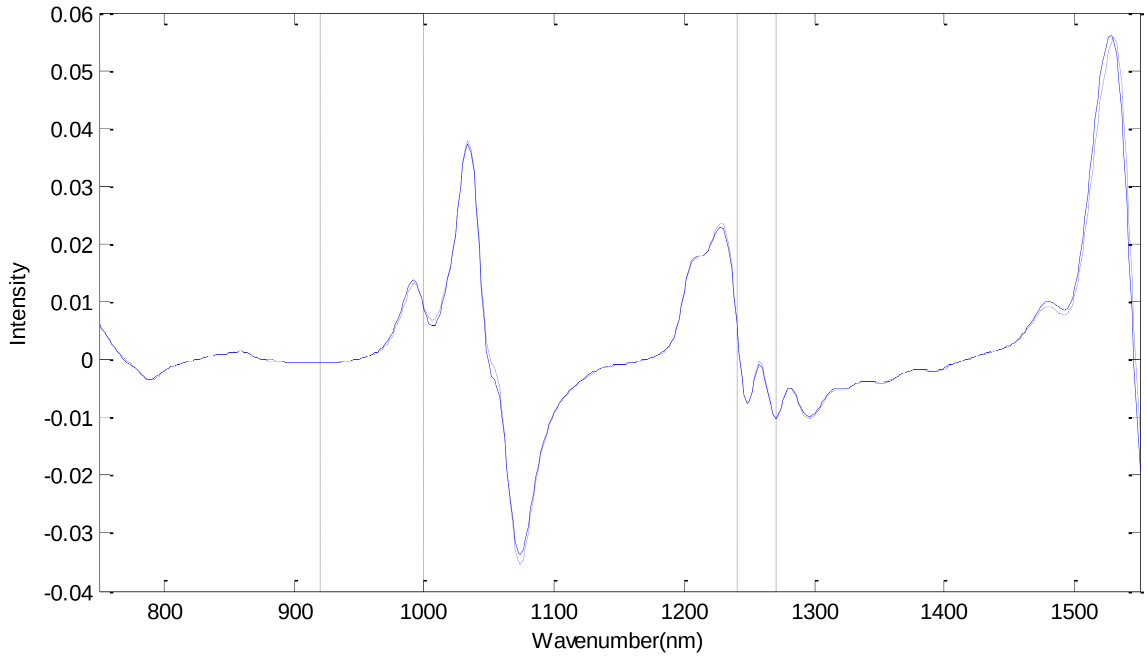

Figure 5: Average spectra of both classes (solid and dashed line) and the two spectral bands in the diesel data set represented by the dashed vertical lines.

5. Conclusion

Facing the dimensionality issue that usually appear in many classification tasks involving spectroscopy, this paper presents a novel method of wavenumber selection for classification purpose based on interval selection. The method resembles the iPLS [30], although it can select intervals of different size, as the limits of the intervals depends on the distance between the average spectra of each class.

The method was applied to two different data sets, combining three classification tools, namely kNN, LDA and PNN, and three selection approaches, forward selection, backward selection and all possible combinations of 1, 2, 3 and 4 intervals. The results proved the method to be robust, leading to improvements in the accuracy of classification for both data sets while reducing the percent of wavenumbers retained in the model.

Further research is needed to adapt the method to tasks with more than two classes, also to extend it to prediction tasks, which are the most common problems faced in analytical chemistry.

References:

[1] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy., Anal. Chim. Acta. 667 (2010) 14–32. doi:10.1016/j.aca.2010.03.048.

[2] R.M. Balabin, S. V. Smirnov, Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, Anal. Chim. Acta. 692 (2011) 63–72. doi:10.1016/j.aca.2011.03.006.

[3] O.Y. Rodionova, a. L. Pomerantsev, NIR-based approach to counterfeit-drug detection, TrAC Trends Anal. Chem. 29 (2010) 795–803. doi:10.1016/j.trac.2010.05.004.

[4] C.M. Bishop, Pattern Recognition and Machine Learning , Springer, 2006.

[5] A.R. Webb, K.D. Copsey, Statistical Pattern Recognition, 3rd ed., Wiley, 2011.

[6] W. Krzanowski, Selection of variables to preserve multivariate data structure, using principal components, Appl. Stat. 36 (1987) 22–33.

[7] I. Jolliffe, Discarding variables in a principal component analysis. I: Artificial data, Appl. Stat. 21 (1972) 160–173.

[8] I.M. Jolliffe, Principal Component Analysis, Springer, 2002.

[9] M.J. Anzanello, R.S. Ortiz, R.P. Limbergerb, P. Mayorga, A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes., J. Pharm. Biomed. Anal. 83 (2013) 209–14. doi:10.1016/j.jpba.2013.05.004.

[10] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., Wiley, 2000.

[11] R. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. (1936).

[12] J. Zhao, P. Yu, L. Shi, S. Li, Separable linear discriminant analysis, Comput. Stat. Data Anal. (2012).

[13] R. Balabin, R. Safieva, E. Lomakina, Gasoline classification using near infrared (NIR) spectroscopy data: comparison of multivariate techniques, Anal. Chim. Acta. 671 (2010) 27–35. doi:10.1016/j.aca.2010.05.013.

[14] K. Fukunaga, Introduction to statistical pattern recognition, 2nd ed., Academic Press, 1990.

[15] E. Fix, J.H. Jr, Discriminatory analysis-nonparametric discrimination: consistency properties, 57 (1951) 238–247.

[16] K. Fukunaga, P. Narendra, A branch and bound algorithm for computing k-nearest neighbors, Comput. IEEE Trans. (1975) 750–753.

[17] S. Tan, Neighbor-weighted K-nearest neighbor for unbalanced text corpus, Expert Syst. Appl. 28 (2005) 667–671. doi:10.1016/j.eswa.2004.12.023.

[18] X. Sun, C.M. Zimmermann, G.P. Jackson, C.E. Bunker, P.B. Harrington, Classification of jet fuels by fuzzy rule-building expert systems applied to three-way data by fast gas chromatography--fast scanning quadrupole ion trap mass spectrometry., Talanta. 83 (2011) 1260–8. doi:10.1016/j.talanta.2010.05.063.

[19] J. Keller, M. Gray, J. Givens, A fuzzy k-nearest neighbor algorithm, Syst. Man Cybern. …. (1985) 580–585.

[20] D. Specht, Probabilistic neural networks, Neural Networks. 3 (1990).

[21] M. Hajmeer, I. Basheer, A probabilistic neural network approach for modeling and classification of bacterial growth/no-growth data, J. Microbiol. Methods. 51 (2002) 217–226. doi:10.1016/S0167-7012(02)00080-5.

[22] L. Rutkowski, S. Member, Adaptive Probabilistic Neural Networks for Pattern Classification in Time-Varying Environment, 15 (2004) 811–827.

[23] K.Z. Mao, K.C. Tan, W. Ser, Probabilistic neural-network structure determination for pattern classification., IEEE Trans. Neural Netw. 11 (2000) 1009–16. doi:10.1109/72.857781.

[24] M. Dyrby, S.B. Engelsen, L. Nørgaard, M. Bruhn, L. Lundsberg-Nielsen, Chemometric Quantitation of the Active Substance (Containing C≡N) in a Pharmaceutical Tablet Using Near-Infrared (NIR) Transmittance and NIR FT-Raman Spectra, Appl. Spectrosc. 56 (2002) 579–585. doi:10.1366/0003702021955358.

[25] M. Fernanda Pimentel, G.M.G.S. Ribeiro, R.S. da Cruz, L. Stragevitch, J.G. a. Pacheco Filho, L.S.G. Teixeira, Determination of biodiesel content when blended with mineral diesel fuel using infrared spectroscopy and multivariate calibration, Microchem. J. 82 (2006) 201–206. doi:10.1016/j.microc.2006.01.019.

[26] M.J.C. Pontes, C.F. Pereira, M.F. Pimentel, F.V.C. Vasconcelos, A.G.B. Silva, Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectrometry and multivariate classification., Talanta. 85 (2011) 2159–65. doi:10.1016/j.talanta.2011.07.064.

[27] S. Li, L. Dai, Classification of gasoline brand and origin by Raman spectroscopy and a novel R-weighted LSSVM algorithm, Fuel. 96 (2012) 146–152. doi:10.1016/j.fuel.2012.01.001.

[28] R.M. Balabin, R.Z. Safieva, Gasoline classification by source and type based on near infrared (NIR) spectroscopy data, Fuel. 87 (2008) 1096–1101. doi:10.1016/j.fuel.2007.07.018.

[29] M. Kim, Y. Lee, C. Han, Real-time classification of petroleum products using near-infrared spectra, Comput. Chem. Eng. 24 (2000) 513–517.

[30] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, … Spectrosc. 54 (2000) 413–419.

[31] M.F. Ferrão, M.D.S. Viera, R.E.P. Pazos, D. Fachini, A.E. Gerbase, L. Marder, Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions, Fuel. 90 (2011) 701–706. doi:10.1016/j.fuel.2010.09.016.

[32] Petrobras Distribuidora, Produtos automotivos - oleo diesel, (n.d.). http://www.br.com.br/wps/portal/portalconteudo/produtos/automotivos/oleodiesel (accessed November 08, 2014).