New chemometric methods for solving classification problems in NIR spectroscopy

Max Egebo[1], Åsmund Rinnan[2], Søren Balling Engelsen[2], Rasmus Bro[2] and Lars Nørgaard[1]*

[1]FOSS Analytical, Team Chemometric Development, Hillerød, DK-3400, Denmark
[2]Faculty of Life Science, Department of Food Science, Frederiksberg, DK-1958, Denmark
*Corresponding author: lno@foss.dk

Introduction

The extremely high correlation between variables that is an inherent feature of near infrared (NIR) spectroscopic data makes multivariate projection methods suitable as data analytical tools. Projection-based methods such as principal component analysis (PCA)[1], partial least squares (PLS)[2] regression and PLS-discriminant analysis (PLS-DA)[3] or soft independent modelling of class analogy (SIMCA)[4] are therefore often used for data exploration, regression and classification respectively.

In classical statistics, the canonical variates analysis (CVA) method was introduced in the 1930s by Fisher[5-6] in order to optimise the group discriminative direction in multivariate space. CVA is an excellent method for estimating discriminative directions, but in contrast to e.g. PCA, the CVA method breaks down if the number of variables is larger than the number of samples and/or if the variables are highly correlated. In this respect, CVA is comparable to multiple linear regression (MLR), which suffers the same limitations.

A modification of CVA called extended canonical variates analysis (ECVA)[7] was recently developed that handles highly correlated data and here we will demonstrate the efficiency of the method,.

Interval PLS (*i*PLS)[8] has proven to be an efficient tool for exploration of all sorts of spectroscopic data including NIR. The power of *i*PLS is the graphical output clearly illustrating to the spectroscopist where interesting spectroscopic areas exists for the problem in question. ECVA is introduced in the interval version (*i*ECVA) as an exploratory tool for spectroscopic interpretation and model optimisation, and it is our hope that it will be as beneficial as the *i*PLS method. With respect to *i*PLS, it should be stressed that it was originally intended as an exploratory tool for spectroscopic or other highly co-linear data structures. It was not intended as a stand-alone variable selection technique since such selection can be performed in a more elegant and optimal way by a multitude of other methods. The same comment holds for the *i*ECVA method.

The ECVA and *i*ECVA methods are applied on a data set compiled in a large study to evaluate vis-NIR spectroscopy for the assessment of the depth of CO_2 stunning of slaughter pigs. Several parameters such as corneal reflex, breathing and convulsions were analysed to assess the depth of the stunning but in this study we will solely focus on a multivariate analysis of the vis-NIR spectra to investigate if there are systematic discriminative patterns among the three slaughterhouses included.

Materials and Methods

Samples and NIR instrumentation

The data set analysed consisted of 162 blood samples from pigs from three commercial slaughterhouses, designated S1, S2 and S3, with known differences in stunning time and concentration of CO_2 stunning gas.

During exsanguinations 1-2 L of blood from each animal was collected in a container. A 0.5 L wash bottle, containing 5 mL 10% EDTA and 10% anticoagulant, was immediately filled with the blood.

The samples were analysed in the wavelength range 400 to 2498 nm (every 2 nm corresponding to 1050 variables) by a reflectance measurement in a 3 mm cuvette on NIRSystems 6500 spectrophotometer (FOSS NIRSystems, Inc., Silver Spring, Maryland, USA fitted with a transport module (NR-6511)) filled with the blood sample. An average of 32 scans was used for each sample. The measurements were carried out over three consecutive days with the same instrument and operator.

Slaughterhouse 3 (S3) used the longest stunning time and the highest CO_2 concentration while stunning time was shortest and CO_2 concentration lowest at S1.

Chemometric methods

The ECVA method is presented in detail in Nørgaard *et al.*[6] while *i*ECVA was introduced in a cancer diagnostics feasibility study based on fluorescence spectroscopy[9]; the reader is referred to these papers for an in-depth description.

Reference paper as:
Egebo, M., Rinnan, Å., Engelsen, S.B., Bro, R. and Nørgaard, L. (2012). New chemometric methods for solving classification problems in NIR spectroscopy, in: Proceedings of the 15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 87-90.

The ECVA and *i*ECVA methods are implemented in MATLAB (The MathWorks, Inc., Massachusetts, USA) and made freely available for non-commercial use at www.models.life.ku.dk. The present toolbox carries the version number 2.5.

Data pre-processing

The data were analysed using a) only mean centering[10] and b) 2nd derivative pre-treated data followed by mean-centering. The 2nd derivative was calculated according to a Savitzky-Golay algorithm[11] with an eleven point window width using a 2nd order polynomial.

Model validation

Four samples (one from S2 and three from S3) were excluded from further analysis due to clearly deviating outlying spectral features.

A calibration set consisting of 118 samples (selected by excluding every 4th sample) and a model independent test set consisting of 40 samples (every 4th sample of the original 158 samples without outliers) were constructed. For the calibration set, ten segment Venetian blinds cross-validation[12] was used for estimating the number of inner PLS components in the ECVA method. This number was used when predicting the test set samples class affiliation.

The calibration set contained thirty-nine S1, forty-two S2 and thirty-seven S3 samples while the test set contained fourteen S1, fourteen S2 and twelve S3 samples.

Results and Discussion

The mean-centred spectroscopic data for all 158 samples are shown in Figure 1. Spectral peaks are observed at 440 and 550 nm in the visual part of the spectra; these peaks are caused by the oxygenated form of haemoglobin[13] while the peak at 760 nm is related to deoxygenated haemoglobin[14-15]. The O-H stretching and bending from water causes the broad peaks at 970 nm (O-H stretch, second overtone), 1190 nm (O-H stretch and bend, combination tone), 1450 nm (O H stretch, first overtone) and 1940 nm (O-H stretch and bend, combination tone).

A PCA was calculated on the mean-centered data set and the score plot of principal component 1 versus principal component 2 (Figure 2) illustrates that S3 is more distinct compared to S1 and S2 that are quite overlapped. The two first components explain 37.8% and 34.0% of the variance respectively; relevant information was available also from higher principal components but no other combination showed a larger degree of discrimination between groups than the first two components.

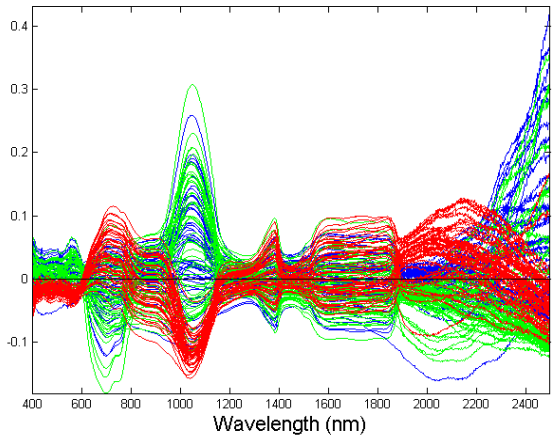

Figure 1. Mean-centred vis-NIR spectra (400 to 2500 nm) of 158 blood samples from slaughterhouses S1 (blue), S2 (green) and S3 (red).

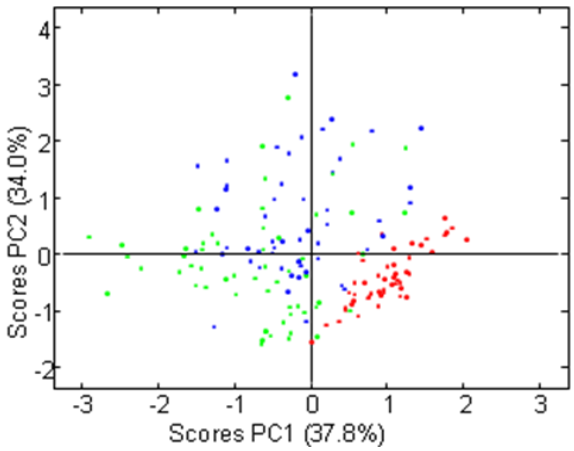

Figure 2. Principal component analysis score plot based on the mean-centred data presented in Figure 1. Slaughterhouse S1 (blue), S2 (green) and S3 (red).

The next step was to focus directly on discrimination between the groups by calculating an ECVA model based on the full spectrum mean-centered data. The model has a minimum number of misclassifications using nineteen PLS components with two cross-validated classification errors. Application of the developed model on the independent test set consisting of 40 samples, resulted in two misclassifications.

Figure 3 illustrates the extended canonical variates plot for component 1 versus 2. A very clear separation between the three slaughterhouses is observed compared to the PCA scope plot (Figure 2). ECVA focuses on the discriminative direction in multivariate space and the extended canonical weights (corresponding to

Reference paper as:
Egebo, M., Rinnan, Å., Engelsen, S.B., Bro, R. and Nørgaard, L. (2012). New chemometric methods for solving classification problems in NIR spectroscopy, in: Proceedings of the 15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 87-90.

loadings in PCA) provide information on spectral ranges relevant for discrimination. As seen in Figure 4, the range from 650 to 1400 nm contains systematic variation in both canonical weights while the low and the high wavelength ranges are noisy. Utilising the properties of the ECVA there is no need for compression or variable selection of the spectral data *before* the analysis since the ECVA is capable of handling highly collinear data. This capability does not guarantee a good model, but the mathematics will not break down.

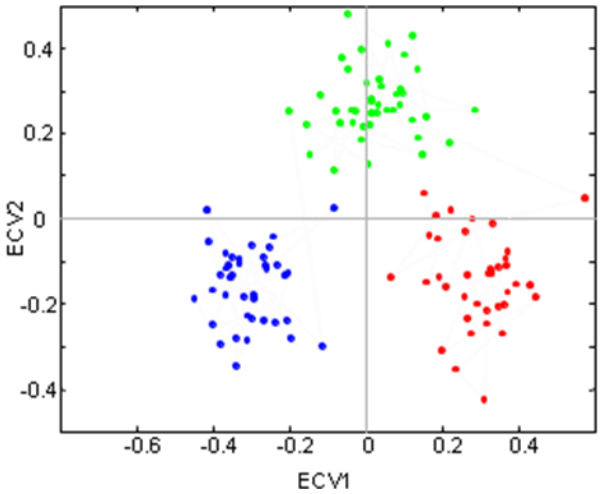

Figure 3. Extended canonical variate number 1 versus number 2. Slaughterhouse S1 (blue), S2 (green) and S3 (red).

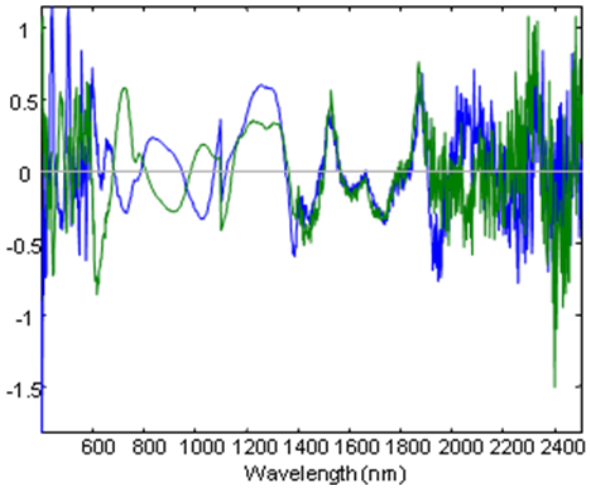

Figure 4. Extended canonical weights number 1 (blue) and 2 (green).

In order to explore approximately which spectral ranges contain discriminative information, an *i*ECVA is calculated on the mean-centered data with 20 spectral intervals. The number 20 could easily be changed according to the problem analysed but it seems a reasonable number in the present study taking into account the variation in the data and the spectral range.

The number of misclassifications is given for each interval by the height of the bars (Figure 5). No single interval outperforms utilising the full spectrum for classification but is it clear that intervals 2, 9 and 11, corresponding to wavelength ranges 506-610 nm, 1248-1352 nm and 1460-1562 nm, are relevant for discriminating between the three slaughterhouses.

Suitable data pre-processing can improve performance in both regression and classification problems based on NIR spectroscopy. In this case, the 2nd derivative is used as a pre-treatment method and an *i*ECVA is calculated on the pre-processed data (Figure 6). The full spectrum classification error is very high (33), due to the noisy ranges at low and high wavelengths. Two intervals provide a classifier with zero misclassifications: intervals 4 and 5 corresponding to 718-822 nm and 824-928 nm respectively.

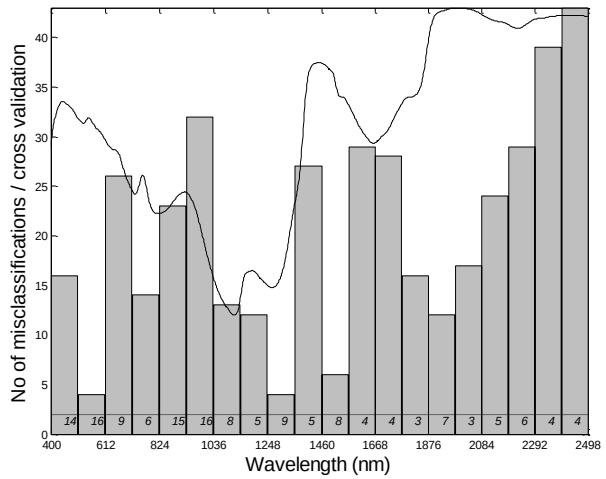

Figure 5. Interval ECVA with 20 intervals on the mean centred spectroscopic data. The spectrum is the average over all 118 calibration samples. Numbers on the bars are the number of inner components in the ECVA model.

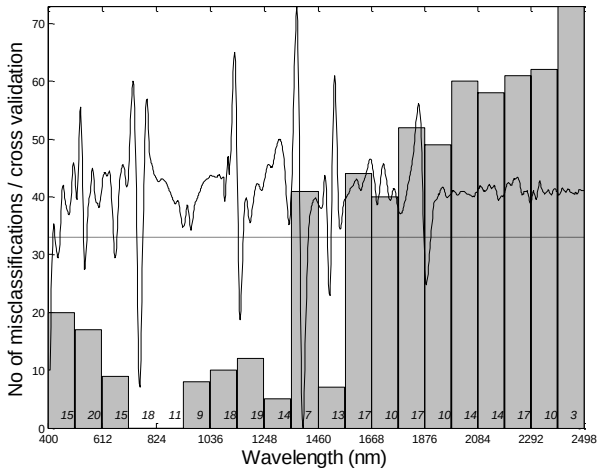

Figure 6. Interval ECVA with 20 intervals on the mean centred and 2nd derivative pre-treated spectroscopic data. The spectrum is the average over all 118 calibration samples.

Reference paper as:
Egebo, M., Rinnan, Å., Engelsen, S.B., Bro, R. and Nørgaard, L. (2012). New chemometric methods for solving classification problems in NIR spectroscopy, in: Proceedings of the 15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 87-90.

The 2nd derivative pre-treatment improved the misclassification error of the calibration set and caused a different spectral range to emerge as the most powerful range for discrimination. The test set misclassification errors based on the two individual models were 0 and 1 respectively (Table 1).

Table 1. Misclassifications for selected ECVA models. All models are mean-centered.

	Sample set	
	Calibration	Test
Number of samples	118	40
Full spectrum	2	2
Full spectrum 2nd derivative	33	12
718-822 nm 2nd derivative	0	0
824-928 nm 2nd derivative	0	1

Conclusion

Extended Canonical Variates Analysis and the interval version of the same method are shown to be relevant and robust alternatives for solving NIR spectroscopy classification problems based on complex food, feed or pharmaceutical sample matrices. The methods are useful also for exploration, interpretation and identification of specific discriminative ranges in NIR spectra.

In this study, the focus was on indicating the efficiency of ECVA and *i*ECVA to help improve classification and interpretation; a more thorough study focusing on the chemical interpretation is under preparation.

*i*ECVA is not suggested as an optimal method variable selection but merely an exploratory tool to be used e.g. in combination with *i*PCA and *i*PLS for investigating the problem under study. The true essence of the word chemometrics lies in the synergistic combination of application knowledge combined with data analytical skills.

References

1. S. Wold, K. Esbensen and P. Geladi, *Chemometr. Intell. Lab. Syst.* **2,** 37-52 (1987).
2. D.M. Haaland and E.V. Thomas, *Anal. Chem.* **60,** 1193-1202 (1988).
3. S. Wold, *Pattern Recogn.* **8,** 127-139 (1976).
4. L. Ståhle and S. Wold, *J. Chemometr.* **1,** 185-196 (1987).
5. R.A. Fisher, *Ann. Eug.* **7**, 179-188 (1936).
6. W.J. Krzanowski, *Principles of Multivariate Analysis.* Revised edn., Oxford University Press, UK (2000).
7. L. Nørgaard, R. Bro, F. Westad and S.B. Engelsen, *J. Chemometr.* **20**, 425-435 (2006).
8. L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck and S.B. Engelsen, *Appl. Spectrosc.* **54**, 413-419 (2000).
9. L. Nørgaard, G. Soletormos, N. Harrit, M. Albrechtsen, O. Olsen, D. Nielsen, K. Kampmann and R. Bro, *J. Chemometr.* **21**, 451-458 (2007).
10. R. Bro and A.K. Smilde, *J. Chemometr.* **17**, 16-33 (2003).
11. A. Savitzky and M. J. E. Golay, *Anal. Chem.* **36**, 1627-1632 (1964).
12. M. Stone, *J. Roy, Stat. Soc. B* **36**, 111-147 (1974).
13. Y.J. Kim, S. Kim, J.W. Kim and G. Yoon, *J. Biomed. Optic.* **6**, 177-182 (2001).
14. S. Wray, M. Cope, D.T. Delpy, J.S. Wyatt and E. O.R. Reynolds, *Biochim. Biophys. Acta* **933**, 184-192 (1988).
15. D. Baykut, M.M. Gebhard, H. Bölükoglu, K. Kadipasaoglu, S. Hennes, O.H. Frazier and A. Krian, *Thorac. Cardiovasc. Surg.* **49**, 162-166 (2001).

Reference paper as:
Egebo, M., Rinnan, Å., Engelsen, S.B., Bro, R. and Nørgaard, L. (2012). New chemometric methods for solving classification problems in NIR spectroscopy, in: Proceedings of the 15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 87-90.