Wavelength selection for the multivariate calibration of near infrared spectroscopic data

D. Jouan-Rimbaud and D.L. Massart

ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussels, Belgium.

Introduction

Near infrared (NIR) spectroscopy is a widely used technique in many different industrial fields. The popularity of this technique is related to its rapidity and precision. The fact that no, or little, sample pretreatment is required is another advantage of this technique, as (i) its is non-destructive, (ii) it presents some environmental benefits (e.g. no use of toxic solvent) and (iii) less sources of errors are introduced. However, as the spectra represent overlapped information of diverse origins, NIR spectra are quite difficult to interpret and the quantitative analysis is also not straightforward. The use of statistical methods, such as principal component analysis (PCA),[1] which decomposes the spectral information onto a few orthogonal variables (called "latent variables") is quite helpful for the interpretation of the spectra. As far as calibration is concerned, methods such as principal component regression (PCR)[1] or partial least squares (PLS)[1] are employed.

However powerful these methods may be, they are not easy, as the chemical interpretation is performed by the study of loading plots, which is not straightforward. For this first reason, multiple linear regression (MLR) could be a favoured method as it involves the modelling of a few wavelengths only, directly related to chemical information. Furthermore, MLR is simple to perform, which is another advantage. For some mathematical reasons, in inverse MLR (the concentration is modelled as a function of the absorbances measured at a few wavelengths) the number of calibration samples should be larger than the number of wavelengths used in the spectra. This is very difficult to achieve when the spectra are given for hundreds of wavelengths: MLR modelling requires the pre-selection of a subset of wavelengths.

In this paper, we present a few methods of wavelength selection which have been applied to two data sets from industry.

The data sets

Data set 1 comes from the pharmaceutical industry and its study was presented in a previous paper.[2] In 11 samples from production batches, the concentration of the active ingredient which is to be determined varies in a very narrow range, so 13 synthetic samples were produced, with an extended concentration range of the active ingredient. The data set is quite heterogeneous (Figure 1). Second-derivative spectra of log (1/*R*) are used for calibration, measured in the range 400–2500 nm. The 24 spectra were separated into two heterogeneous sets, one for calibration (14 samples) and one for prediction (10 samples).

Data set 2 comes from the oil industry. Spectra of 87 samples of polyether polyols were recorded. The data are separated into two main clusters, depending on the chemical structure of

From Near Infrared Spectroscopy: The Future Waves

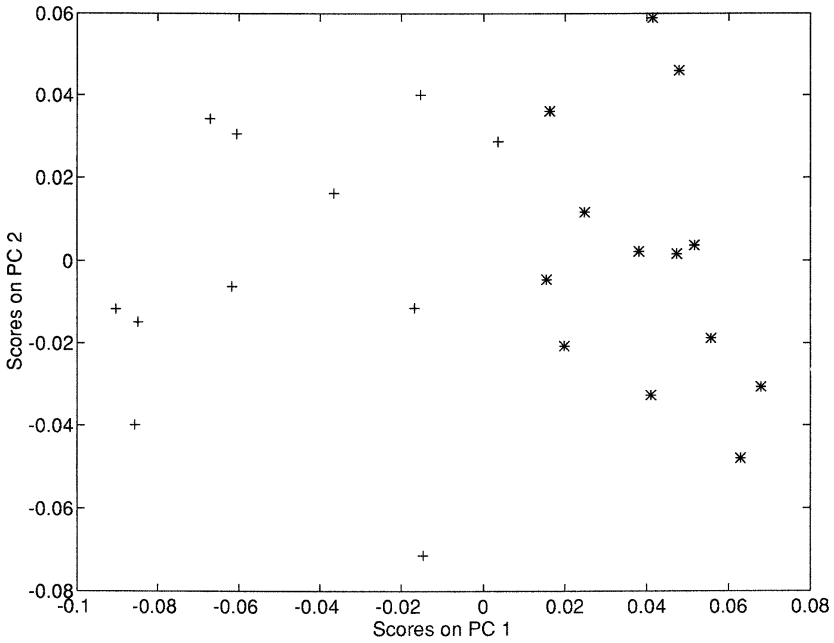

Figure 1. PC1-PC2 score plot of the spectral data (* synthetic samples, + production samples).

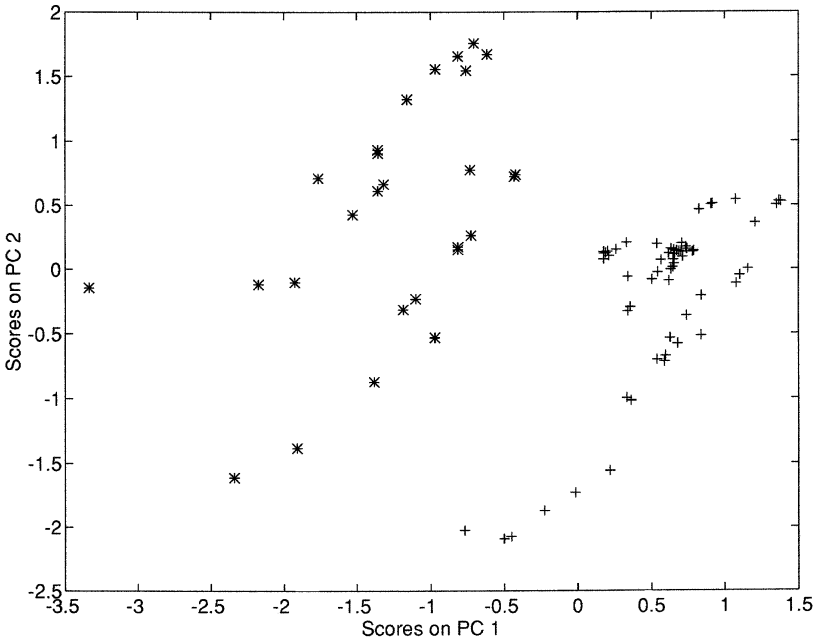

Figure 2. PC1-PC2 score plot of the spectral data.

From Near Infrared Spectroscopy: The Future Waves

the data, but smaller sub-clusters can be distinguished (Figure 2). The calibration aims at relating the hydroxyl number of the samples to their NIR spectra.

Wavelength selection and MLR modelling

Stepwise selection[3] was applied to data set 1. As too many wavelengths were present in the spectra (1050), a pre-selection of wavelengths took place:

- The 100 wavelengths with highest correlation with the active ingredient's concentration were selected. The stepwise selection applied on these 100 wavelengths retained two variables, which constitute subset 1.
- The stepwise selection was applied on the 100 wavelengths with highest covariance with concentration and a two-wavelengths subset was found (subset 2), different from subset 1.

A PCR with selection of PCs was performed, and only PC 3 and 6 were selected.[2] The wavelengths with high loadings on these PCs were selected, but they were too numerous to enable an inverse MLR modelling, so two methods of selection among the 31 pre-selected wavelengths were applied:

- selection of the most correlated wavelengths (subset 3).
- stepwise selection (subset 4): two wavelengths were selected, different from those in subset 1 and 2.

MLR models were built with the calibration set and the different subsets of wavelengths and the prediction set was predicted from these models. The prediction error is estimated by the root mean square error of prediction (*RMSEP*) and the results are shown in Table 1.

The *RMSEP* is defined as:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

From this table, we conclude that MLR on very few wavelengths can perform as well as PCR or PLS on the whole wavelength range.

Yet, for data set 2, these methods did not succeed to select subsets with which MLR models had a satisfactory predictive ability. Indeed, this data set is quite complex and probably requires a more powerful tool to select the wavelengths. Genetic algorithms have already been used successfully in feature selection,[4,5] and particularly on spectroscopic data.[6,7]

Table 1. *RMSEP* for the different MLR models compared to PCR or PLS.

Model	*RMSEP*
PCR (6 PCs)	0.3893
PLS (4 factors)	0.4527
MLR with subset 1	0.2911
MLR with subset 2	0.4817
MLR with subset 3	0.3810
MLR with subset 4	0.5277

From Near Infrared Spectroscopy: The Future Waves

Table 2. Comparison of MLR with PLS.

Model	*RMSEP*
PLS (7 factors)	1.78
MLR (7 variables)	1.39

Applied to data set 2, we found a subset of seven variables, which yielded a model with a good predictive ability (estimated by cross-validation). The comparison of this model with PLS is shown in Table 2.

The first remark was that the dimensionality of MLR and PLS was the same (seven variables). Furthermore, it seems MLR performs better than PLS.

A forward selection was applied to the 7-wavelengths subset,[7] so that the seven wavelengths were ordered. Comparison of 1- to 7- variables MLR model with 1- to 7-factors PLS model was performed and showed that the role of the first variables in the MLR model is similar to the role of the first factors in PLS.

Conclusion

Provided an appropriate method of wavelength selection is used, it is possible to perform the calibration of near infrared data by means of multiple linear regression. Depending on the data set, more or less simple methods of wavelength selection can be used. Even when the data set is very complex, a subset of very few wavelengths can yield satisfactory results. The advantages of MLR are its simplicity and its ease of interpretation. Yet, one drawback of MLR is that it may be less robust than PLS, for example if a shift occurs in some future spectra to predict, so care should be given on this aspect.

References

1. H. Martens and T. Næs, *Multivariate Calibration.* John Wiley & Sons, Chichester (1991).
2. D. Jouan-Rimbaud, B. Walczak, D.L. Massart, I.R. Last and K.A. Prebble, *Analytica Chimica Acta* **304,** 285 (1995).
3. N.R. Draper and H. Smith, *Applied Regression Analysis,* 2nd Edition. John Wiley and Sons, New York (1981).
4. R. Leardi, R. Boggia and M. Terrile, *Journal of Chemometrics* **6,** 267 (1992).
5. R. Leardi, *Journal of Chemometrics* **8,** 65 (1994).
6. C.B. Lucasius, M.L. Beckers and G. Kateman, *Analytica Chimica Acta* **286,** 135 (1994).
7. D. Jouan-Rimbaud, D.L. Massart, R. Leardi and O.E. de Noord, *Anal. Chem.* **67,** 4295 (1995).

From Near Infrared Spectroscopy: The Future Waves