SELECTION OF A MULTIVARIATE CALIBRATION METHOD

0. Aim of this document

Different types of multivariate calibration methods are available. The aim of this document is to help the user select the proper approach on a rational basis. Of the many methods available, a set of six methods was chosen. Depending on the data structure some of these methods should be preferred to others.

1. Methods to be considered for multivariate calibration

Many methods for multivariate calibration have been proposed. It turns out that many of the methods perform similarly. To avoid confusion due to use of many different methods, it is suggested that only the following should be considered:
 - Multiple linear regression (MLR)
 - Principal component regression (PCR)
 - Partial least squares (PLS)
 - Neural networks (NN)
 - Locally weighted regression (LWR)
 - Radial basis functions combined with PLS (RBF-PLS)
They have been included on the basis of theoretical considerations, confirmed by an intercomparison of their performances carried out on near infrared data sets with different data structures. Their main advantages and disadvantages are described in the following sections

2. Data and data structures

2.1 Data:

There are n samples which were spectroscopically measured at p wavelengths. The data are present in a nxp data matrix $\mathbf{X}$. Moreover each sample is characterized by the values of the concentration or other characteristic, y. These form together the $\mathbf{y}$ vector. The object of the multivariate calibration is to develop a model $\mathbf{y} = f(\mathbf{X})$, which can be used to predict the y-values of a new sample.

 2.2. Data structures.

The following terms are used to describe data structures:
- When the density of the data points, either in the y or the X space is not unimodal, so that more than one distinct group of data can be seen, then the data are said to be heterogeneous or clustered. Otherwise they are called homogeneous. Clusters contain objects that are more similar to each other than to objects outside of the group and, therefore, one may decide to make separate models for each of the clusters. If one prefers a single model for all data together, then it may be interesting to apply a so-called local modeling technique. This starts out with all data points but gives higher weights in the model to those data that are closest to the data point which is being predicted, so that members of the same cluster are more important in the prediction than samples belonging to other clusters. A model that gives equal weight to all samples is called global. Of the methods decribed higher, LWR and RBF-PLS are considered to be local methods, MLR, PCR, PLS and NN are global methods. Local models should be avoided when the number of calibration samples is small. Since only some samples are given high weights (or are

used at all), the number of samples that has an influence on the model may be simply too small
- When the relationship between **y** and some or all of the **X** variables is non linear, then the data are said to show <u>non linearity</u>. PLS and PCR are able to some extent to model non linear behaviour. NN are called universal approximators, which means that they are able to model also strongly non linear behaviour. Non linear relationships can be locally approximated by linear models, so that local methods such as LWR and RBF-PLS can be applied also for non linear data.
- Outliers in the y or X space are a problem for all multivariate calibration models. What to do with them is not discussed here, but should be considered for each of the methods separately. At the selection stage, when the first preliminary investigation of the data (see further) is carried out, sometimes samples with very extreme characteristics may be detected. Such data points constitute <u>gross outliers</u> and they are often due to errrors. These could influence the conclusions to such an extent that one needs to decide whether such a sample should be retained. It is possible for instance that all data except the extreme one show linear behaviour, but that on the basis of the gross outlier one would conclude that there is non linearity. The decision to retain the sample must be made on the basis of practicality and of chemical reasoning. It should be clear however, that the region in data space described by the gross outlier is described by a single data point. It may then be preferable to limit the prediction range so that the gross outlier is not relevant anymore.
- In principle one should not make predictions for data that are not inside the <u>calibration domain</u>, i.e. concentrations predicted should fall within the concentration range (the **y** calibration domain) defined by the calibration samples. If the model is correct this also means that the sample to be predicted will show a signal, e.g. absorbance, within the signal range (the **X** calibration domain) defined by the calibration samples. In multivariate calibration, this is not always easy to ascertain. It is for instance possible that a sample within the **y** calibration domain includes a source of spectral variation, not present in the calibration samples, so that it is outside the **X** calibration domain. In such cases, one may be required to <u>extrapolate</u> the model outside the calibration domain. This should be avoided if at all possible, but this is not always possible. Some methods are less robust towards extrapolation than others.

2.3. Preliminary investigation of data structure

When contemplating the development of a multivariate calibration model, one should first ask whether one should expect one of the data structures described above. For instance, when the data concern a few grades of the same product and the difference between grades are larger than within grades, then clusters should be expected. Additionally one should decide whether extrapolation is allowed or expected and inquire about the number of samples that will be available to construct the model.

The following plots can help in reaching conclusions about the data structure.
- a histogram of the y-values. This will make apparent clusters and gross outliers in **y**.
- an overlay plot of the spectra. This will make apparent gross outliers and clear clusters in **X**.

- score plots on the first principal components, e.g. PC1-PC2, PC2-PC3. This will provide additional insight in clustering or the presence of extreme data in **X**.
- a plot of the first few PC, e.g. PC1-PC5 or PLS components against **y**. Strong non linearities will probably show up on one of these plots. Some PC may not be relevant to the calibration. Therefore, the fact that no relation is seen for some PC should not surprise.

It is recommended that, even if one has already decided to use a specific method (for instance, because it is the only one included in the available software), these plots should be obtained.

It should be understood that at this stage no exhaustive search into the presence of non-linearities, clusters and outliers is carried out. Minor non-linearities, clustering and less gross outliers do not need to be detected at this stage: they can be handled once the method has been selected in ways that are appropriate for that method.

3. Selection of methods

3.1. MLR:

MLR is the method which is best known from a statistical point of view. The method is based on selection of variables. It should certainly be preferred when the selection of variables is simple, i.e. when some variables are rather selective for the compounds or characteristics being determined (Note only for the text for our partners, not for the guidelines: give Raman and UV examples). When MLR is used with strongly collinear variables, as is the case with spectroscopic variables, there is what might seem a statistical paradox. The regression coefficients for different selected collinear wavelengths have relatively little meaning for interpretation purposes, but the model performs well, both in calibration and in prediction, provided that the model is linear and that the prediction is performed within the calibration domain. In those conditions it is probably the method to be recommended.

3.2. PCR:

PCR is a well known method that, of the methods described, is closest to MLR. It is a two step procedure. In the first step, one determines principal components, that are linear combinations of the original variables. They can be considered as new variables that summarize in an optimal way the variation present in the spectra. Linear combinations of original variables, combined so that the combinations optimize a certain criterion (here maximize explained variation in the data) are also called latent variables. In the second step, MLR is applied to the newly obtained latent variables. When collinearity between original variables occurs, principal component plots often allow better interpretation of the variations observed in the data set than plots of original variables selected by MLR. As a modelling method, it is somewhat less performant than MLR when performing prediction within the calibration domain and when the model is indeed linear, but it is more reliable if extrapolation may be required. It is very comparable in performance to PLS, but it is less rapid. On the other hand, it is easier to understand and to explain. It is a linear method, but it is able to perform quite well for moderately non linear data. As MLR, it is a global method.

3.3. PLS:

PLS is the method that is used often used in multivariate calibration. It resembles PCR, but works in one step. It also determines latent variables, that are linear combinations of the original variables, but the criterion applied is maximal covariance between the y - values and the spectral variables. Because of this criterion the algorithm yields models by an iterative procedure which is perceived by the user as a single regression step. As explained above, it generally performs equally well as PCR and has the same properties towards difficult data structures. As stated in 3.2., it is faster, but less easy to explain.

3.4. NN

Only backpropagation networks are considered. NN start by making linear combinations of the original variables. These combinations are then made non linear through the application of transfer functions. It follows that the method performs better than the linear methods described in the preceding sections when there is a pronounced non linearity in the relation between $\mathbf{y}$ and $\mathbf{X}$. They can perform as well as the linear methods, when the data are linear. Their main disadvantage is that extrapolation outside the calibration domain can lead to very bad results. Additionally, it is more difficult to use NN for interpretation purposes.

3.5. LWR

Locally weighted regression applies PCR or PLS, combined with a weighting scheme such that calibration samples closest to the sample to be predicted are given higher weight. Many variants are possible and the variant that probably should be preferred is one which has PLS for all samples with equal weights as the limiting case. In that case, when the data are linear and not clustered, the local method becomes automatically a global method. LWR was shown to be very performant when analysing clustered or non linear data, but it has not been studied to the same extent as the preceding methods, so that less diagnostics are available and that it is known less well in which cases the method should be preferred and when it should be avoided. One disadvantage is that it is more time consuming than other methods, because it requires that several sets of latent variables be determined. There also are indications for instance that it is less robust towards wavelength shifts than other methods.

3.6. RBF-PLS

This is a local method. It seems to perform particularly well for difficult data structures. Too little is known however about its practical use. It is not clear yet in which circumstances exactly it works well and what pitfalls may occur. The recommendation to consider it, should be understood as a recommendation for further investigation into the advantages and disadvantages of the method.

3.7 Decision scheme

The considerations given above lead to the decision scheme of Figure 1.

4. Methods that are not considered

In the intercomparison, many other methods were tried out, namely non-linear variants of PLS, PCR and MLR, nearest-neighbour methods, methods in which the regression is carried out on Fourier coefficients or wavelets .

Some of these do not yield good results, several others perform equally well as (some of) the methods selected, but are less easy to understand or are less generally used. For this reason their use is not recommended for users that are not experienced in chemometrics.

Fig. 1

```
┌─────────────────────┐
│ Data set acquisition│
│ and pre-processing  │
└─────────────────────┘
          │
          ▼
┌─────────────────────────────┐
│ Data inspection :           │◄──────────────────────────┐
│ - PC score plots            │                            │
│ - Plot of most correlated PCs vs y                       │
│ - Histogram of y-values     │                            │
└─────────────────────────────┘                            │
          │                                                 │
          ▼                                                 │
        ╱╲                                          ┌───────────────┐
       ╱  ╲                                         │ Remove outliers│
      ╱Obvious╲        yes                          └───────────────┘
      ╲outliers╱───────────────────────────────────────┘
       ╲detected╱
        ╲╱
          │ no
          ▼
        ╱╲
       ╱Calibration╲
      ╱space covers ╲        no
      ╲expected     ╱──────────────────┐
       ╲prediction ╱                   ▼
        ╲space ╱                      ╱╲
         ╲╱                          ╱Accurate╲
          │                         ╱prediction╲
          │                         ╲required outside╱
          │                          ╲calibration space╱
          │                           ╲╱
          │ yes                        │ yes              │ no
          │                            ▼                  │
          │                 ┌──────────────────────────┐ │
          │                 │ Predicted y-values outside the│ │
          │                 │ calibration space will indicate that a│ │
          │                 │ different model must be applied│ │
          │                 └──────────────────────────┘ │
          │                            │                  │
          │◄───────────────────────────┘                  │
          ▼                                                ▼
        ╱╲                                               ╱╲
       ╱Non linearity╲                                  ╱Non linearity╲
      ╱expected      ╲                                 ╱expected and/or╲
      ╲and/or detected╱                                ╲detected       ╱
       ╲╱       │ no                                    ╲╱      │ no
        │        ▼                                       │       ▼
   yes  │      ╱╲                                   yes  │     ╱╲
        │     ╱Clustering╲                               │    ╱Clustering╲
        │     ╲tendancy  ╱                               │    ╲tendancy  ╱
        │      ╲╱    │                                   │     ╲╱    │
        │  yes │  no │                                   │  yes │  no │
        │      ▼     ▼                                   │      ▼     ▼
        │  ┌────────┐                                    │  ┌────────┐
        │  │Compare MLR and│                             │  │Apply PCR/PLS│
        │  │LWR (including │                             │  │(no MLR) │
        │  │full PCR/PLS   │                             │  └────────┘
        │  │option)        │        ┌──────────┐         │
        │  └────────┘       │Apply LWR │         │
        ▼                            │(including full│   ▼
   ┌──────────┐                      │PCR/PLS option)│ ┌──────────┐
   │Apply LWR │  ┌──────────┐        │and/or NN │     │Apply LWR │
   │(including full│ │Apply preferably│ └──────────┘     │(including full│
   │PCR/PLS option)│ │MLR, alternatively│               │PCR/PLS option)│
   │and/or NN │   │try PLS/PCR│                          └──────────┘
   └──────────┘   └──────────┘
```

Data set acquisition and pre-processing

Data inspection :
- PC score plots
- Plot of most correlated PCs vs y
- Histogram of y-values

Obvious outliers detected — yes → Remove outliers

no

Calibration space covers expected prediction space — no →

Accurate prediction required outside calibration space

yes → Predicted y-values outside the calibration space will indicate that a different model must be applied

no

Non linearity expected and/or detected

yes → Apply LWR (including full PCR/PLS option) and/or NN

no → Clustering tendancy

yes → Compare MLR and LWR (including full PCR/PLS option)

no → Apply preferably MLR, alternatively try PLS/PCR

Non linearity expected and/or detected

yes → Apply LWR (including full PCR/PLS option) and/or NN

no → Clustering tendancy

yes → Apply LWR (including full PCR/PLS option)

no → Apply PCR/PLS (no MLR)