Robust preprocessing and model selection for spectral data

Sabine Verboven*, Mia Hubert[†], and Peter Goos[‡]

March 2012

Abstract

To calibrate spectral data one typically starts with preprocessing the spectra, and then applies a multivariate calibration method such as principal component regression (PCR) or partial least squares regression (PLSR). In the model selection step the optimal number of latent variables is determined in order to minimize the prediction error. To protect the analysis against the harmful influence of possible outliers in the data, robust calibration methods have been developed. In this paper we focus on the preprocessing and the model selection step. We propose several robust preprocessing methods as well as robust measures of the root mean squared error of prediction (RMSEP). To select the optimal preprocessing method, we summarize the results for the different RMSEP values by means of a desirability index (DI), which is a concept from industrial quality control. These robust RMSEPs are also used to select the optimal number of latent variables. We illustrate our newly developed techniques through the analysis of a real data set containing NIR-measurements of samples of animal feed.

Keywords: Robust preprocessing, robust prediction error, desirability index, robust calibration, NIR

*University of Antwerp, Department of Environment, Technology & Technology Management, Prinsstraat 13, 2000 Antwerp, Belgium

[†]KU Leuven, Department of Mathematics, Celestijnenlaan 200B, 3001 Heverlee, Belgium

[‡]University of Antwerp, Department of Environment, Technology & Technology Management and StatUa Statistics Center, Prinsstraat 13, 2000 Antwerp, Belgium; Erasmus Universiteit Rotterdam, Erasmus School of Economics, Postbus 1738, 3000 DR Rotterdam, The Netherlands

1 Introduction

Building a practically useful calibration model using principal component regression (PCR) or partial least squares regression (PLSR) involves preprocessing, model selection, and model validation. A common difficulty in the process is that real data are often contaminated with atypical samples, called outliers, that may have a detrimental impact on the estimation of the calibration model parameters. To withstand this unwanted effect of outlying observations, robust estimators have been developed. They have the property to produce accurate estimates even when outliers are present. Such highly robust methods have e.g. been proposed in [1],[2] and [3]. Overviews of their use, applicability and implementation in MATLAB (The MathWorks Inc., Natick, MA) and R are provided in [4],[5] and [6].

In this paper, we focus on comparing different preprocessing methods for NIR-spectral data and on the determination of the optimal number of latent variables for the preprocessed data set for a calibration model. An important feature of our work is that we consider robust versions of the most well-known preprocessing methods.

To compare the different preprocessing techniques and to determine the number of components, we define and use several robust root mean squared error of prediction (RMSEP) values. The idea behind this is to avoid having the selection of the best preprocessing technique be overly sensitive to outliers in the data. We then summarize the results for the different RMSEP values by means of a desirability index (DI), a concept from industrial quality control to produce a one-number summary of a range of scores on different dimensions. This DI allows us to make the preprocessing procedure fully automatic, which saves time as it does not require any user intervention.

After fixing the appropriate preprocessing method we continue the search for the optimal number of latent variables in the calibration models. Looking at the curves of the different RMSEPs for models with a varying number of latent variables and their standard errors, we can establish the model with sufficiently low prediction error. This leads to a final model for the data where one can look at the outliers by inspecting the outlier maps [2].

First we describe the real data set, containing NIR measurements of samples of animal

feed, on which we will illustrate our methodology (Section 2). Then we explain how to robustify several preprocessing techniques (see Section 3). In Section 4 we define the different robust root mean squared error of prediction measures. Section 5 is devoted to the selection of the best preprocessing method by means of the DI, and to the determination of the optimal number of latent variables.

Throughout the paper, we denote the observed spectra as $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^t$ for $i = 1, \dots, n$, hence n indicates the number of spectra and p the number of wavelengths. The observed concentrations are denoted as $\mathbf{y} = (y_1, \dots, y_n)'$.

2 Real data set

The data set on which we will apply our new procedure is provided by a Belgian animal feed company and consists of 895 Near InfraRed (NIR) reflectance spectra of animal feed samples measured every 2nm on a Foss 5000 instrument over 700 wavelengths with range 1100nm-2498nm. These raw spectra are depicted in Figure 1. From this figure we notice some outlying spectra, as well as some increasing trend and heteroscedasticity.

The feed samples (which are taken after production) vary from pelleted cattle feed to coarse mixes for poultry. There are in fact six animal categories: cattle, pig, poultry, horses, sheep and others (e.g. small farm animals). Furthermore, the data samples in this study were not ground before measurement which causes some heterogeneity in the data. For each sample several quality parameters are inspected among which the humidity level is the most important one. To be able to store the feed well, the humidity level should not exceed 12%. On the other hand the animal feed must be moist enough to meet customer demands. So the company aims to be as close to the 12% specification as possible by improving their calibration models. The reference method to determine humidity level in feed samples is the drying oven method. The samples are heated at 103 degrees Celcius for four hours, and afterwards the weight loss is measured. It is believed that the weight loss is due to the vaporization of water.

The goal of the calibration model is thus to predict the humidity level as a linear combination of the wavelength intensities in the NIR-reflectance spectrum at a new animal feed sample.

[Figure 1]

Note that when we analysed this data set using the robust calibration methods RPCR [1] and RSIMPLS [2], we had to decide a priori on the minimal proportion of samples α that are expected to be uncontaminated. To ensure a high level of robustness, we first took $\alpha = 0.5$. We then looked at different outlier diagnostics, and concluded that the data set contained several types of outliers (see also the detailed analysis in Section 5) but the overall proportion of outliers appeared to be much smaller than 50%, so we then fixed $\alpha = 0.75$.

3 Robust preprocessing techniques

There are several well-known preprocessing techniques for NIR data like e.g. standardization, taking first and second derivatives, standard normal variate transformation (SNV), multiplicative scattering correction (MSC), smoothing techniques etc. See for example [7], [8], [9] and references therein. Typically the different methods are evaluated for their optimality by the root mean squared error of prediction (RMSEP). But real life data are often contaminated and this contamination will influence the classical preprocessing techniques as well. Therefore, we consider robust versions of the most well-known preprocessing techniques which are summarized briefly in the next subsection.

Note that contamination in calibration data can occur in different ways. Often, some samples have a deviation spectrum. Other samples might be outlying with respect to the linear PCR or PLS model, because their response value is unusual. Both types of contamination can also occur at the same sample. These different types of outlyingness will be illustrated on the real data example in Section 5. As the preprocessing step only involves the spectra, and not the response value, robust preprocessing (instead of classical preprocessing) is only necessary when there are outlying spectra.

3.1 Robust standardisation

In robust standardisation, a spectrum is first robustly centered by subtracting the columnwise median of all spectra and then divided by the columnwise median absolute deviation (MAD)

which is a robust estimator of scale. Let

$$\text{med}(X)_j = \text{med}_{i=1,\dots,n} x_{ij}$$

and

$$\text{mad}(X)_j = 1.4826 \text{med}_{i=1,\dots,n} (|x_{i,j} - \text{med}(X)_j|)$$

then the elements $x_{i,j}^*$ of the transformed spectrum $\mathbf{x}^*$ are computed as:

$$x_{i,j}^* = \frac{x_{i,j} - \text{med}(X)_j}{\text{mad}(X)_j}$$

Note that also other robust measures of location and scale could be used here, such as the Q_n estimator of scale [10] or τ -estimators which are described in Section 4.4. Here, we applied estimators which are easy to compute and which are available in most statistical software.

3.2 Robust standard normal variate transformation

The standard normal variate (SNV) transformation has been proposed for removing the multiplicative interference of scatter and particle size [11]. Each spectrum is standardized by its own average and standard deviation. In the robust version, the centering is done by subtracting the row-wise median, while the robust scaling is done using the row-wise MAD. The transformed spectrum now becomes

$$x_{i,j}^* = \frac{x_{i,j} - \text{med}(X^t)_i}{\text{mad}(X^t)_i}.$$

Note that the MATLAB library TOMCAT contains this robust SNV transformation using the Q_n estimator instead of the MAD [12].

3.3 Robust multiplicative scattering correction

Multiplicative scattering correction (MSC) has been proposed in [13] and is used to compensate for additive and multiplicative effects in spectral data. Whereas MSC is based on the average spectrum, as well as on OLS regression, our robust MSC is based on the median spectrum and robust LTS regression. More precisely, for each sample ($i = 1, \dots, n$) the slope β_i and the intercept α_i in the regression

$$x_{i,j} = \alpha_i + \beta_i \text{med}(X)_j + e_{i,j} \tag{1}$$

with $j = 1, \dots, p$ are estimated by the robust LTS regression [14] estimator. The LTS estimator minimizes the sum of the h smallest squared residuals, where typically $h \approx 0.5p$ or $h \approx 0.75p$ (here, the regression is done on p values). In our example, we always use $h = [0.75p]$, which is appropriate if at most 25% of the wavelengths of a spectrum are contaminated.

An element $x_{i,j}^*$ of the corrected spectrum $\mathbf{x}_i^*$ is then calculated as:

$$x_{i,j}^* = \frac{x_{i,j} - \hat{\alpha}_i}{\hat{\beta}_i}$$

where $\hat{\alpha}_i$ and $\hat{\beta}_i$ are the robustly estimated regression coefficients. Thus, the spectrum is shifted vertically and scaled proportionally to make it as close as possible to the median spectrum.

3.4 Robust detrending

Detrending spectra accounts for the variation in baseline shift and (curvi)linearity of spectra by using a second degree polynomial to correct the data [11]. A second degree polynomial

$$x_{i,j} = \alpha_i + \beta_i \lambda_j + \gamma_i \lambda_j^2 + e_{i,j}$$

with λ_j the j -th wavelength can be used to standardise the variation in curvilinearity. The coefficients α_i, β_i and γ_i are robustly estimated using a LTS-regression of the spectrum $\mathbf{x}_i$ on the wavelengths. The transformed spectrum $\mathbf{x}_i^*$ is computed as

$$x_{i,j}^* = x_{i,j} - (\hat{\alpha}_i + \hat{\beta}_i \lambda_j + \hat{\gamma}_i \lambda_j^2)$$

where $\hat{\alpha}_i, \hat{\beta}_i$, and $\hat{\gamma}_i$ are the LTS-estimates of α_i, β_i , and γ_i .

3.5 Robust offset correction

Offset correction only removes the baseline shift [15]. Each spectrum is then corrected by subtracting either its absorbance at the first wavelength (or another arbitrarily chosen wavelength) or the median value in a selected range from l to u :

$$x_{i,j}^* = x_{i,j} - \text{med}_{1 \leq l \leq u \leq p} (x_{i,l:u})$$

with $i = 1, \dots, n$. In our examples we always used $l = 1$ and $u = p$.

3.6 Combined preprocessing methods

Finally we consider the combination of offset correction followed by MSC (denotes as offset-MSC), SNV followed by detrending (SNV-detrend) and SNV followed by taking first and second order differentiation (SNV-d1 and SNV-d2). On our real data set, this yields for example the preprocessed offset-MSC spectra in Figure 2 and the SNV-d1 spectra in Figure 3.

[Figure 2]

[Figure 3]

Other interesting preprocessing methods could consist of smoothing techniques, for which also robust alternatives have been proposed (such as the running median smoother [16] or the LOESS technique [17]) but this will not be pursued in this paper.

4 Robust root mean squared error of prediction

4.1 Classical RMSEP

To evaluate the preprocessing methods the root mean squared error of prediction (RMSEP) statistic is investigated. It measures the accuracy of the predictions of the calibration model. Obviously, the preprocessing technique leading to the smallest prediction error will be the preferred one. Given a test set of size n_t , the classical RMSEP is defined as

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

where $\hat{y}_i = \hat{\theta}_0 + \hat{\boldsymbol{\theta}}^t \mathbf{x}_i$ is the predicted response of the i th test spectrum. The regression coefficients $(\hat{\theta}_0, \hat{\boldsymbol{\theta}})$ are estimated on a training sample, typically by means of principal component regression (PCR) or a partial least squares regression (PLSR), whereas the prediction error RMSEP is computed using an independent test set. Here, we always randomly split the data set into a training set that contains 75% of the samples and a test set that contains the remaining 25%. Note that if the number of samples is small, also cross-validation methods can be used to estimate the prediction ability of the calibration model, as e.g. done in [1] and [18].

A first attempt to robustify the RMSEP is obtained by using robust calibration methods, such as the robust PCR method proposed in [1] or the robust partial least squares regression RSIMPLS [2]. But even when a robust calibration method is applied, the RMSEP is still not appropriate. As the spectra from the test set are usually just a random subset of the full data set, this test set can contain outlying samples. These outliers can have large residuals with respect to the model that is estimated via a robust calibration method which tries to fit well the regular observations. Hence the outliers in the test set will unnecessarily increase the RMSEP. Therefore we consider several robust alternatives, all based on robust parameter estimates $(\hat{\theta}_0, \hat{\boldsymbol{\theta}})$.

4.2 Outlier-free RMSEP

The outlier-free RMSEP, as introduced in [2], is defined as

$$\text{RMSEP}_{\text{outfree}} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2} \quad (3)$$

where m is the number of *non-outlying observations* in the test set (so m is often smaller than n_t). To determine whether a test sample is outlying or not, its standardized residual based on the training estimates $(\hat{\theta}_0, \hat{\boldsymbol{\theta}})$ is computed and compared to a cutoff value, see [1] and [2]. It is thus a hard rejection rule.

4.3 Trimmed RMSEP

The trimmed RMSEP has been introduced in [3] and is defined as

$$\text{RMSEP}_{\text{trimh}} = \sqrt{\frac{1}{h} \sum_{i=1}^h (y - \hat{y})_{[i:n_t]}^2} \quad (4)$$

where $(y - \hat{y})_{[i:n_t]}^2$ is the i th ordered squared residual and h ranges from $[0.5n_t]$ to n_t . It expresses how well the h/n_t best predictions perform (see also [19]). We consider in particular $h = [0.5n_t]$, $h = [0.75n_t]$ and $h = [0.95n_t]$, and denote the corresponding RMSEP values by $\text{RMSEP}_{0.5}$, $\text{RMSEP}_{0.75}$ and $\text{RMSEP}_{0.95}$. Note that by definition $\text{RMSEP}_{0.5} \leq \text{RMSEP}_{0.75} \leq \text{RMSEP}_{0.95}$ when they are computed on the same test set.

4.4 Robust mean-squared-error RMSEP

Finally, we also use an RMSEP value following the ideas of Bianco and Boente [20]. We refer to it as the mean-squared-error RMSEP ($\text{RMSEP}_{\text{mse}}$) and define it as follows:

$$\text{RMSEP}_{\text{mse}} = \sqrt{\hat{\mu}^2(r_i) + \hat{\sigma}^2(r_i)}, \quad (5)$$

with $r_i = y_i - \hat{y}_i$ the residuals from the test set, and $\hat{\mu}$ and $\hat{\sigma}$ robust estimators of location and scale. Formula (5) expresses the well-known fact that the mean squared error equals the sum of the squared bias and variance. As in [20], we consider for $\hat{\mu}$ and $\hat{\sigma}$ univariate τ -estimators of location and scale [21]. For univariate data $R = r_1, \dots, r_{n_t}$ they are defined as

$$\hat{\mu}(R) = \frac{\sum_i w_i r_i}{\sum_i w_i}$$

with $w_i = W\left(\frac{r_i - \text{med}(R)}{\text{mad}(R)}\right)$ and $W(x) = \left(1 - \left(\frac{x}{4.5}\right)^2\right)^2 I(|x| \leq 4.5)$, whereas

$$\hat{\sigma}(R)^2 = \frac{\text{mad}(R)^2}{n_t} \sum_i \rho\left(\frac{r_i - \hat{\mu}(R)}{\text{mad}(R)}\right)$$

with $\rho(x) = \min(x^2, 3^2)$. We see that $\hat{\mu}$ and $\hat{\sigma}$ downweight outliers by the choice of the weight function W for location and the bounded function ρ for scale.

5 Model selection

5.1 Desirability index for the selection of the suboptimal rank

All the RMSEP values from Section 4 depend on the number of latent variables k in the robust calibration method. To determine the best preprocessing method, we should optimize the RMSEP over all preprocessing methods and all ranks k . We propose to do this in a two-step procedure. First, for each preprocessing method, we determine in an automatic way its suboptimal rank by computing the five robust RMSEP values for each k and combining them using the desirability index (DI). Next, we consider for each preprocessing method the RMSEP values for their suboptimal rank, and use the DI again to decide upon the optimal preprocessing method.

We first explain how we determine the suboptimal rank for a fixed preprocessing method. We start by randomly splitting the full data set into a training and test set. Then we

apply the preprocessing procedure on the training data, and apply a robust calibration method for a range of number of latent variables $k = 1, \dots, k_{max}$. Next, we apply the same preprocessing procedure on the test data, but based on the results from the training set. Consider e.g. the robust standardization as preprocessing technique. The test samples are then standardized according to the median and the mad of the training samples. For each k , we can then compute the fitted values of the test samples according to the robust calibration parameters, as well as the five robust RMSEP values defined in Section 4: $RMSEP_{outlfree}$, $RMSEP_{0.5}$, $RMSEP_{0.75}$, $RMSEP_{0.95}$ and $RMSEP_{mse}$. To avoid that the resulting RMSEP values depend highly on the specific partitioning into training and test set, we have repeated five times the random splitting, and took the average RMSEP values over the five runs. Doing so, we are still left with several robust RMSEP values, that all measure the prediction ability in a different way. Typically these RMSEP decrease up to a certain value of k and then start to increase again, but they do not necessarily all reach their minimum at the same k value.

Consider for example Table 1 which displays the resulting RMSEP values, obtained on our real data set, preprocessed according to the robust offset correction and calibrated with RSIMPLS for $k = 1$ up to $k_{max} = 12$ PLS components. The minimum $RMSEP_{outlfree}$ value is reached at $k = 12$, the minimum $RMSEP_{0.5}$ at $k = 11$, while $RMSEP_{0.75}$, $RMSEP_{0.95}$ and $RMSEP_{mse}$ are in favor of selecting $k = 10$ latent variables. To decide upon the optimal number of components in an automatic way, we prefer to combine all RMSEP values.

[Table 1]

To this end, we introduce a concept from industrial quality control, the desirability index (DI). This index was introduced by Harrison [22] as a method for multicriteria optimization (see also [23]). The goal is to summarize the results for different optimisation criteria in one single performance measure. Therefore, the results for each of the criteria are assigned a unitless value, called the desirability coefficient (DC), between 0 and 1 by aid of a desirability function. Next, these coefficients are combined into a DI by taking the geometric mean of the DC values.

In our case, the different optimisation criteria are the $C = 5$ (averaged) RMSEP values, and the first goal is to find for each the preprocessing technique the number of components k

that yield the smallest overall RMSEP. In general, if the different choices of k are ordered in R rows and the different robust RMSEP values in C columns, a DC value can be computed as

$$\text{DC}_{r,c} = \frac{\min_r(\text{RMSEP}_{r,c})}{\text{RMSEP}_{r,c}} \quad (6)$$

for every cell (r, c) with $r = 1, \dots, R$ and $c = 1, \dots, C$, i.e. for every combination of rank and RMSEP criterion in the resulting table. A $\text{DC}_{r,c}$ value is close to 1 if the corresponding rank k gives rise to an RMSEP value close to the minimum RMSEP of the column c . The desirability index for each row r of the table is then computed as

$$\text{DI}_r = \prod_{c=1}^C \text{DC}_{r,c}. \quad (7)$$

In this approach, an equal importance is attached to each of the RMSEP values. Obviously, the rank with the highest DI is the preferred one. In the artificial example of Table 2 the first method attains 50% desirability compared to 30% for the second method. Looking at the individual RMSEP values this seems logical since the first method attains the lowest prediction errors on two out of three measures.

[Table 2]

The RMSEP values of the real data set, preprocessed with the offset correction, yields desirability coefficients and desirability indices, as exposed in Table 3. Here we conclude that we should consider $k = 10$.

[Table 3]

Note that we consider this selection of k a suboptimal solution, as it is fast and fully automatic but it does not yet penalize large models with many components. In the final stage of the model selection step (Section 5.3) we will apply a more refined rank selection procedure.

5.2 Desirability index for the selection of the optimal preprocessing method

Once we have obtained the suboptimal rank for each preprocessing method, we will now determine the optimal preprocessing technique, again with the help of the desirability indices. The table on which the desirability coefficients and the desirability indices will be computed, now consists of m rows, with m the number of preprocessors under study. The columns contain for each method the five (averaged) RMSEP values at their suboptimal rank obtained in the previous step. We now select that preprocessing method with the largest desirability index, i.e. the one which yields globally the smallest averaged robust prediction errors at its suboptimal rank, thereby measuring the prediction error in five different ways.

In our data example, we considered 9 preprocessing methods as described in Section 3: robust standardisation, robust SNV, robust MSC, robust detrending, robust offset correction, offset correction followed by MSC, SNV followed by detrending, SNV followed by first order differentiation and SNV followed by second order differentiation. For comparison we also include the analysis on the raw data, so without applying any preprocessing. The resulting RMSEP values are displayed in Table 4. We have also added a column which indicates the corresponding chosen rank. The minimum value for each RMSEP value is underlined. We see that the automatic rank selection selects the maximal number of considered components for many preprocessors, but not all. Moreover it is clear that SNV followed by first order derivation yields globally the smallest prediction errors, i.e. when we compute the DI on the five different prediction errors, we see that SNV-d1 has the maximal DI value of 0.937. The second best preprocessing method is Offset-MSC with a DI of 0.891. The raw data only achieve a DI of 0.31, hence preprocessing seems really useful here.

From Table 4 we also notice that $\text{RMSEP}_{0.75} < \text{RMSEP}_{\text{outlfree}} < \text{RMSEP}_{0.95}$ for each preprocessing method. This is an indication that the number of regular observations in the test sets of size $n_t = n - [0.75n] = 224$ lies between $[0.75n_t]$ and $[0.95n_t]$, and thus that the proportion of outliers in the data lies between 5% and 25%.

[Table 4]

5.3 Selecting the optimal number of latent variables

Once the most appropriate preprocessing method is chosen, the calibration model can be optimized for the number of latent variables. Again we look at the different robust RMSEP values obtained from the previous analysis, but now we visualize the results by plotting them on one graph. In this way we obtain different curves, one curve for each robust RMSEP value, which is an evaluation of the prediction error for the calibration model in $k = 1, \dots, k_{max}$ latent variables. For each curve we now look at the value k where the RMSEP value is minimal (or close to minimal). Alternatively we can add standard errors to the RMSEP values, which are obtained from the five splits in training and test set. According to [25] and [6] we can then take the most parsimonious model for which the prediction mean is smaller than twice the standard error of the RMSEP where the minimum is attained.

In the worst case this procedure might give five different rank values, one for each RMSEP measure. Depending on the application, one might then decide to take a final decision based on only one RMSEP measure (here, our preference goes to RMSEP_{mse} as it does not require to fix the trimming proportion), or by considering all curves and all numerical results. Note that this will typically yield a different rank than the one obtained in the first stage of our analysis as carried out in Section 5.1 where the rank determination was done in a completely automatic way. Our last step here needs more manual intervention, but this is less of an issue as it only needs to be done on one preprocessed data set. It is in particular interesting as it might yield a model with fewer components.

Having established the optimal preprocessing and the optimal subspace dimension k , the final calibration model can be built on the whole dataset using a robust calibration method such as RPCR and RSIMPLS. The regression coefficients from this final calibration model are the ones to be used for future predictions of the response of interest.

Let us illustrate this procedure on the animal feed data set. The resulting RMSEP values are depicted in Figure 4. We see that all curves show the same tendency. It seems that at least 6 components should be retained leading to RMSEP values which are all smaller than 0.4%. This is in line with the company's preference to end up with a calibration method that has an RMSEP smaller than 0.4%. The selection rule based on the standard errors of the RMSEP values yields $k_{opt} = 9$ based on $\text{RMSEP}_{0.75}$, $\text{RMSEP}_{0.95}$ and RMSEP_{mse} . The outlier free $\text{RMSEP}_{outlfree}$ results in $k_{opt} = 6$, whereas $\text{RMSEP}_{0.5}$ yields $k_{opt} = 11$. Based on these results and the graphical output, we would opt for a minimum of 6 and a maximum

of 9 components.

[Figure 4]

When we perform RSIMPLS with $k = 9$ components on the animal feed data set, we obtain the outlier maps in Figure 5. The first outlier map shows the outlyingness of the spectra. On the vertical axis the orthogonal distance to the PLS-score space is depicted, whereas the horizontal axis shows the score distance, i.e. the robust Mahalanobis distances of the spectra in the 9-dimensional score space. The second plot depicts the standardized residuals versus the robust distances. (See [24, 2] for the precise definition of the outliers maps.) It can be seen that there are many outlying observations. In particular we notice on the right figure that observation 248 has a huge score distance (SD) but a small absolute standardized regression residual. This is an illustration of a good regression leverage point. Sample 383 on the other hand has both a large SD and a large standardized residual, which indicates that it is a bad regression leverage point. Besides there are several vertical outliers such as case 357. They do not have an outlying spectrum, but an unusual large absolute residual.

[Figure 5]

When we apply classical SIMPLS on the robustly preprocessed data, we obtain the outliers maps in Figure 6. On the left plot we see that fewer outlying spectra are flagged than using RSIMPLS. Also the regression outlier map on the right shows only a small number of samples with outlying residual. Sample 248 is again recognized as a good leverage point. Sample 383 on the other hand is also flagged as a good leverage point, which implies that according to the classical SIMPLS analysis, its humidity level is not abnormal. The robust analysis in Figure 5 marked that sample as a bad leverage point, thus having an unusual humidity level according to its spectrum. This again illustrates the power of robust calibration methods to detect outlying observations.

[Figure 6]

6 Conclusion

We have presented several robust preprocessing methods, as well as different robust measures of the prediction ability of a calibration method. To combine all the information provided by these measures, and to select the optimal preprocessing method, we introduced the desirability index. Moreover the RMSEP values were used to select the optimal number of latent variables. This allowed us to perform the model selection step in an automatic way, which is time saving and which simplifies the model development process. The analysis of a real data set on animal feed showed the benefits of using this robust calibration procedure.

Acknowledgments

When performing this research, Sabine Verboven was funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT-Vlaanderen), project no. OZM/070221. We thank Marnix De Schrijver for providing the animal feed data set and for his feedback on our analysis. We also acknowledge two anonymous reviewers for their constructive comments which have led to an improved paper.

References

- [1] Hubert M. and Verboven S. A robust PCR method for high-dimensional regressors. *Journal of Chemometrics* 2003; **17**, 438–452.
- [2] Hubert, M. and Vanden Branden, K. Robust methods for Partial Least Squares regression. *Journal of Chemometrics* 2003; **17**, 537–549.
- [3] Serneels, S., Croux, C., Filzmoser, P., Van Espen, P.J. Partial robust M-regression. *Chemometrics and Intelligent Laboratory Systems* 2005; **79**, 55–64.
- [4] Rousseeuw, P.J., Debruyne, M., Engelen, S., Hubert, M. Robustness and outlier detection in chemometrics. *Critical Reviews in Analytical Chemistry* 2006; **36**: 221–242.
- [5] Verboven, S. and Hubert, M. LIBRA: a MATLAB Library for Robust Analysis. *Chemometrics and Intelligent Laboratory Systems* 2005; **75**, 127–136.

- [6] Filzmoser, P., Todorov, V. Review of robust multivariate statistical methods in high dimension. *Analytica Chimica Acta* 2011; **705**: 2–14.
- [7] Beebe, K.R., Pell, R.J., and Seasholtz, M.B. *Chemometrics: A Practical Guide*. John Wiley & Sons, 1998.
- [8] Thennadil, S.N., and Martin, E.B. Empirical preprocessing methods and their impact on NIR calibrations: a simulation study. *Journal of Chemometrics* 2005; **19**: 77–89.
- [9] Xu, L., Zhou, Y.-P., Tang, L.-J., Wu, H.-L., Jiang, J.-H., Shen, G.-L., and Yu, R.-Q. Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. *Analytica Chimica Acta* 2008; **616**, 138–143.
- [10] Rousseeuw, P.J., Croux, C., Alternatives to the median absolute deviation. *Journal of the American Statistical Association* 1993; **88**, 1273–1283.
- [11] Barnes, R.J., Dhanoa, M.S., and Lister, S.J. Standard normal variate transformation and detrending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy* 1989; **43**: 772–777.
- [12] Daszykowski, M., Serneels, S., Kaczmarek, K., Van Espen, P., Croux, C., and Walczak, B. TOMCAT: a MATLAB toolbox for multivariate calibration techniques. *Chemometrics and Intelligent Laboratory Systems* 2007; **85**, 269–277.
- [13] Geladi, P., MacDougall, D., and Martens, H. Linearization and scatter correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy* 1985; **39**: 491–500.
- [14] Rousseeuw, P.J. Least median of squares regression. *Journal of the American Statistical Association* 1984; **79**, 871–880.
- [15] Brereton, R.G. *Chemometrics: data analysis for the laboratory and chemical plant*. John Wiley & Sons, 2003.
- [16] Tukey, J. W. *Exploratory Data Analysis*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1977.
- [17] Cleveland, W.S. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 1979; **74**, 829–836.

- [18] Engelen, S. and Hubert, M. Fast model selection for robust calibration methods. *Analytica Chimica Acta* 2005; **544**, 219–228.
- [19] Khan, J.A., Van Aelst, S., and Zamar, R.H. Fast robust estimation of prediction error based on resampling. *Computational Statistics and Data Analysis* 2010; **54**, 3121–3130.
- [20] Bianco, A. and Boente, G. Robust Estimators under semi-parametric partly linear autoregression: asymptotic behaviour and bandwidth selection. *Journal of Time Series Analysis* 2004; **28**, 274–306.
- [21] Yohai, V.J. and Zamar, R. High breakdown point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association* 1988; **86**, 403–413.
- [22] Harrison, J. The desirability function. *Industrial Quality Control* 1965; **12**, 494–498.
- [23] Derringer, G.C. and Suich, D. Simultaneous optimization of several response variables. *Journal of Quality Technology* 1980; **12**, 214–219.
- [24] Hubert, M., Rousseeuw, P.J., and Vanden Branden, K. ROBPCA: a new approach to robust principal component analysis. *Technometrics* 2005; **47**, 64–79.
- [25] Hastie, T., Tibshirani, R. and Friedman, J. The Elements of Statistical Learning. Springer, New York, 2001.

components	RMSEP _{outlfree}	RMSEP _{0.5}	RMSEP _{0.75}	RMSEP _{0.95}	RMSEP _{mse}
1	0.683	0.278	0.469	0.707	0.763
2	0.556	0.254	0.419	0.722	0.699
3	0.462	0.203	0.332	0.496	0.538
4	0.361	0.167	0.282	0.452	0.472
5	0.348	0.161	0.268	0.423	0.446
6	0.321	0.153	0.264	0.492	0.446
7	0.316	0.140	0.241	0.402	0.404
8	0.317	0.138	0.238	0.385	0.392
9	0.306	0.136	0.237	0.395	0.393
10	0.292	0.132	0.223	0.380	0.377
11	0.286	0.130	0.233	0.437	0.394
12	0.282	0.132	0.235	0.453	0.406

Table 1: RMSEP values (in %) for RSIMPLS models with $k = 1, \dots, 12$ components, pre-processed according to the robust offset correction.

	RMSEP ₁	RMSEP ₂	RMSEP ₃	
preprocess 1	0.10	0.15	0.20	
preprocess 2	0.20	0.25	0.10	
	DC _{.,1}	DC _{.,2}	DC _{.,3}	DI
preprocess 1	1	1	0.5	0.5
preprocess 2	0.5	0.6	1	0.3

Table 2: Artificial example to illustrate the desirability index.

components	DC _{.,1}	DC _{.,2}	DC _{.,3}	DC _{.,4}	DC _{.,5}	DI
	RMSEP _{outlfree}	RMSEP _{0.5}	RMSEP _{0.75}	RMSEP _{0.95}	RMSEP _{mse}	
1	0.413	0.469	0.475	0.537	0.494	0.024
2	0.507	0.513	0.532	0.526	0.539	0.039
3	0.610	0.643	0.672	0.767	0.699	0.141
4	0.781	0.779	0.791	0.842	0.798	0.323
5	0.810	0.809	0.833	0.899	0.844	0.414
6	0.878	0.855	0.847	0.773	0.844	0.415
7	0.894	0.931	0.928	0.947	0.933	0.682
8	0.890	0.947	0.939	0.988	0.961	0.752
9	0.921	0.958	0.942	0.962	0.957	0.765
10	0.967	0.991	1	1	1	0.958
11	0.986	1	0.959	0.870	0.955	0.786
12	1	0.985	0.949	0.838	0.927	0.726

Table 3: Desirability coefficients and desirability indices for RSIMPLS models with $k = 1, \dots, 12$ components, preprocessed according to the robust offset correction.

	k	DC _{.,1} RMSEP _{outlfree}	DC _{.,2} RMSEP _{0.5}	DC _{.,3} RMSEP _{0.75}	DC _{.,4} RMSEP _{0.95}	DC _{.,5} RMSEP _{mse}	DI
raw	8	0.320	0.145	0.248	0.421	0.417	0.310
Std	8	0.329	0.148	0.257	0.417	0.428	0.282
Offset	10	0.292	0.132	0.223	0.380	0.377	0.572
Detrend	12	0.282	0.129	0.242	0.517	0.415	0.333
SNV	11	0.278	0.122	0.206	0.342	0.344	0.768
MSC	12	0.288	0.129	0.209	0.344	0.347	0.679
Offset-MSC	12	<u>0.256</u>	0.118	<u>0.197</u>	0.345	0.342	0.891
SNV-detrend	12	0.281	0.122	0.210	0.355	0.355	0.692
SNV-d1	12	0.273	<u>0.114</u>	0.198	<u>0.332</u>	<u>0.328</u>	0.937
SNV-d2	12	0.387	0.156	0.269	0.424	0.446	0.204

Table 4: Robust RMSEP values (in %) and desirability indices obtained to the raw data and to several preprocessed data, at the suboptimal k for each preprocessing method. The minimal value for each RMSEP value is underlined.

Figure 1: The raw animal feed data measured on the Foss5000 instrument.

Figure 2: The animal feed data preprocessed with offset correction followed by MSC (offset-MS)

Figure 3: The animal feed data preprocessed with SNV and first derivatives (SNV-d1).

Figure 4: Robust RMSEP curves for the preprocessed animal feed data. For the RMSEP_{mse} criterion also the confidence band (RMSEP estimate ± 2 times the standarderror) is indicated.

Figure 5: Outlier maps for the preprocessed animal feed data set, calibrated with RSIMPLS.

Figure 6: Outlier maps for the preprocessed animal feed data set, calibrated with SIMPLS.