Analytica Chimica Acta 938 (2016) 44–52

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Boosting model performance and interpretation by entangling preprocessing selection and variable selection

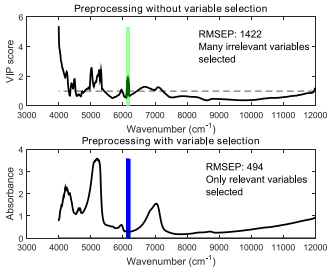
Jan Gerretzen [a, b], Ewa Szymańska [a, b], Jacob Bart [c], Antony N. Davies [c, d], Henk-Jan van Manen [c], Edwin R. van den Heuvel [e], Jeroen J. Jansen [a], Lutgarde M.C. Buydens [a, *]

[a] Radboud University, Institute for Molecules and Materials, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
[b] TI-COAST, P.O. Box 18, 6160 MD Geleen, The Netherlands
[c] AkzoNobel, Supply Chain, Research & Development, Strategic Research Group – Measurement & Analytical Science, Zutphenseweg 10, 7418 AJ Deventer, The Netherlands
[d] SERC, Sustainable Environment Research Centre, Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, CF37 1DL, UK
[e] Eindhoven University of Technology, Den Dolech 2, 5600 MB Eindhoven, The Netherlands

HIGHLIGHTS

- A generic approach for preprocessing selection and variable selection is proposed.
- Variable selection has been integrated in the process of preprocessing selection.
- This integration leads to improved predictive model performance.
- It also enables correct interpretation of the model.
- Appropriate preprocessing aids in extracting the true relevant variables.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 17 June 2016
Received in revised form
27 July 2016
Accepted 9 August 2016
Available online 15 August 2016

Keywords:
Design of experiments
Variable selection
Preprocessing selection
Partial least squares
Chemometrics

ABSTRACT

The aim of data preprocessing is to remove data artifacts—such as a baseline, scatter effects or noise—and to enhance the contextually relevant information. Many preprocessing methods exist to deliver one or more of these benefits, but which method or combination of methods should be used for the specific data being analyzed is difficult to select. Recently, we have shown that a preprocessing selection approach based on Design of Experiments (DoE) enables correct selection of highly appropriate preprocessing strategies within reasonable time frames.

In that approach, the focus was solely on improving the predictive performance of the chemometric model. This is, however, only one of the two relevant criteria in modeling: interpretation of the model results can be just as important. Variable selection is often used to achieve such interpretation. Data artifacts, however, may hamper proper variable selection by masking the true relevant variables. The choice of preprocessing therefore has a huge impact on the outcome of variable selection methods and may thus hamper an objective interpretation of the final model. To enhance such objective interpretation, we here integrate variable selection into the preprocessing selection approach that is based on DoE.

We show that the entanglement of preprocessing selection and variable selection not only improves the interpretation, but also the predictive performance of the model. This is achieved by analyzing several experimental data sets of which the true relevant variables are available as prior knowledge. We

* Corresponding author.
E-mail address: chemometrics@science.ru.nl (L.M.C. Buydens).

http://dx.doi.org/10.1016/j.aca.2016.08.022

show that a selection of variables is provided that complies more with the true informative variables compared to individual optimization of both model aspects.

Importantly, the approach presented in this work is generic. Different types of models (e.g. PCR, PLS, …) can be incorporated into it, as well as different variable selection methods and different preprocessing methods, according to the taste and experience of the user. In this work, the approach is illustrated by using PLS as model and PPRV-FCAM (Predictive Property Ranked Variable using Final Complexity Adapted Models) for variable selection.

1. Introduction

In chemometric data analysis, it is important that data variation due to data artifacts is removed from the data prior to construction of a chemometric model. This variation is not related to the ultimate data goal, such as regression or classification and as such hampers chemometric model performance. Examples of such variation include time misalignment, commonly encountered in chromatographic data, or baseline and scatter effects, often present in spectroscopic data. Data *preprocessing* aims to remove this 'irrelevant' variation: it transforms the original data into preprocessed data, which has been cleaned from uninformative variation.

Data from each analytical chemical platform—such as infrared or nuclear magnetic resonance spectroscopy, mass spectrometry or separation sciences such as gas chromatography—are associated with their own sources of uninformative variation. Many preprocessing *methods* have been developed for each platform, which aim to remove a single source of uninformative variation from the data [1–6]. Since data often contains multiple sources of uninformative variation, multiple preprocessing methods often need to be applied in what we have defined previously as a preprocessing *strategy* [7]. A strategy consists of consecutive preprocessing *steps* (e.g. scatter correction or smoothing), where a different preprocessing method is applied for each step in the strategy.

In previous work, we have shown that the influence of preprocessing on chemometric model performance may be considerable [8]. Care must be taken as preprocessing using strategies that combine methods of widespread use in the literature may be detrimental to the overall information content in the data. Appropriate preprocessing selection is therefore a major issue in chemometrics. However, currently available preprocessing selection approaches are seriously lacking and likely lead to a suboptimal selection of a preprocessing strategy [8]. Therefore, we have previously developed a systematic approach based on Design of Experiments (DoE), to specifically evaluate which preprocessing steps are relevant for a given data set [7]. This information is then subsequently used to introduce the most appropriate preprocessing method for each step deemed relevant by the DoE.

This earlier work, however, only used the prediction accuracy to evaluate the quality of different preprocessing strategies. This was a logical first step, as it provided an unbiased basis to evaluate model quality that did not require any prior knowledge and was therefore most widely applicable. Interpretation of the constructed models, i.e. the relative importance of each measured variable to the prediction, was not taken into account. Interpretability, however, is also a very relevant part in chemometric modeling, often even the most important goal of the analysis. Therefore, our aim is to select a preprocessing strategy for a given data set, which improves not only model performance, but also model interpretation.

Many approaches are available regarding the importance of variables in Partial Least Squares (PLS) models, on which we will focus in this work. The most straightforward approaches are so-called filter methods [9]. Filter methods are applied on the output of the PLS algorithm (e.g. regression coefficients, scores, loadings) and transform these into variable importance measures. Well-known examples include the Variable Importance in Projection (VIP), the Selectivity Ratio (SR) and significance Multivariate Correlation (sMC) [10,11]. Based on the outcome of such a filter method, variables can be selected by e.g. setting a threshold on the value of the variable importance measure. For example, when using VIP variables are often deemed relevant if their VIP score is >1.

However, as we will show in this work, the application of filter methods in the process of preprocessing selection does not enhance model interpretability. This is due to the fact that the ultimately selected preprocessing strategy is applied to all variables in the data, including those that may hamper the model. Ideally, a preprocessing strategy should be chosen that removes artifacts from the chemically relevant variables only. It is easy to imagine that this may require a different preprocessing strategy, consisting of different preprocessing steps and methods. The only way to find an appropriate preprocessing strategy that enhances both model interpretation and model performance, is therefore to *entangle* preprocessing selection with variable selection.

In this work, we provide an example of how the selection of preprocessing and variable selection can be entangled, using our DoE-based approach for preprocessing selection. Model predictive performance is expected to improve even more compared to models for which preprocessing has been optimized without variable selection: indeed, many uninformative variables have been removed from the data and thus cannot hamper the model anymore. Secondly, the correct combination of a preprocessing strategy and variable selection should enhance model interpretation by highlighting the true chemically relevant variables. Both advantages will be proven here.

The example we provide is based on another class of variable selection methods in PLS: wrapper methods [9]. They extend the concept of filter methods by starting from a PLS model based on all variables, followed by iteratively removing variables from the data and refitting a PLS model on the reduced data. Variable removal may, for instance, be based on a variable importance measure from a filter method. Our example uses a wrapper method from the Predictive Property-Ranked Variable (PPRV) family of methods [12,13]. This method was chosen because it was shown to lead to improved results compared to other commonly used variable selection methods.

A large selection of variable selection methods exists, containing for example iPLS (interval PLS), UVE-PLS (Uninformative Variable Elimination PLS) and IPW PLS (Iterative Predictor Weighting PLS)—see e.g. Refs. [9,14–18] for more details. Our aim in this work is not to provide a comprehensive comparison of these variable selection methods—such comparisons may be found elsewhere, e.g. Refs. [19–21]. We aim to show that entangling preprocessing selection with variable selection boosts both model performance *and*

model interpretation. We provide a generic approach to do so, in which the model, the variable selection method and different preprocessing steps and methods can all be selected by the user based on e.g. the characteristics of the data or the taste and experience of the user. The focus of the examples will be on spectroscopic data sets.

2. Experimental

2.1. Methods

In this section, we will first extensively describe the original DoE approach as described in Refs. [7], after which we will provide all required details on the integrated variable selection algorithm, PPRV-FCAM.

2.1.1. The original DoE approach

The aim of the original DoE approach was to evaluate which preprocessing *steps* are relevant for the data under study and which are not, by focusing solely on improving model performance.

For spectroscopic data, the four commonly applied preprocessing steps are baseline correction, scatter correction, noise reduction by smoothing and scaling, also often applied in this order [8]. These four steps are evaluated using a two-level full factorial design, where each factor in the design represents a preprocessing step. The low level in this design always represents "do nothing" (i.e. do not perform the specific step), while the high level equals one of the available methods for each step (see Table 1).

This design consists of $2^4 = 16$ experiments in total. Data are split in a training set and test set and the training set is preprocessed according to the methods specified in each of the 16 experiments. A PLS model is subsequently built for each preprocessed training data set, using single cross-validation to optimize the number of latent variables in the model. Each model is then applied to the corresponding test set—which has been preprocessed accordingly—leading to 16 RMSEP values. These are the responses for the design.

Using standard effect calculations, the effect value of each preprocessing step can be calculated. For example, if the effect value of the baseline correction step is negative, then a decrease in RMSEP (and hence, a better model in terms of performance) is expected when performing baseline correction. Thus, baseline correction is a step that should be considered further. Preprocessing steps that have a nonnegative or zero effect value are excluded from further investigation. Effect values of second-order interactions between preprocessing steps are also taken into account. A negative effect value of the second-order interaction between, for example, baseline correction and scatter correction implies that an additional decrease in RMSEP is expected when performing both baseline correction and scatter correction.

A bootstrap procedure is applied to estimate the significance of effects [7]. Bootstrapping creates artificial subsets of similar size of the original data matrix by resampling the original samples. Some samples may therefore be present multiple times in such a subset, while others may not be present. In our procedure, 150 bootstrapped data sets are created based on the original data set. The complete approach is repeated for each bootstrapped data set and effect values are calculated based on the bootstrapped data sets. The pooled standard deviation—based on the variances in RMSEP in each of the 16 rows in the DoE—is used to estimate the significance of each effect.

Next, the most appropriate preprocessing *method* should be found for each preprocessing step deemed relevant using the design. This is done using a scheme in which the most appropriate preprocessing method for each step is sequentially selected. For example, suppose that baseline correction and scatter correction are the two relevant steps. First, the baseline correction method leading to the lowest RMSEP is selected from among the available methods. The list of preprocessing methods we used for each step can be found in Ref. [8]. Next, the most appropriate scatter correction method is selected, while the baseline correction method is fixed to the method already selected.

It should be noted that the order in which the preprocessing steps are applied is fixed. As also discussed in Refs. [7], our order is the order in which the different preprocessing steps are commonly applied. A user is free to change the application order by changing the ordering of the columns in the DoE. In this work, we chose the original order of applying the different preprocessing steps.

2.1.2. Entangling variable selection with preprocessing selection

For our current approach, we integrated variable selection into the original DoE-based preprocessing selection approach described in the previous section. To do so, we replaced PLS with the wrapper method PPRV-FCAM [12,13]. PPRV methods iteratively remove

Table 1
Design matrix as used in the DoE approach.

Experiment	Baseline[a]	Scatter[b]	Smoothing[c]	Scaling[d]	Response (RMSEP)
1	+ (AsLS)	+ (SNV)	+ (smoothing)	+ (Pareto)	
2	+ (AsLS)	+ (SNV)	+ (smoothing)	− (meancenter)	
3	+ (AsLS)	+ (SNV)	− (do nothing)	+ (Pareto)	
4	+ (AsLS)	+ (SNV)	− (do nothing)	− (meancenter)	
5	+ (AsLS)	− (do nothing)	+ (smoothing)	+ (Pareto)	
6	+ (AsLS)	− (do nothing)	+ (smoothing)	− (meancenter)	
7	+ (AsLS)	− (do nothing)	− (do nothing)	+ (Pareto)	
8	+ (AsLS)	− (do nothing)	− (do nothing)	− (meancenter)	
9	− (do nothing)	+ (SNV)	+ (smoothing)	+ (Pareto)	
10	− (do nothing)	+ (SNV)	+ (smoothing)	− (meancenter)	
11	− (do nothing)	+ (SNV)	− (do nothing)	+ (Pareto)	
12	− (do nothing)	+ (SNV)	− (do nothing)	− (meancenter)	
13	− (do nothing)	− (do nothing)	+ (smoothing)	+ (Pareto)	
14	− (do nothing)	− (do nothing)	+ (smoothing)	− (meancenter)	
15	− (do nothing)	− (do nothing)	− (do nothing)	+ (Pareto)	
16	− (do nothing)	− (do nothing)	− (do nothing)	− (meancenter)	

[a] AsLS: Asymmetric Least Squares baseline estimation [28].
[b] SNV: Standard Normal Variate [29].
[c] Smoothing implies Savitzky-Golay smoothing (window width 9 px, 3rd order polynomial).
[d] The low level represents meancentering instead of do nothing, because meancentering is customary for many PLS models.

variables, until no more variables can be removed without significantly influencing model performance. The remaining variables are selected and thus relevant according to the model. The key feature of PPRV methods is that they may adjust the complexity of the model (i.e. the number of latent variables, LVs) during the removal of variables, whereas many other variable selection methods optimize the number of LVs based on the full-spectrum model and do not alter this anymore during variable removal.

In general, PPRV methods start with building a PLS model on the complete training data set. Wold's criterion is used to optimize the number of LVs during cross-validation [12]: if the difference in RMSECV (Root Mean Square Error of Cross-Validation) between a model based on a and $a+1$ LVs is less than 2%, a LVs are selected as optimal.

Using a predictive property of the model (e.g. regression coefficients, loadings, VIP score), the variable having the lowest importance to the model is removed from the data and a new model is built. This procedure continues until all but one variable have been removed from the data. Model performance (RMSECV) for each model is stored during removal of the variables. When this process is finished, the lowest RMSECV is obtained from among all models built ($RMSECV_{\min}$). This is considered the optimal model. However, it may be that models with even more removed variables are not statistically different in terms of RMSECV from this optimal model. This is evaluated by using a one-tailed F-test:

$$RMSECV^2_{\text{crit}} = F_{\alpha,N_{\text{train}},N_{\text{train}}} \times RMSECV^2_{\min} \quad (1)$$

In this equation, $F_{(\alpha,N_{\text{train}},N_{\text{train}})}$ is given at significance level α (in this work, $\alpha = 0.05$). N_{train}, the number of samples in the training set, represents the degrees of freedom of both the numerator and denominator. In this way, models are sought with an even higher number of variables removed than the optimal model, while having an RMSECV not higher than $RMSECV_{\text{crit}}$. The final model is then the model with the most variables removed, while not differing significantly from the optimal model in terms of RMSECV.

To complete the procedure, a PLS model is built on the preprocessed training set with all variables removed as indicated by the final model. The same variables are also removed from the preprocessed test set and the PLS model is applied to it. The resulting RMSEP is used as response in the DoE.

In the foregoing, we have not yet described how the model complexity changes during variable removal and which predictive property to use for variable removal. Andries et al. investigated different ways of reducing the complexity during variable removal and also different predictive properties. For the former, it appeared that reducing model complexity with so-called FCAM led to the highest predictive accuracy [12]. In FCAM, variables are removed until the number of variables left equals the number of LVs as determined for the complete data set. From that moment, the complexity is reduced by one until the complexity equals 1 (and hence a single variable is left). Andries et al. furthermore found that variables should be removed based on the lowest absolute regression coefficient, since that led to the best predictive performance [13]. Therefore, in this work, we have used PPRV-FCAM with the absolute regression coefficients as predictive property.

2.1.3. Variable selection without PPRV-FCAM

The original DoE approach does not contain form of any variable selection. A filter method was applied to the results of the original approach, to show the advantages of entangling preprocessing selection with variable selection. Basically, this can be seen as a non-entangled version of preprocessing selection and variable selection. First, the complete PLS model is built and only then the relevant variables are determined. The VIP criterion [10] was chosen for this purpose, being one of the most commonly used variable importance methods in PLS. In this method, each variable receives a VIP score, based on its importance in the projections used to find n latent variables. A variable with a VIP score larger than a threshold of 1 is considered important.

3. Data

Two spectroscopic data sets with three different responses were analyzed in this work. The first data set originates from industrial practice and relates to latex samples, while the second is a publicly available data set about corn [22]. For both data sets, the true chemically relevant variables are known based on prior knowledge.

3.1. Latex data set

The latex data set consists of 196 near-infrared (NIR) spectra of acrylic latex samples, measured in aqueous conditions. The NIR spectra were recorded on a Bruker Matrix-F NIR spectrometer, coupled with optical fibers to an optical immersion probe. Spectra were acquired at 16 cm^{-1} resolution and addition of 64 scans per sample. Each spectrum contains 1037 variables, measured in a wavenumber range of 4000–12000 cm^{-1} (see Fig. 1). The spectral regions around 4200 cm^{-1} and 5000 cm^{-1} are relatively noisy because of the high absorbance in these regions. We did not remove these regions, because it would lead to a non-continuous signal, which may hamper appropriate preprocessing—especially derivatives are largely influenced by this. Moreover, variables in these regions should not be deemed relevant after appropriate preprocessing, so this provides an additional quality measure for our new approach.

For each sample, the concentrations of butyl acrylate (BA) and styrene (S) were measured in ppm units by headspace gas chromatography (GC) analysis. The true relevant variables in the NIR spectra are found at around 6160 and 6145 cm^{-1}, representing the vinylic C–H stretch overtone bands for BA and S, respectively [23,24]. The data set was randomly split in 150 training samples and 46 test samples; 10-fold cross validation (CV) was performed on the training samples to optimize the number of LVs, both for the original and new approach.

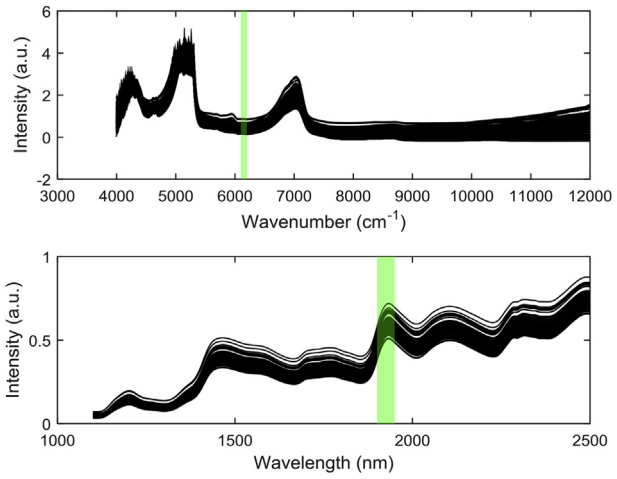


Fig. 1. Upper panel: latex data set. Lower panel: corn data set. In both panels, the shaded green area indicates the location of the known relevant variables. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Corn data set

For this data set, 80 corn samples have been measured using an 'm5' NIR spectrophotometer. The data set is freely available from the Eigenvector Research website [22]. The samples have been measured in the wavelength range 1100–2498 nm with 2 nm intervals, leading to 700 variables (see Fig. 1). Four response variables are provided in this data set. In this work, only the response 'moisture' is used. For dry food samples such as corn, it is known that they show absorption due to water at around 1900–1950 nm [12]. The data set was randomly split in 70 training samples and 10 test samples. Also here, 10-fold CV was performed for optimization of the number of LVs.

4. Results & discussion

All programming was performed using MATLAB (version 8.4.0 (R2014b), The MathWorks Inc., Natick, MA, USA).

4.1. Original approach (VIP) – latex data

Fig. 2 shows the main effects and second-order interaction effects for the latex data set based on the original DoE approach, including error bars highlighting the significance of effects based on 150 bootstrap samples. For prediction of BA, smoothing (Sm) and scaling (Sg) seem to be the relevant preprocessing steps, since they reduce RMSEP (i.e. they show a negative effect value). Baseline correction (B) only has a slightly negative effect value, but all interactions that involve B have a positive effect value, so B is excluded. Scatter correction (St) in itself is already not beneficial, and all its interactions are also either positive or insignificant.

For prediction of S, we can reason in a similar way that B, Sm and Sg are the relevant steps: St has a positive effect and a large positive effect for the interaction with Sg and is therefore excluded. All three other steps have negative effects and also their mutual interactions have negative effect values and are thus considered relevant.

After sequential optimization, we find that the most appropriate preprocessing strategy for prediction of BA consists of smoothing with Savitzky-Golay (window 11 px, 2nd order polynomial) and level scaling, leading to an RMSEP of 1228 (see Table 2). Similarly, for S we find smoothing with Savitzky-Golay (window 9 px, 2nd order polynomial) and level scaling, leading to an RMSEP of 1422 (Table 2). Figs. 3 and 4 show the variables that are determined relevant using VIP scores for prediction of BA and S, respectively. All variables above the dashed line are considered to be important.

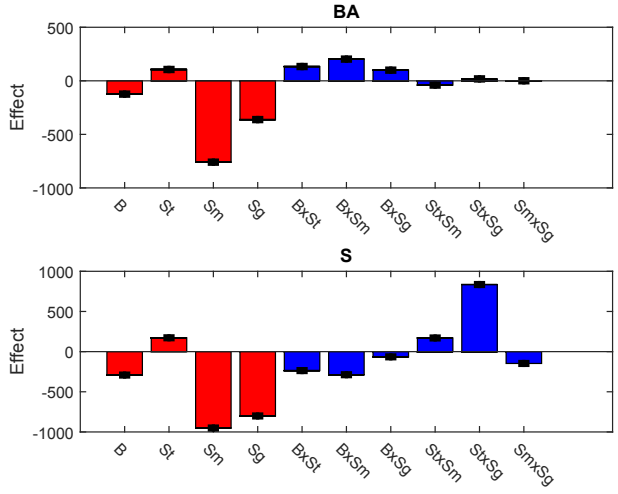


Fig. 2. Main effects and second-order interactions for the latex data set, based on the original DoE approach. Upper panel: prediction of butyl acrylate (BA), lower panel: prediction of styrene (S). The error bars are based on 150 bootstrap samples. The horizontal axis is labeled with B (baseline correction), St (scatter correction), Sm (smoothing) and Sg (scaling) and their second order interactions. Main effects are shown in red, interactions in blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 3 shows that the true relevant variables for BA (around 6160 cm^{-1}) are not determined relevant in a model based on the raw data. Appropriate preprocessing increases the importance of variables in that region, but many more variables are deemed important as well. This obviously hampers a correct interpretation of the model. Moreover, many variables have a higher VIP score than the true relevant variables (see e.g. around 5000 cm^{-1} and around 4200 cm^{-1}), indicating that the variables around 6160 cm^{-1} are not the most important ones in the constructed model. Because these variables are in the noisy regions, model interpretation should be done very carefully.

Similar observations hold for the important variables when predicting S (Fig. 4). Again, the true relevant variables are not deemed relevant in a PLS model on the raw data. After preprocessing, they become relevant, but many more relevant variables are found. Although both models have improved in terms of RMSEP after preprocessing (Table 2), they clearly have not improved in terms of model interpretation.

4.2. Entangling preprocessing selection and variable selection – latex data

Effect values for prediction of BA and S using the enhanced approach, i.e. by entangling variable selection and preprocessing selection, are given in Fig. 5. For prediction of BA, the preprocessing steps Sm and Sg are relevant. Also B is included, since the interactions with Sm and Sg have a negative effect. After sequential optimization, the most appropriate preprocessing strategy for BA prediction is found to be baseline correction with a 2nd derivative, followed by smoothing (window width 9 px, polynomial order 4) and meancentering. This strategy is different from the one found when using VIP, indicating that the addition of variable selection may influence the preprocessing strategy, as already outlined in the introduction section. The RMSEP of the corresponding model equals 475, much lower than the RMSEP value based on the full spectrum model with the most appropriate preprocessing (1228). When also taking RMSEP values of the raw data for BA into account (Table 2), we can conclude that the lowest RMSEP is obtained when entangling preprocessing selection with variable selection. Simultaneous optimization of a preprocessing strategy and variable selection thus clearly enhances model performance.

The effects for prediction of S are less straightforward to interpret. All main effects have a negative value and all interactions have a positive value. Therefore, we concluded that all preprocessing steps may be relevant and hence all are included in the sequential optimization step. The most appropriate strategy is ultimately found to be baseline correction via detrending with a 4th order polynomial, smoothing and meancentering. Just as with prediction of BA, this is a different strategy compared to the situation without variable selection. The corresponding RMSEP value is 494 (Table 2), again much lower than what was achieved with the original approach (1422). So, also in this case, the simultaneous optimization of preprocessing and variable selection is highly beneficial for the predictive performance of the model.

The selected variables for both BA and S using the enhanced approach are shown in Figs. 6 and 7, respectively. In both figures, one can see that the true relevant variables are selected after preprocessing: the variables corresponding to the vinylic BA band at

Table 2
Summary of results. For each data set, model performances of the original ('PLS') and extended approach are listed ('PPRV-FCAM'), together with the number of variables deemed relevant by both approaches. For the original approach, this has been determined by using VIP. 'Raw data' indicates the result without preprocessing and 'Appropriate preprocessing' the results after applying the DoE and sequential optimization of the relevant preprocessing steps.

Data	Method		RMSEP	No. Var	Selected variables
Latex BA	PLS	Raw data	3073	231	After preprocessing, true relevant variables are
		Appropriate preprocessing	1228	264	found (around 6160 cm^{-1}), but also many other irrelevant variables
	PPRV-FCAM	Raw data	1723	8	After preprocessing, true relevant variables are found (around 6160 cm^{-1})
		Appropriate preprocessing	475	6	
Latex S	PLS	Raw data	4460	214	After preprocessing, true relevant variables are
		Appropriate preprocessing	1422	242	found (around 6145 cm^{-1}), but also many other irrelevant variables
	PPRV-FCAM	Raw data	2010	8	After preprocessing, true relevant variables are found (around 6145 cm^{-1})
		Appropriate preprocessing	494	8	
Corn moisture	PLS	Raw data[a]	0.0051	261	No preprocessing required; true relevant variables
		Appropriate preprocessing[a]	0.0051	261	found (around 1900 nm), but also many more
	PPRV-FCAM	Raw data[b]	0.0003	2	No preprocessing required; true relevant variables found (around 1900 nm)
		Appropriate preprocessing[b]	0.0003	2	

[a,b]: These represent identical models, since no preprocessing was required.

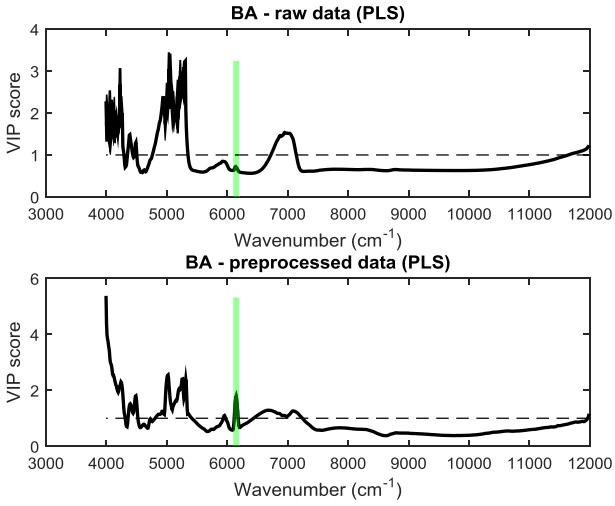


Fig. 3. Important variables for a PLS model built on the raw data (upper panel) and the appropriately preprocessed data (bottom panel) when predicting BA. Variables above the dashed line (VIP score 1) are considered important. The shaded green area indicates the location of the known relevant variables. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

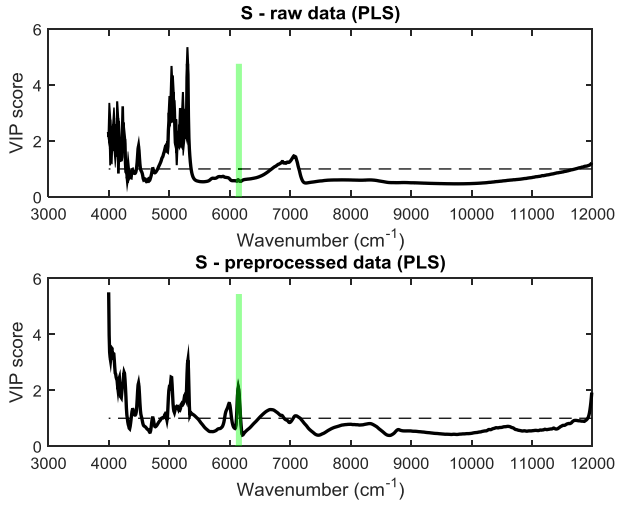


Fig. 4. Important variables for a PLS model built on the raw data (upper panel) and the appropriately preprocessed data (bottom panel) for the prediction of S. Variables above the dashed line (VIP score 1) are considered important.

6160 cm^{-1} are retained, as well as those for the vinylic S band at 6145 cm^{-1}. Without preprocessing, PPRV-FCAM models do not retain any of these variables (top panels in Figs. 6 and 7), indicating that appropriate preprocessing is required to highlight the true relevant variables. Moreover, *only* the true relevant variables are selected in the final models. There are no variables selected that are outside the known relevant region—and hence also not in the noisy regions—for prediction of either BA or S.

The enhanced approach thus improves both on predictive performance and in model interpretability, compared to the original approach. The lowest RMSEP values are found when simultaneously optimizing preprocessing and variable selection. Traditional PLS-based models were not able to unambiguously select the true relevant variables after appropriate preprocessing (original approach), since many uninformative variables were also deemed relevant. The enhanced approach, on the other hand, only selected the true relevant variables in combination with proper preprocessing, clearly showing the added value of entangling variable selection and preprocessing selection.

To further confirm these conclusions, we also applied PPRV-FCAM on data preprocessed with the preprocessing strategy found with the original approach. The majority of the variables deemed relevant in this way do not correspond to the true relevant variables (see Fig. 8). This again confirms that the selection of an appropriate preprocessing strategy and variable selection are strongly related and should therefore be entangled.

4.3. Corn data set

Fig. 9 shows the main effects and second-order interaction effects for predicting moisture in the corn data set using both the original and enhanced approach, including error bars highlighting the significance of effects. B and St decrease model performance for both approaches, judging from their positive effect values (i.e. increase in RMSEP). The only preprocessing step that may be slightly beneficial is smoothing (Sm) for the new approach, so this is the only preprocessing step considered relevant—Sg has a nonsignificant effect. For PLS-based models, no preprocessing seems to be required.

After sequential optimization of Sm, it appears that two different settings for the smoothing step lead to an equal RMSEP: no smoothing and smoothing using Savitzky-Golay with a window width of 5 px and a 4th order polynomial. The setting for no smoothing is chosen, such that the final models for both the original and new approach are built on the raw, meancentered data.

The true relevant wavelength region for this data set is

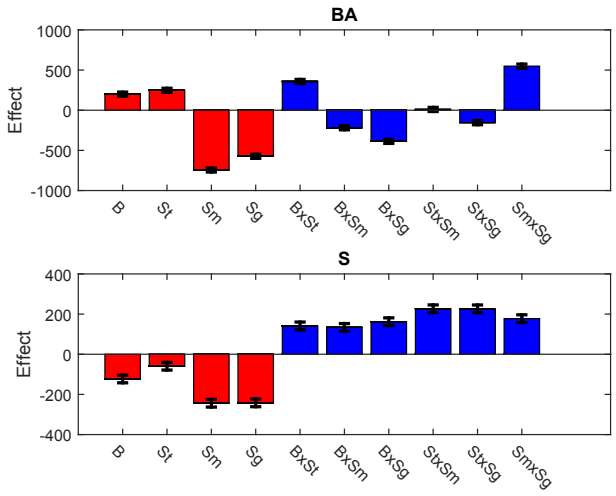


Fig. 5. Main effects and second-order interactions for the latex data set, based on the extended DoE approach. Top panel: butyl acrylate (BA), bottom panel: styrene (S). The error bars are based on 150 bootstrap samples.

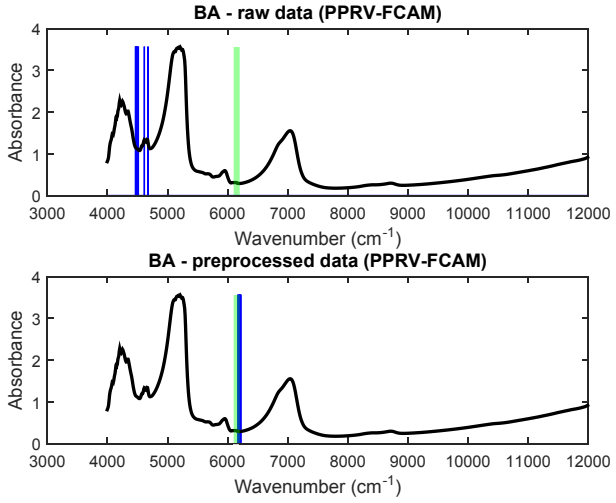


Fig. 6. Important variables for a PPRV-FCAM model built on the raw data (upper panel) and the appropriately preprocessed data (bottom panel) when predicting BA. The relevant variables according to the model are indicated with vertical blue lines; one of the original spectra is shown in black. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

1900–1950 cm^{-1}. Fig. 10 shows that both approaches indicate variables in this area. However, according to the VIP from the PLS model in the original approach, many more variables are important (in total 261 variables out of 700, see Table 2), which does not comply with the true relevant region. The new approach using PPRV-FCAM bases its regression model on just two variables—1908 cm^{-1} and 2108 cm^{-1}—and leads to a lower RMSEP as well (Table 2).

Also this data set shows the advantage of entangling preprocessing selection and variable selection, as is done in the enhanced approach. First, this approach has clearly shown that no preprocessing was required for this data set. Second, the final model is based on only two variables, of which one is in the known relevant interval, clearly enhancing model interpretability. Finally, predictive performance is improved by using the enhanced approach compared to using the original, PLS-based approach without variable selection.

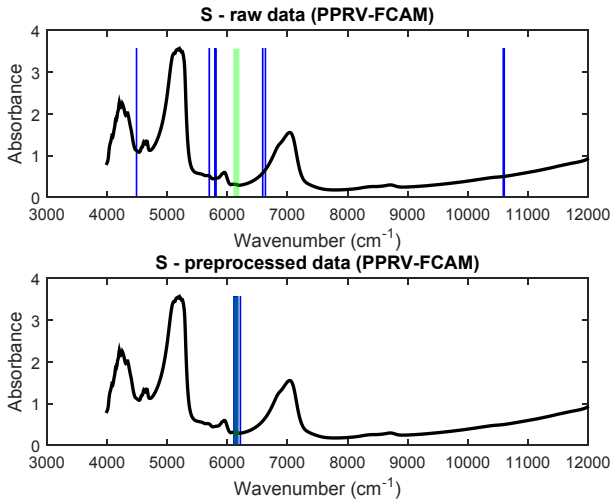


Fig. 7. Important variables for a PPRV-FCAM model built on the raw data (upper panel) and the appropriately preprocessed data (bottom panel) when predicting S. The relevant variables according to the model are indicated with vertical blue lines; one of the original spectra is shown in black. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.4. Other discussion points

Calculation time of the enhanced approach is somewhat longer compared to the original approach, taking approximately 30 min on a standard personal computer (original approach: 10–15 min). In the original approach, a single PLS model was built using cross-validation for each row in the DoE. In the new approach, however, many more PLS models need to be constructed for each row in the DoE, since a PLS model has to be rebuilt every time a variable is removed from the data. Since this rebuilding does not involve cross-validation to optimize the number of LVs, the increase in computation time is limited to approximately a factor two.

In this work, we have solely considered variable selection for interpretation of the model. However, more aspects may play a role in model interpretation [25]. One of these aspects is prior knowledge about the relevance or irrelevance of certain variables. This may occur, for example, when the data contain a region where detector saturation has taken place (i.e. known irrelevant variables). If such prior knowledge is available, the approach can be extended further to take this information into account as well.

For this purpose, one may add a second response variable to the DoE which expresses the 'quality' of the selected variables. For example, in the saturated region case, this second response variable could be represented by the percentage of all selected variables that are outside the saturated region. The higher this number, the more the selected variables comply with the prior knowledge. Of course, other definitions are possible as well.

For each of the two response variables (i.e. RMSEP and selected variable quality), effects can be calculated and interpreted separately. A user can then decide whether baseline correction should be performed if, for example, the effects indicate a little loss in variable quality, but a large gain in model performance.

It is also possible to combine multiple responses into a single response. In the context of DoE, this is often performed by using a desirability approach [26,27]. In such an approach, each response variable is transformed into a dimensionless value d between 0 and 1, and these are subsequently combined into a single response D, using

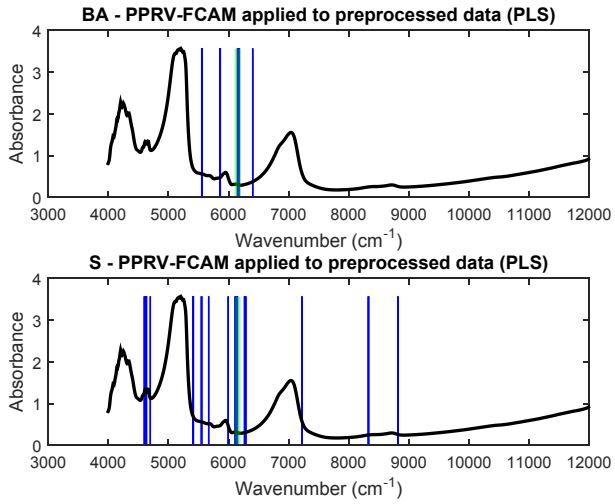


Fig. 8. Variables deemed relevant by PPRV-FCAM for the latex data (upper panel: BA, lower panel: S). The data were preprocessed with a strategy found using normal PLS.

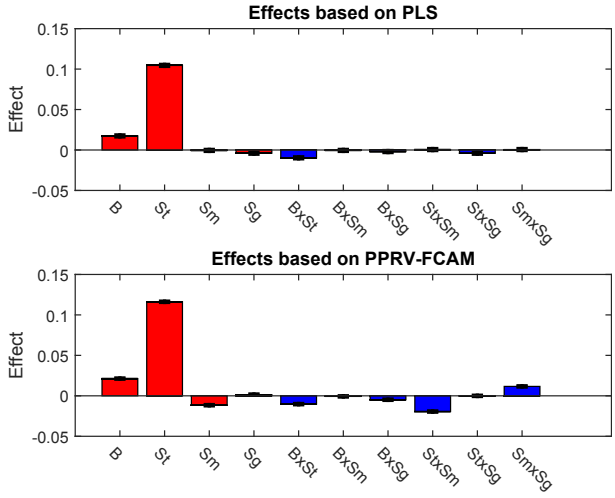


Fig. 9. Main effects and second-order interactions for the corn data set using the moisture response variable—upper panel: effects based on DoE and PLS, lower panel: effects based on DoE and PPRV-FCAM. The error bars are based on 150 bootstrap samples.

$$D = \sqrt[k]{\prod_{i=1}^{k} d_i} \quad (2)$$

The parameter k equals 2 for the current saturated region example. For correct interpretation, we would recommend to not only interpret effect values based on D, but also the effects based on the individual constituents of D.

5. Conclusion

In this work, we have shown that entangling preprocessing selection and variable selection enhances not only model performance, but also model interpretation. Our DoE-based preprocessing selection approach can be used to entangle these two aspects. The developed DoE-based approach is generic, such that different types of models, different types of variable selection methods and different preprocessing steps and methods can be incorporated into it. For illustration purposes, in this work we integrated variable

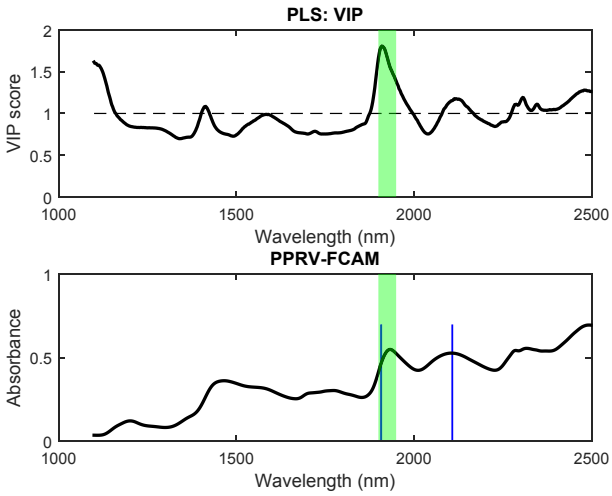


Fig. 10. Selected variables for PLS (upper panel) and PPRV-FCAM (lower panel) for the corn data set. In the upper panel, variables above the dashed are considered relevant, while in the lower panel the variables indicates with a red bar are relevant. Both figures are based on the raw data, since no preprocessing was required.

selection using PPRV-FCAM into the approach using PLS as model.

Our results showed that the entanglement of variable selection and preprocessing selection was beneficial for the construction of interpretable and accurate models. The predictive performance of PLS models improved when variable selection was used in the construction of the model. Secondly, appropriate preprocessing also led to an improvement in predictive performance. However, simultaneously optimizing variable selection and preprocessing is the most beneficial, since the lowest RMSEP values were obtained in this way.

Model interpretation did not improve when solely optimizing preprocessing or variable selection. Again, we were able to extract the true relevant variables from the data only when optimizing preprocessing and variable selection simultaneously. Therefore, to obtain accurate and interpretable models, we recommend combining the optimization of preprocessing with variable selection. In this work, we presented a generic approach for this purpose.

Acknowledgement

This research received funding from the Netherlands Organization for Scientific Research (NWO) in the framework of Technology Area COAST.

References

[1] Z.J. Wu, A review of statistical methods for preprocessing oligonucleotide microarrays, Stat. Methods Med. Res. 18 (2009) 533–541.
[2] A. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common preprocessing techniques for near-infrared spectra, Trac-Trend Anal. Chem. 28 (2009) 1201–1222.
[3] A. Rinnan, Pre-processing in vibrational spectroscopy - when, why and how, Anal. Methods-UK 6 (2014) 7124–7129.
[4] M. Daszykowski, I. Stanimirova, A. Bodzon-Kulakowska, J. Silberring, G. Lubec, B. Walczak, Start-to-end processing of two-dimensional gel electrophoretic images, J. Chromatogr. A 1158 (2007) 306–317.
[5] A. Smolinska, L. Blanchet, L.M.C. Buydens, S.S. Wijmenga, NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review, Anal. Chim. Acta 750 (2012) 82–97.
[6] J. Gerretzen, L.M.C. Buydens, A.O. Tromp-van den Beukel, E. Koussissi, E.R. Brouwer, J.J. Jansen, E. Szymanska, A novel approach for analyzing gas chromatography-mass spectrometry/olfactometry data, Chemom. Intell. Lab. Syst. 146 (2015) 290–296.
[7] J. Gerretzen, E. Szymańska, J.J. Jansen, J. Bart, H.-J. van Manen, E.R. van den

Heuvel, L.M.C. Buydens, Simple and effective way for data preprocessing selection based on design of experiments, Anal. Chem. 87 (2015) 12096–12103.
[8] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L.M.C. Buydens, Breaking with trends in pre-processing? TrAC Trends Anal. Chem. 50 (2013) 96–106.
[9] T. Mehmood, K.H. Liland, L. Snipen, S. Saebo, A review of variable selection methods in partial least squares regression, Chemom. Intell. Lab. Syst. 118 (2012) 62–69.
[10] M. Farres, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. Chemom. 29 (2015) 528–536.
[11] T.N. Tran, N.L. Afanador, L.M.C. Buydens, L. Blanchet, Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC), Chemom. Intell. Lab. Syst. 138 (2014) 153–160.
[12] J.P.M. Andries, Y. Vander Heyden, L.M.C. Buydens, Improved variable reduction in partial least squares modelling based on Predictive-Property-Ranked Variables and adaptation of partial least squares complexity, Anal. Chim. Acta 705 (2011) 292–305.
[13] J.P.M. Andries, Y. Vander Heyden, L.M.C. Buydens, Predictive-property-ranked variable reduction in partial least squares modelling with final complexity adapted models: comparison of properties for ranking, Anal. Chim. Acta 760 (2013) 34–45.
[14] M. Forina, C. Casolino, C.P. Millan, Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems, J. Chemom. 13 (1999) 165–184.
[15] J.A.F. Pierna, O. Abbas, V. Baeten, P. Dardenne, A backward variable selection method for PLS regression (BVSPLS), Anal. Chim. Acta 642 (2009) 89–93.
[16] V. Centner, D.L. Massart, O.E. deNoord, S. deJong, B.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration, Anal. Chem. 68 (1996) 3851–3858.
[17] 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, Appl. Spectrosc. 54 (2000) 413–419.
[18] K.A. Le Cao, D. Rossouw, C. Robert-Granie, P. Besse, A sparse PLS for variable selection when integrating omics data, Stat. Appl. Genet. Mol. 7 (2008).
[19] G. McLeod, K. Clelland, H. Tapp, E.K. Kemsley, R.H. Wilson, G. Poulter, D. Coombs, C.J. Hewitt, A comparison of variate pre-selection methods for use in partial least squares regression: a case study on NIR spectroscopy applied to monitoring beer fermentation, J. Food Eng. 90 (2009) 300–307.
[20] L. Xu, W.J. Zhang, Comparison of different methods for variable selection, Anal. Chim. Acta 446 (2001) 477–483.
[21] C. Abrahamsson, J. Johansson, A. Sparen, F. Lindgren, Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets, Chemom. Intell. Lab. Syst. 69 (2003) 3–12.
[22] NIR of Corn Samples, accessed December 2, 2015.
[23] D. Chicoma, V. Carranza, C. Sayer, R. Giudici, In line monitoring of VAc-BuA emulsion polymerization reaction in a continuous pulsed sieve plate reactor using NIR spectroscopy, Macromol. Symp. 289 (2010) 140–148.
[24] G.E. Fonseca, M.A. Dube, A. Penlidis, A critical overview of sensors for monitoring polymerizations, Macromol. React. Eng. 3 (2009) 327–373.
[25] F. Westad, F. Marini, Validation of chemometric models - a tutorial, Anal. Chim. Acta 893 (2015) 14–24.
[26] M.A. Bezerra, R.E. Santelli, E.P. Oliveira, L.S. Villar, L.A. Escaleira, Response surface methodology (RSM) as a tool for optimization in analytical chemistry, Talanta 76 (2008) 965–977.
[27] L.V. Candioti, M.M. De Zan, M.S. Camara, H.C. Goicoechea, Experimental design and multiple response optimization. Using the desirability function in analytical methods development, Talanta 124 (2014) 123–138.
[28] P.H.C. Eilers, Parametric time warping, Anal. Chem. 76 (2004) 404–411.
[29] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777.