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

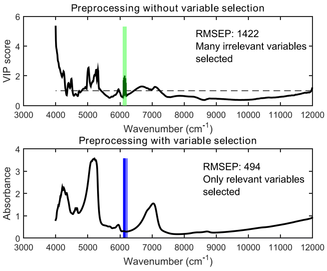
Jan Gerretzen [a, b], Ewa Szymańska [a, b], Jacob Bart [c], Antony N. Davies [c, d], Henk-Jan van Manen [c], Edwin R. van den Heuvel [e], Jeroen J. Jansen [a], Lutgarde M.C. Buydens [a, *]

[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*