Chemometrics and Intelligent Laboratory Systems 149 (2015) 53–65

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Fused stagewise regression — A waveband selection algorithm for spectroscopy☆,☆☆

B. Malli *, T. Natschläger

Software Competence Center Hagenberg, Softwarepark 21, 4232 Hagenberg, Austria

ARTICLE INFO

Article history:
Received 5 November 2014
Received in revised form 29 July 2015
Accepted 4 September 2015
Available online 11 September 2015

Keywords:
Robust variable selection
Wavelength selection
Fused Lasso
Double repeated cross validation
Spectroscopy

ABSTRACT

While partial least squares (PLS) and principal component regression (PCR), the most popular regression techniques in chemometrics, may theoretically be able to deal with large numbers of possibly correlated variables, as occurring in the analysis of spectroscopic data, the importance of performing some form of variable selection in practical applications has been widely discussed and acknowledged. In this work we address this problem via proposing a sparse regression algorithm, referred to as fused stagewise regression (FSR), which iteratively performs a selection of connected regions of variables (wavelengths), while being quite easy to implement and interpret, due to its resemblance to typical steps in iterative manual feature selection procedures. We evaluate the proposed variable selection technique on a publicly available benchmark data set and compare the performance of PLS models built on the determined selection to ones obtained by state-of-the-art feature selection methods from the fields of chemometrics and machine learning. In order to ensure robust feature selection, we integrate the individual selection methods into an extensive repeated cross validation procedure. For the data set under investigation, it is shown that FSR performs at least as good as state-of-the-art approaches and well within the range of variable selections provided by experts.

1. Introduction

A critical step in the analysis of spectroscopic data is the choice of an appropriate variable/wavelength selection used for chemometric modelling. Projection-based regression methods, such as PLS or PCR, are designed to deal with a very large number of possibly correlated variables. In practice, though, and particularly for near-infrared (NIR) and mid-infrared (MIR) applications, it is known that a removal of irrelevant and noisy wavelengths can improve prediction performance, yield simpler models and allow for better model interpretation (see e.g. [1,2] and references therein). Ideally, a determination of relevant variables should be based upon expert knowledge about the chemical properties of the substance under analysis. In practice, though, such a selection can only rarely be considered an objective process, as it is frequently heavily influenced by the experience of the expert carrying out the analysis. Thus, different experts may select significantly different wavelengths, possibly also resulting in highly deviating models and corresponding performances. This issue becomes particularly obvious when analysing highly complex target substances or properties, where a selection of relevant spectral regions solely based on chemical insight may be extremely difficult, if possible at all. Because of this, a universally applicable algorithm working independently of expert knowledge can often bear advantages.

This has been known for a long time and, as an exhaustive search of the perfect combination of variables is generally impossible, resulted in the use and development of a huge variety of different variable selection algorithms: starting with forward and backward selection methods [1], through genetic algorithms [3,4], up to, among others, interval PLS (iPLS) [5,6] or moving window PLS (MWPLS) approaches (see e.g. [2] and references therein). In contrast to techniques that pick combinations of single variables, iPLS and MWPLS select connected regions of wavelengths (= wavebands) and are thus of particular interest in spectroscopy, a field where adjacent variables are usually highly correlated.

In this contribution, we aim at the introduction of a new waveband selection algorithm that is based on the idea of the Fused Lasso [7], a regression method developed in the field of machine learning. In an iterative manner, the newly developed approach seeks for connected regions of variables relevant for predicting the target property at hand and weights them in an appropriate way, resulting in a vector of regression coefficients that can either be used for prediction as is done for any other conventional regression technique, but incorporating implicit waveband selection, or serve in a pre-step to regression, purely to determine important input variables. In order to allow for a fair comparison with state-of-the-art variable selection methods, only the latter

☆ The research reported in this article has been partly supported by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry of Science, Research and Economy, and the Province of Upper Austria in the frame of the COMET center SCCH.

☆☆ Selected papers presented at the 3rd European Conference on Process Analytics and Control Technology, 6–9 May 2014, Barcelona, Spain.

* Corresponding author. Tel.: +43 7236 3343 832.
E-mail addresses: Birgit.Malli@scch.at (B. Malli), Thomas.Natschlaeger@scch.at (T. Natschläger).

http://dx.doi.org/10.1016/j.chemolab.2015.09.004
0169-7439/

Table 1
Method specific parameters to be determined in inner CV loop.

Method	Parameters
FusedStage	$\tau \in [0, 1]$: Minimum correlation parameter $\eta \in [0, 2]$: Relative step size $\phi \in [0, 1]$: Relative shrinkage $\delta \in [0, 1]$: Only for FusedStageCUT; Cutoff parameter
GAPLS	None: The number of evaluations could be decided within the inner CV loop, but experiments have shown that the prior selection of a good value once in advance is possible. The same was found true for l, the number of variables to combine in mean building.
FusedSLEP	λ_1: l_1 − norm regularization parameter λ_2: Regularization parameter controlling the influence of the l_1 − norm on differences of successive regression coefficients.
Lasso	λ_1: l_1 − norm regularization parameter
ElNet	λ_1: l_1 − norm regularization parameter λ_{2el}: l_2 − norm regularization parameter
MWPLS	w: Window width ρ: Importance thresholding parameter Additionally, the maximum number of PLS components m could be chosen within the inner CV loop, but was found possible to be set suitably (10) in advance.
Expert	None
NoSel	None

approach[1] is studied and PLS is employed in the regression step.[2] Results for a benchmark data set from literature are given in Section 7. The complete algorithm, its basic ideas and recommendations for parameter selection, possibly incorporating expert knowledge, are presented in the following sections.

2. Background — Lasso and the Fused Lasso

Our proposed fused stagewise regression algorithm is based on the idea of a regression technique referred to as the Fused Lasso [7]. This technique represents a generalization of the popular Lasso method [8] which solves the optimization problem stated in Eq. (1), where $\boldsymbol{X}$ denotes the mean centred $n \times p$ design matrix (rows corresponding to samples and columns to variables/wavelengths) — in a spectroscopic application each row of $\boldsymbol{X}$ corresponds to a spectrum, y the mean centred $n \times 1$ vector of target observations (reference vector), β the $p \times 1$ vector of regression coefficients and λ_1 a non-negative regularization parameter determining the influence of the l_1 regularization term.

$$\beta^L = \underset{\beta \in \mathbb{R}\, p}{\operatorname{argmin}} \left(\|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right). \tag{1}$$

Thus, the Lasso aims for a solution, β^L, minimizing the sum of squared residuals, while attempting to keep $\lambda_1\|\beta\|_1 = \sum_{j\,=\,1,\ldots,p} \|\beta_j\|$ small. It has been shown [8] that the use of the l_1 norm in the penalty term $\lambda_1\|\beta\|_1$ introduces sparsity, i.e. it enforces solutions β^L with $\beta_j = 0$ for many coordinates j, thereby resulting in an implicit variable selection. A closer look at the solution reveals, however, that in case of correlated adjacent variables, as found in spectroscopy, the Lasso tends to select a set of single wavelengths instead of connected regions. From a chemical (thinking in terms of functional groups) and statistical (use of redundancies) point of view, the selection of wavebands, i.e. connected regions of wavelengths, would be favourable.

This is actually the idea behind a method referred to as the Fused Lasso [7] which incorporates an additional regularization term and solves

$$\beta^{FL} = \underset{\beta \in \mathbb{R}\, p}{\operatorname{argmin}} \left(\|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j| \right). \tag{2}$$

As was already the case for the Lasso solution, the term $\lambda_1\|\beta\|_1$ introduces sparsity in the regression coefficients, resulting in $\beta_j = 0$ for certain j. In the same manner, the use of the l_1 norm in the additional penalty term $\lambda_2 \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j|$ introduces sparsity in the differences $(\beta_{j+1} - \beta_j)$ of consecutive coordinates, resulting in $\beta_{j+1} = \beta_j$ for many j (i.e. β_js to be equal to their immediate neighbouring entries). As such, the combination of these two penalty terms leads to a piecewise constant solution β^{FL} containing connected regions of zero and non-zero entries, where the latter ones exhibit only a limited number of different values and only few significant jumps between neighbouring coefficients [7].

Due to this property, the Fused Lasso appears particularly suitable for employment in spectroscopic analysis. The problem of solving the optimization objective in Eq. (2), though, was found to be considerably more challenging than the solution of the Lasso problem [9–11]. An excellent review of existing algorithms, incorporating information on applicability, limitations and computational considerations, was provided by Yu et al. [11]: For a long time only special cases, demanding a certain structure or properties of $\boldsymbol{X}$, could be solved efficiently [12–14]. Solutions for a more general setting, including the one occurring in spectroscopy, could only be provided via the introduction of numerous auxiliary variables and constraints [7] in order to enable the application of a general purpose quadratic programming solver. As has been noted by Tibshirani et al. as well as others, e.g. [15,11], this computational approach does not scale well with the number of considered variables, p. Concretely, it is stated that for values of $p > 2000$ and $n > 200$ speed could become a practical limitation, particularly if five or tenfold cross validation is carried out. In 2010, Liu et al. published an algorithm for the efficient solution of this problem [15]. The complexity of the proposed algorithm, though, makes it difficult to follow propagated steps and interpret finally selected bands. Recently developed algorithms [16–18,11] bear the advantage to provide solutions to an even more general optimization problem, but are, in our opinion, equally complex and difficult to follow. This led us to develop a forward stagewise regression (FSR) algorithm that is based on the idea of the Fused Lasso but designed in a way that allows for considerably easier interpretation and implementation. A comparison of finally selected wavebands will be carried out for the analysed data set in Section 7.

Table 2
Criteria used to determine best parameters/parameter combinations.

Method	Criterium
FusedStage	Minimization of the FSR MSECV (i.e. the MSECV obtained if the FSR predictions are compared against the measured reference values; same notation also for methods below).
GAPLS	None
FusedSLEP	Minimization of the (SLEP) Fused Lasso MSECV.
Lasso	SLEP: minimization of the (SLEP) Lasso MSECV. MATLAB: Determine the largest λ_1, for which the corresponding (Matlab) Lasso MSECV is within one standard error of the minimum (see also http://www.mathworks.de/help/stats/lasso.html). This value is referred to as $\lambda_{1,1SE}$ below.
ElNet	SLEP: minimization of the (SLEP) Elastic Net MSECV. MATLAB: determine $\lambda_{1,1SE}$ to all available λ_{2el} and compute (Matlab) elastic net MSECVs for all ($\lambda_{1,1SE}$, λ_{2el}) pairs. Use the combination with minimal obtained MSECV.
MWPLS	Minimization of the PLS MSECV obtained using the selected variables only and a fixed number (set to 10 in results below) of PLS components.

[1] The former approach is intended to be studied in a follow-up publication.
[2] The same approach is adopted for all other discussed regression techniques performing implicit variable selection (i.e. for the Lasso, Elastic Net and Fused Lasso).