Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 78 (2005) 74–80

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Comparison of PLS algorithms in gasoline and gas oil parameter monitoring with MIR and NIR

Cândida C. Felício[a], Lígia P. Brás[a], João A. Lopes[a], Luís Cabrita[b], José C. Menezes[a,*]

[a]Centre for Chemical and Biological Engineering, IST, Technical University of Lisbon, Av. Rovisco Pais, P-1049-001, Lisbon, Portugal
[b]Quality Control, PETROGAL Sines Refinery, P-7521-952 Sines, Portugal

Received 30 March 2004; received in revised form 22 December 2004; accepted 24 December 2004
Available online 25 February 2005

Abstract

The objective of the present work was to compare different partial least squares algorithms upon mid- and near-infrared data. The applied techniques were single PLS, multiblock PLS and serial PLS. The comparison was made upon values of Q_y^2 determined in leave one out validation, mean square error of prediction using 80% of the data as calibration set and 20% as validation, and 95% confidence intervals for this parameter. A comparison between regression coefficients for all algorithms was also performed, after selecting the number of latent variables.

In order to perform this study, three parameters were used: flash point in gas oil, benzene and research octane number in gasoline.

Serial PLS gave the best results in all analysed cases, followed by one single PLS with MIR or NIR. Multiblock PLS gave intermediate results between both single PLS. However, for the studied parameters, the best calibration model was single PLS, since the results were quite accurate and achieved in less time.

Keywords: Gasoline; Gas oil; Partial least squares; Multiblock partial least squares; Mid-infrared spectroscopy; Near-infrared spectroscopy

1. Introduction

Recently, Brás et al. [11] evaluated the potential benefits of combining NIR and MIR spectra for the prediction of protein, moisture, fat and fiber content of soybean flour. These authors compared MIR- and NIR-based partial least squares (PLS) models with those obtained with multiblock PLS where both spectral data sets were combined for the modelling of the same flour's property.

The present work applies a similar approach to assess the advantages of combining NIR and MIR spectroscopic methods for the prediction of important parameters of gasoline (RON, benzene) and gas oil (flash point). Furthermore, the modelling performance of the different PLS-based regression techniques was compared: single PLS (PLS), multiblock PLS (MB-PLS) and serial PLS (S-PLS).

There are a number of parameters in gasoline and gas oil that must be analysed before these products can be commercialised, either due to vehicles motors specifications or to environmental requirements. The gasoline and gas oil samples analysed in the present work were obtained at the most important Portuguese Refinery (Petrogal's refinery from GALPENERGIA S.A.).

Throughout the latest years, directives have been approved concerning the limits of the research octane number (RON) and benzene; while flash point is an important parameter relating engine's performances, hence the importance of accurate measuring methods for these parameters.

The octane number is a fuel-performance property of gasoline that indicates the resistance of a motor fuel to knock. This property is determined by the measurement of a standard knock intensity in a standardized single cylinder variable compression ratio engine, where the

---

* Corresponding author. Tel.: +351 218 417 347; fax: +351 218 419 197.
E-mail address: bsel@ist.utl.pt (J.C. Menezes).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2004.12.009

sample's performance is compared to accepted references, as specified in ASTM D2699 Method [1]. As a gasoline additive, benzene increases the octane rating and reduces knocking. Its negative health effects and the possibility of benzene entering the groundwater have led to a strict regulation of gasoline's benzene content. Benzene is determined by gas chromatography, accordingly to the ASTM D3606 Method [2]. The flash point is the temperature at which the gas oil will ignite inside the cylinder. It is important that this parameter is within certain limits: if the temperature is higher than the specified, the gas oil will not ignite, and if it is lower, the engine is wasting energy. The flash point takes about 20 min to determine in traditional analysis, ASTM D93 [3].

Gasoline and gas oil are clear liquids and are easily measurable by vibrational spectroscopic methods shortening the analysis time and utilizing smaller sample volumes and cheaper measurement instruments. Hence, vibrational spectroscopic techniques such as near-infrared (NIR) and mid-infrared (MIR) spectroscopy combined with multivariate calibration methods have been extensively used in the analysis of physical and chemical properties of these petrochemical products [4–9].

The MIR region of the electromagnetic spectrum (4000–400 cm^{-1}) is related to transitions between vibrational states of molecules, while overtone and combination bands of molecular vibrations are responsible for the NIR spectrum (12,500–4000 cm^{-1}) [10]. For this reason, the peaks of a MIR spectrum are more specific, sharper and better defined that those of a NIR spectrum.

2. Theory

2.1. Partial least squares

In partial least squares, the predictor block **X** and the response vector ***y*** are decomposed using a given number of latent variables [12,13], according to Eqs. (1) and (2):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

$$\boldsymbol{y} = \mathbf{T}\boldsymbol{q}^{\mathrm{T}} + \boldsymbol{f} \tag{2}$$

T and **P** are, respectively, the scores and loadings for **X**, $\boldsymbol{q}$ is the loading vector for $\boldsymbol{y}$, and **E** and $\boldsymbol{f}$ are the residuals for **X** and $\boldsymbol{y}$, respectively. These residuals are obtained by deflation for each new latent variable $r(\mathbf{E}_r=\mathbf{E}_{r-1}-t_r p_r^t$ and $\boldsymbol{f}_r=\boldsymbol{f}_{r-1}-t_r \boldsymbol{q}_r)$.

2.2. Multiblock partial least squares

Multiblock partial least squares regression consists on the application of partial least squares on explanatory variables blocked into meaningful subsets for the prediction of the same parameter [14–16]. In the present case there are two predictor blocks generated by different measurement methods, corresponding to the NIR and MIR spectral data sets. In terms of prediction, the results of MB-PLS and single PLS are the same when the latter is applied on a matrix consisting of both blocks. The advantage of multiblock is the combination of information from different sources or methods, and in particular, the separation of each block's contribution to the description of the response variable. The method employed in this work was proposed by Westerhuis and Coenegracht [16]. Eqs. (1) and (2) also apply to multiblock PLS, but in this case, **T** is the matrix of super-scores.

The major difference between the MB-PLS and simple PLS algorithms is that MB-PLS determines a matrix of scores, called super-scores (**T**) which contains the scores for each block (S) computed separately.

In the multiblock algorithm, the block-weights (u) and the block-scores (s) are determined for both blocks (MIR and NIR). In Eqs. (3), (4), (6) and (7), the subscript r is the latent variable. In the first iteration (r=1), matrices $\mathbf{E}_0^{\mathrm{MIR}}$ and $\mathbf{E}_0^{\mathrm{NIR}}$ are the same as the original $\mathbf{X}^{\mathrm{MIR}}$ and $\mathbf{X}^{\mathrm{NIR}}$ spectra matrices and $\boldsymbol{f}_0$ is the original vector of the quality variable (flash point, benzene or RON). Eq. (3) describes how the weights for each block (block-weights) are obtained. The matrices containing each block of deflated data ($\mathbf{E}^{\mathrm{MIR}}$ and $\mathbf{E}^{\mathrm{NIR}}$) are regressed against the residuals for the independent variable $\boldsymbol{f}$ yielding the block-weights (u^{MIR} and u^{NIR}). The block-scores are obtained multiplying each block of deflated data by the corresponding block-weights.

$$u_r^{\boldsymbol{b}} = \frac{\mathbf{E}_{r-1}^{\boldsymbol{b}}\boldsymbol{f}_{r-1}}{\|\mathbf{E}_{r-1}^{\boldsymbol{b}}\boldsymbol{f}_{r-1}\|} \tag{3}$$

$$s_r^{\boldsymbol{b}} = \mathbf{E}_{r-1}^{\boldsymbol{b}}u_r^{\boldsymbol{b}} \tag{4}$$

The block-scores are then paired according to Eq. (5).

$$S_r = \left[s_r^{\mathrm{MIR}}|s_r^{\mathrm{NIR}}\right] \tag{5}$$

Using the block-scores the super-weights (w) and the super-scores (t) are obtained by regression.

$$w_r = \frac{S_r^t\boldsymbol{f}_{r-1}}{S_r^t\boldsymbol{f}_{r-1}} \tag{6}$$

$$t_r = S_r w_r \tag{7}$$

Once the super-scores are created, the algorithm proceeds like in single PLS.

In order to make new predictions ($\hat{\boldsymbol{y}}$), a regression vector ($\boldsymbol{b}$) can be obtained like in simple PLS. New data is paired in order to obtain a global matrix ($\mathbf{X}$=[$\mathbf{X}^{(\mathrm{MIR})}|\mathbf{X}^{(\mathrm{NIR})}$]), which can be used to obtain new predictions (Eq. (8)).

$$\hat{\boldsymbol{y}} = \mathbf{X}\boldsymbol{b} \tag{8}$$