Available online at www.sciencedirect.com

SCIENCE DIRECT

Chemometrics and Intelligent Laboratory Systems 75 (2005) 91–99

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Multiblock PLS as an approach to compare and combine NIR and MIR spectra in calibrations of soybean flour

Lígia P. Brás, Susana A. Bernardino, João A. Lopes, José C. Menezes *

Centre for Chemical and Biological Engineering, IST, Technical University of Lisbon, Av. Rovisco Pais, P-1049-001, Lisbon, Portugal

Received 3 March 2004; received in revised form 11 May 2004; accepted 26 May 2004
Available online 10 August 2004

Abstract

The present work is aimed at investigating the potential benefits of simultaneously combining near-infrared (NIR) and mid-infrared (MIR) spectral regions for use in calibration development of soybean flour quality properties (crude protein and moisture). NIR and MIR spectra were analysed separately using single partial least squares (PLS), and then both spectral data sets were utilized together in the modelling by applying two multiblock methodologies based on PLS regression: Multiblock PLS (MB-PLS) and Serial PLS (S-PLS). Utilizing the concept of net analyte signal (NAS), models constructed from NIR or MIR data were compared in terms of analytical figures of merit (sensitivity (SEN), selectivity (SEL) and limit of detection (LOD)).

When utilized alone, the MIR spectra gave models with considerably inferior prediction power and analytical figures of merit than NIR-based models. The multiblock methodology revealed to be very useful, since it helped to determine if there was distinctive information in each spectral data set and evaluate the relative importance of each data set. The results pointed out the existence of additional information in the MIR spectra not present in the NIR spectra. Although several works have already been reported comparing NIR and MIR spectroscopic techniques using different multivariate regression techniques, at the best of our knowledge none of them applied the approach of multiblock PLS. Therefore, the present work intends to explore this direction, using the NIR and MIR spectra as predictor blocks to model flour's properties in a parallel mode (MB-PLS) or in a serial mode (S-PLS).

Keywords: Multiblock PLS; Near-infrared spectroscopy; Mid-infrared spectroscopy; Complex natural raw materials; Soybean flour

1. Introduction

Soybean flour is composed of starch, protein, moisture and a small portion of lipids and fibre. It is extensively used in the food industry as additive and to increase the nutritional value of processed products. All the abovementioned properties characterise the nutritional value of the flour and, therefore, its quality. In Portugal, maize and wheat are the most utilized cereals for the preparation of animal rations; soybean is also used, even though this cereal is more expensive because Portugal is deficient in its production having a strong dependence on external markets (specially from Latin America). In Portugal, the crude protein content of soybean flour utilized in animal rations is determined from the nitrogen content measured by the Kjeldahl method (Portuguese standard NP 2030 (1996)), whereas the moisture content is determined by the loss of weight after drying (Portuguese standard NP 875 (1994)), in conformity with the European Directives. These methods are time consuming and require the use of chemical reagents. In contrast, near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques are fast and non-destructive, offering some advantages over conventional methods that had made them widely used [1,2]. However, spectroscopic methods require calibration before they can be used for quantitative measurements. This involves comparing the spectral measurements with the corresponding values of the property of interest as determined by a reference or standard method. Although the construction of the calibration model using chemometric tools can be time consuming, once the model is build, analyses can be made within a short period of time [1–3].

* 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 © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2004.05.007

The NIR range of the electromagnetic spectrum extends from 12,500 to 4000 cm^{-1} and is flanked by the mid-infrared region (4000–400 cm^{-1}) to longer wavelengths. While the MIR region relates essentially to transitions between vibrational states of molecules, the NIR spectrum records the so-called overtone and combination bands of molecular vibrations [2]. This fundamental difference makes interpretation of specific functionality easier for MIR spectra (since NIR bands are broader and less defined) and can be seen as an advantage for the replacement of NIR by MIR spectroscopy. Previous works have shown that MIR spectroscopy in the reflectance mode can perform as well or even better than NIR spectroscopy for the determination of properties of wheat samples [4], forages and by-products [5–7].

This work compares and combines MIR reflectance and NIR diffuse reflectance spectroscopy to quantitatively determine the crude protein and moisture content of soybean flour utilized in animal feeds. Firstly, in order to assess the individual modelling ability of each spectroscopic technique, partial least squares (PLS) regression models were constructed for the flour components considering the MIR and NIR data sets separately. By applying the concept of net analyte signal (NAS), the best PLS models obtained for each spectral data set were compared in terms of sensitivity, selectivity and limit of detection. Then, in order to investigate the presence of unique and relevant information in each spectral data set, a multiblock approach was applied, where both NIR and MIR spectra were used together for modelling, in a parallel mode (Multiblock PLS, MB-PLS) or in a serial mode (Serial-PLS, S-PLS). It should be noticed that, at the best of our knowledge, this is the first work that applies the concept of multiblock PLS regression to compare NIR and MIR spectroscopy.

The concept of using several predictor blocks in the principal component analysis and in PLS regression was introduced by H. Wold in 1982 [8]. In MB-PLS, the descriptors are split into meaningful blocks of variables that are then related to the response block in a parallel manner [9,10]. MB-PLS can improve models' interpretability, being potentially very useful in modelling and monitoring a variety of chemical processes [11–13]. Recently, we have applied the concept of MB-PLS for the modelling of an industrial pharmaceutical process [14,15]. The first work considered the active pharmaceutical ingredient (API) production stages (inoculum growth and fermentation), whereas in the latter work the API's isolation stages were also encompassed. S-PLS can be seen as an alternative multiblock PLS algorithm where the separated predictor blocks are modelled serially, i.e. the block models are calculated using the $\mathbf{Y}$ residuals from the previous block model [16]. Since the blocks are treated separately in S-PLS, it is possible to determine if additional blocks have any significant modelling power.

The concept of NAS was proposed by Lorber [17] as the basis of a new procedure for characterising analytical performance of multivariate calibrations where there is no single variable (wave number) that is unique or fully selective for a particular analyte. NAS describes the part of a spectrum that the model relates to the predicted quantity. This concept received much interest in multivariate calibration (see Ref. [18] for a review on NAS definitions and calculation methods), since NAS calculations can be utilized for estimating analytical figures of merit, such as sensitivity, selectivity and limit of detection that can be used for method comparison or to study the quality of a given analytical technique.

2. Theory

2.1. Notation

Throughout this work, matrices are represented by bold capital letters (e.g. $\mathbf{X}$), vectors are columns and are noted in bold lowercase letter (e.g. $\mathbf{x}$), vector and matrix elements in non-bold letters and indices in italic subscript characters (e.g. y_i and X_{ij}). Predictions are represented with a hat (e.g. $\hat{\mathbf{y}}$). The dimensions of some relevant matrices are: the calibration data matrix $\mathbf{X}$, $(I \times J)$ and the vector of calibration concentrations for analyte k, $\mathbf{y}_k$, $(I \times 1)$, where I is the number of calibration samples and J is the number of recorded wave numbers. The subscript i indicates the ith calibration sample (e.g. x_i). In Theory dedicated to the figures of merit, the subscript A (indicating the number of latent variables) will be used when it is necessary to distinguish between two quantities represented by the same letter (e.g. $\mathbf{x}$ and $\mathbf{x}_A$). The symbol $\|\cdot\|$ denotes the Euclidean norm, $\mathbf{I}$ is an appropriately dimensioned identity matrix. The superscripts T, + and * denotes transposition, pseudo-inverse and net signal, respectively.

2.2. MB-PLS

Multiblock PLS is an extension of the PLS method, a class of regression models motivated by the attempt to find the relationship between explanatory and response variables by assuming that they are generated by a common set of underlying factors [19,20]. In multiblock PLS (MB-PLS), the predictors are separated into subsets or blocks, accordingly to a meaningful criterion or process knowledge. For prediction purposes, it is preferred to combine all variables in a single block, as in conventional PLS [21]. However, the MB-PLS strategy of blocking variables produces more interpretable results, as it allows focusing on separate subsets of the data. The algorithm employed in the presented work can be found in the paper of Westerhuis and Coenegracht [22]. As in PLS, the following equations apply to the MB-PLS method:

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

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