Chemometrics and Intelligent Laboratory Systems 164 (2017) 76–82

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Comparison of CCA and PLS to explore and model NIR data

Fernando Gatius*, Carlos Miralbés, Calin David, Jaume Puy

Departament de Química, Universitat de Lleida (UdL), Av. Rovira Roure, 191, Lleida 25198, Spain

ARTICLE INFO

Keywords:
Canonical Correlation Analysis
Partial Least Squares
NIR calibration
Regularization

ABSTRACT

Partial Least Squares (PLS) regression is the most widely used technique for developing NIR calibrations. PLS uses several factors to reach the optimum models which can be helpful in a physical interpretation of the sources of correlation between *x* and *y* variables. However, it suffers from later factors not arising in the order of the explained variance. Canonical Correlation Analysis (CCA) overcomes this problem by selecting the latent variables as the directions of maximum *x*-*y* correlation. Calibration of moisture, crude protein, dry gluten and resistance of dough to deformation of wheat flour samples from NIR spectra is here studied using PLS-1, PLS-2, CCA-1 and CCA-2. The calibration set contains 429 samples while 215 extra independent samples are used for the validation set. It is shown that a 2-D CCA-2 calibration model gathers the highest explained variance between the models studied. When particular calibration models of each property are compared, CCA requires regularization to avoid instability of the regression coefficients. A regularization term that tends to reduce the regression coefficients and the Durbin-Watson test or the Test for Runs to select the regularization parameter have been used. Both statistical tests led to similar values of the regularization parameter and the resulting regression coefficients and RMSEP of the CCA-1 models became similar to those of the PLS-1 models.

1. Introduction

NIR spectroscopy is nowadays recognized as a valuable technique for the quality control of very different materials (Davies and Garrido-Varo [1]). Accuracy, precision, short time of analysis and limited sample preparation are among the main advantages recognized to the NIR spectroscopy. In addition to the specific characteristics of NIR spectrometers, the development of calibration techniques has played a key role in the expansion of NIR applications. Partial Least Squares (PLS) regression (Wold [2,3]; Wold et al. [4]), first developed in the field of econometrics in the 1960s by Wold, is the regression technique commonly used for the prediction of quality parameters from the spectral information provided by NIR spectrometers. This technique is based on the assumption that the observed data of the dependent variable are generated by a process driven by a small number of factors or latent variables defined as the directions of maximum covariance between both independent and dependent variables (Martens and Naes [5]; Rosipal and Krämer [6]).

Commonly, robust and accurate PLS regression models require several factors reflecting the multivariate dependence of the property of interest. The study of the loadings of each factor allows the identification of the factor and the splitting of the total variance in the variance explained by each factor allows assessing the importance of each factor in the model. However, as the number of factors increases, a higher number of samples in the calibration and validation sets are required in order to ensure a complete and accurate description of the population of interest. Additionally, while we would like that the importance of successive PLS factors decrease in terms of the explained variance, it is found in some cases late PLS factors that show higher explained variance than earlier ones. This behavior, although reasonable in PLS-2, might suggest some inefficiency of the PLS-1 procedure.

Alternatively, there is another classical calibration procedure almost unknown for most NIR users at least if we consider the NIR applications reported in the literature. It is called Canonical Correlation Analysis (CCA) and it is based on defining as latent variables the directions of maximum correlation between the two sets of variables. CCA was developed by Hoteling [7] for regression purposes. Using the correlation as a driving force to define the latent variables, it ensures the maximum explained variance in each factor, a result quite attractive for determining the best model (Mardia et al. [8]). For the particular case of only one *y*-variable being described in the calibration model, the model is quite simple since only one factor is possible. As the explained variance is maximized, the sum of squares of the residuals is minimum and then the result coincides with that of multilinear regression (MLR). It is then clear that CCA models for one *y*-variable suffer from some drawbacks: instability in the regression coefficients and the need of a larger number of samples than *x*-variables. These requirements are overcome with the application of the so-called regularized CCA

* Corresponding author.
E-mail address: gatius@quimica.udl.cat (F. Gatius).

http://dx.doi.org/10.1016/j.chemolab.2017.03.011
Received 6 December 2016; Received in revised form 17 March 2017; Accepted 24 March 2017
Available online 28 March 2017

(Tikhonov et al. [9]; Guo and Mu [10]). In this case, which is the usual in the application of NIR spectroscopy, the solution is looked as a minimization of a least squares condition to which additional requirements called regularization factors are added to reduce the variance of the parameter estimates. It seems interesting to test these models poorly used in the development of NIR calibrations and to compare the resulting results with those of PLS.

Thus, both PLS and CCA techniques will be used to develop NIR calibration models of four quality parameters of wheat flour. Some advantages and drawbacks of using these techniques will be discussed. While PLS is a widely used technique to predict moisture and protein of wheat (Pawlinski and Williams [11]) and wheat flour (Delwiche [12,13]), CCA has never been applied to this kind of samples.

1.1. PLS and CCA principles

Let us introduce briefly PLS and CCA as minimization problems.

Let X and Y be the matrices of dimensions $n \times p$ and $n \times q$ respectively, whose columns correspond to variables and whose rows correspond to the samples (or experimental units). The vector x_k is the kth column of the matrix X and contains the measurements of samples on the independent variable x_k centered with respect to the average of each variable in the samples set. The vector y_j is the jth column of the matrix Y and contains the measurements of samples on the dependent variable y_j.

PLS focuses on the covariation between independent and dependent variables and tries to find the new variables (called the PLS factors) that maximize this parameter as is shown in Eq. (1):

$$\rho_{PLS} = \max_{wPLSx, wPLSy} \operatorname{cov}(Xw_{PLSx}, Yw_{PLSy}) \tag{1}$$

where ρ_{PLS} is the maximum covariance. This equation is subject to the constraint $w_{PLSi}^T w_{PLSi} = 1$ $(i = x, y)$, where w_{PLSx} and w_{PLSy} are the so called PLS loading (weight) vectors. This operation can be performed by means of the classical NIPALS algorithm (Martens and Naes [5]), but also performing eigenvalue-eigenvector decomposition (Lindgren et al. [14]). The maximum is achieved having w_{PLSx} and w_{PLSy} as the largest eigenvectors of the matrices $X^T Y\, Y^T X$ and $Y^T X\, X^T Y$ respectively. To obtain subsequent weights, the algorithm is repeated with deflated X and Y matrices (subtracting the contribution of each found PLS factor).

As Eq. (1) indicates, PLS factors are obtained ordered by decreasing values of the covariance between the PLS factor and the y-variable when only one y-variable is considered. This rank order can, then, be altered when the PLS factors are ranked according to the respective correlation (explained variance) to the y-variable. Notice that the correlation depends not only on the covariance but also on the standard deviation of the scores, this last influence being the responsible of solutions of Eq. (1) ranking in different order according to correlation.

On the other hand, CCA can be seen as the problem of finding basis vectors for two sets of variables such that the correlations between the projections of the variables into such basis vectors are mutually maximized. In other words, CCA consists in solving the problem of finding the so called CCA loading normalized vectors w_{CCAx} and w_{CCAy} that maximize the correlation between the linear combinations in both X and Y subspaces:

$$\begin{aligned} \rho_{CCA} &= \max_{wCCAx, wCCAy} cor\,(Xw_{CCAx}, Yw_{CCAy}) \\ &= \max_{wCCAx, wCCAy} \frac{\operatorname{cov}(Xw_{CCAx}, Yw_{CCAy})}{SDev\,(Xw_{CCAx})\,SDev\,(Yw_{CCAy})} \end{aligned} \tag{2}$$

where ρ_{CCA} is the first canonical correlation. If more than one dependent variable is involved in the study, higher order canonical variables and canonical correlations can be found as a stepwise problem, maximizing the correlation and ensuring orthogonality with previous CCA loading normalized vectors.

The CCA formulation is the most attractive procedure when the explanation of the y-variability through the x-variables is envisaged, since the correlation between y and x is just the variability of y that can be explained by x.

Which is, then, the reason for the low popularity of CCA among NIR users? The vectors w_{CCAx} and w_{CCAy} that are the solutions of the problem written in Eq. (2) are, in fact, the largest eigenvectors of the matrices $(X^TX)^{-1}X^TY\,(Y^TY)^{-1}Y^TX$ and $(Y^TY)^{-1}Y^TX\,(X^TX)^{-1}X^TY$ respectively while the subsequent CCA loading vectors are the eigenvectors of the same matrix in the order of decreasing eigenvalues. Notice the presence of the $(X^TX)^{-1}$ and $(Y^TY)^{-1}$ in the procedure that allows finding w_{CCAx} and w_{CCAy}. Thus, the collinearity of the x or y measurements amplifies the noise of the measurements in the calculation of w_{CCAx} and w_{CCAy} when the inverses of these matrices have to be computed. A usual way to reduce this problem is based on the Tikhonov regularization also known as Ridge Regression in the statistical field. In the regularized CCA, additional conditions are imposed in the inversion of the $(X^TX)^{-1}$ like, for instance, smoothness of the estimates when we expect a smooth contribution of the different wavelengths to the CCA loadings as it is the case in the NIR field. The most simple regularization factor is the addition to X^TX of the extra term λI, where λ is the so called regularization parameter that can be tuned by a validation procedure. This regularization term aims at finding an optimum model with the smallest norm of the estimate set of parameters. In this way, both, the propagation of the errors in x-measurements and the arising of spurious peaks in the regression coefficients are more controlled.

Different selection procedures for λ have been suggested in the literature. A first procedure consists in selecting λ so that the Root Mean Square Error of the model calculated on a prediction test, RMSEP, becomes minimum. In the present case, the application of this procedure for the calculation of the regularization parameter leads to models with very low RMSEP but noisy regression coefficients for all the properties. This indicates that with this strategy, the resulting regularization parameter is too small since we are only focusing our interest in reducing the RMSEP. We should then look for a higher λ so that regularization plays the desired role but still giving a statistically good fit to the data. Two procedures based on a test for the residuals of the model will be examined in this work to determine λ: the Durbin-Watson test (Draper and Smith [15]; Durbin and Watson [16]) and the Test for Runs (Draper and Smith [15]; Swed and Eisenhard [17]). The optimum model should have the highest λ keeping the residuals uncorrelated when the samples are ordered according to the value of the property under calibration.

2. Materials and methods

PLS and CCA methods have been applied to the calibration of quality parameters of wheat flour samples using NIR spectra.

2.1. Sample description

Samples used in this work are selected from the commercial samples received by a flour company that mills 900,000 kg of wheat per day. The origin is from different countries, mainly France and USA, and different years. From a set of 3000 samples, 429 were selected for the calibration, 215 for the validation and 100 for the prediction sets by looking at the scores of a PCA model from the spectral information. The selection was done in order to include samples covering all the area of the scores plot. Samples selected from the y-values were also included in order to cover the whole range of y-values. Characteristic values of these sets are given in Table 1.

2.2. NIR spectra

NIR spectra of the wheat flour samples were collected with a near infrared spectrometer FOSS NIR System 6500 using the ISIscan (version 2.83) routine operation software from FOSS. These spectra contain the