Chemometrics and Intelligent Laboratory Systems 148 (2015) 32–50

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Nonlinear partial least squares regressions for spectral quantitative analysis

Yanxia Wang [a], Hui Cao [a,*], Yan Zhou [b], Yanbin Zhang [a]

[a] State Key Laboratory of Electrical Insulation and Power Equipment, School of Electrical Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi Province, 710049, China
[b] School of Energy and Power Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China

ARTICLE INFO

Article history:
Received 11 February 2015
Received in revised form 28 August 2015
Accepted 29 August 2015
Available online 5 September 2015

Keywords:
Spectral quantitative analysis
The NPLS model
The internal NPLS
The external NPLS
The nonlinear components extracted NPLS
The kernel NPLS

ABSTRACT

As a traditional linear regression model, the partial least squares (PLS) could not handle the nonlinearities in spectral quantitative analysis. This paper focuses on four types of nonlinear partial least square (NPLS) models: the internal NPLS model, the external NPLS model, the nonlinear components extracted NPLS model and the kernel NPLS model. The internal NPLS model adopts the neural network as the nonlinear function to describe the inner relation. For the external NPLS model, the PLS regression is performed on the extended input matrix which contains the nonlinear terms of the independent variables. The nonlinear components extracted NPLS model extracts the nonlinear principal components by selecting the weight vectors with PLS, and then the nonlinear relationship between the nonlinear principal components and the dependent variables is established. For the kernel NPLS model, the original input is transformed into a high-dimensional space by the nonlinear kernel functions and the PLS regression model is built in the new feature space. The 10-fold root–mean–squares error of cross validation is the criterion to decide the optimal parameters of these models. The performance of the different regressions is demonstrated by three real spectral datasets: the meat dataset, the flue gas dataset of gas-fired plant and the flue gas dataset of coal-fired plant. The results suggest that the internal NPLS model with the radial basis function neural network, the external NPLS model with the radial basis function neural network and the kernel NPLS model with polynomial kernel function have a higher predictive ability for spectral quantitative analysis in most cases.

1. Introduction

Spectroscopy is a non-destructive, fast, and informative measurement technique [1,2]. In recent years, spectral analysis has increasingly been adopted in a variety of chemical fields such as in the pharmaceutical [3], food [4], biomedical [5], agricultural [6,7] and petrochemical [8] areas. Based on the wavelength signals, spectral analysis is to predict the component concentrations by regression models. Partial least squares (PLS) is a well-known multivariate regression algorithm for quantitative analysis of spectral data [9,10]. The traditional PLS only extracts the linear correlations between the independent variables (the wavelength signals) and the dependent variables (the component concentrations) [11]. In practice, however, the ambient conditions, the instrument variation, and the analyte characteristics may result in nonlinearities in spectra [12]. In order to improve the predictive accuracy, nonlinear partial least squares (NPLS) has been investigated [13,14]. There are many types of NPLS models and they could be classified as follows.

The first type is the internal NPLS model. Instead of the original inner linear function, it adopts nonlinear functions to describe the relationship between the latent variables [15]. The quadratic partial least squares algorithm modifies the relationship between the independent latent variables and the dependent latent variables to be nonlinear [16,17]. Nevertheless, because the predefined form of the quadratic function is fixed, the nonlinearity of the model may be limited [18]. With the spline inner function, the spline partial least squares is to fit nonlinearity, while this algorithm may suffer from over-fitting or local minima [19–21]. Neural network is an adaptive nonlinear dynamic system [22–24]. With a suitable combination of weights and activation functions, the neural network could approximate any nonlinear function in sufficient accuracy [25–30]. Using the neural network as the inner function, the internal NPLS would have the advantages of the robustness of PLS as well as the nonlinear capability of the neural network.

The second type is the external NPLS model with extended independent input. The independent input matrix is augmented to include the nonlinear terms of the independent variables. Then, PLS regression is performed on the extended input matrix [31,32]. The quadratic and cross-product combinations of input variables are used as additional input terms to PLS model to build nonlinearity in reference [33]. Since the selection of the nonlinear terms and the relevant coefficients rely on prior knowledge, this method is restricted in the application of spectral analysis. Reference [34] provides a version of NPLS, which

* Corresponding author.
E-mail address: huicao@mail.xjtu.edu.cn (H. Cao).

http://dx.doi.org/10.1016/j.chemolab.2015.08.024
0169-7439/

transforms the original variables from the input layer into the hidden layer with a neural network, and then PLS is utilized to relate the outputs of the hidden layer to the dependent variables. This algorithm may lose the information of the original input matrix. Considering the relationship between the independent variables and the dependent variables, we use the outputs of the hidden layer in a neural network as the nonlinear extensions of the input matrix. The nonlinear extensions and the original input variables constitute the new input matrix of the external NPLS model.

The third type is the nonlinear components extracted NPLS model which adopts two neural networks. By selecting the weight vectors with PLS, the algorithm uses a neural network to extract the nonlinear principal components, namely, the latent variables of independent variables [35,36]. Then the nonlinear relationship between the nonlinear principal components and the dependent variables is established by another neural network [37]. Principal component analysis is a linear technique mapping multi-dimensional data into lower dimensions [38]. As a combination of the principal component analysis and neural networks, nonlinear principal component analysis is proposed [39]. In reference [40], based on the neural networks, the linear PLS is integrated with the nonlinear principal component analysis to establish a robust estimation method. With the same approximate property as the neural network, the robust estimation method could handle nonlinear problem well.

The fourth type is the kernel NPLS model. According to Cover's theorem, the nonlinear data structure is more likely to be linear after high-dimensional nonlinear mapping [41]. For the kernel NPLS model, the original input data are transformed into a high-dimensional feature space by the nonlinear kernel functions. Then, a liner PLS regression model is established in the new feature space [42,43]. Several versions of kernel NPLS have been investigated in literature. The covariance kernel partial least squares is presented in [44]. Reference [45] describes the kernel partial least squares algorithm with quadratic and cubic polynomial kernel function. Gaussian kernel function, a special case of radial basis function kernels, is introduced to kernel partial least squares in [46]. The main advantage of the kernel NPLS method is that it does not involve nonlinear optimization procedure and could possess low computational complexity similar to that of linear PLS [47]. In addition, it can handle a wide range of nonlinearities by different kinds of kernel functions and adjustable parameters [48]. Whereas, the kernel NPLS is not resistant to bad leverage points (outlier in the space of input variables) [49].

In this paper, we emphasize four types of nonlinear partial least squares regressions for spectral quantitative analysis. They are the internal NPLS model, the external NPLS model, the nonlinear components extracted NPLS model and the kernel NPLS model. For the internal NPLS and external NPLS, two popular neural networks, the back propagation neural network and the radial basis function neural network are employed. The regression models in this article are listed as follows:

- The partial least squares (PLS)
- The back propagation neural network (BPNN)
- The radial basis function neural network (RBFNN)
- The internal NPLS model with BPNN (BPI-NPLS)
- The internal NPLS model with RBFNN (RBFI-NPLS)
- The external NPLS model with BPNN (BPE-NPLS)
- The external NPLS model with RBFNN (RBFE-NPLS)
- The nonlinear components extracted NPLS model (NC-NPLS)
- The kernel NPLS model with polynomial kernel function (Poly-KNPLS)
- The kernel NPLS model with Gaussian kernel function (Gaus-KNPLS)
- The kernel NPLS model with Sigmoid kernel function (Sig-KNPLS)
- The kernel NPLS model with exponential kernel function (Exp-KNPLS)
- The kernel NPLS model with Fourier kernel function (Fou-KNPLS).

To evaluate the effectiveness of these models, the root–mean–squares error of cross validation, the root–mean–squared error of prediction, the squared correlation coefficient of prediction and the squared correlation coefficient of cross validation are presented. The experiments are conducted with three real datasets: a meat dataset, a flue gas dataset of gas-fired plant and a flue gas dataset of coal-fired plant. The organization of this paper is as follows: Section 2 reviews the relevant algorithms. The experimental datasets and procedure are provided in detail in Section 3. In Section 4, the experiments results are discussed. Finally, Section 5 concludes the paper.

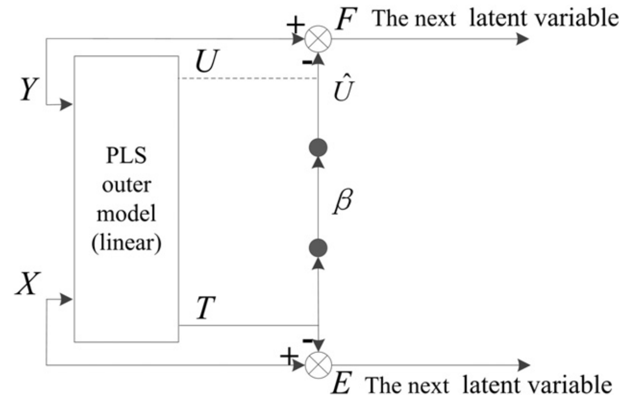

Fig. 1. The structure of the PLS model.

2. Relevant algorithms

For the quantitative analysis of spectral data, the absorbance of each wavelength is the independent input matrix $X \in R^{n \times m}$ and the component concentration is the dependent output matrix $Y \in R^{n \times s}$.

2.1. PLS

As is shown in Fig. 1, the regression equations of PLS can be demonstrated by:

$$\begin{cases} X = TP^T + E = \sum_{i=1}^{r} t_i p_i^T + E \\ Y = UQ^T + F = \sum_{i=1}^{r} u_i q_i^T + F \end{cases} \quad (1)$$

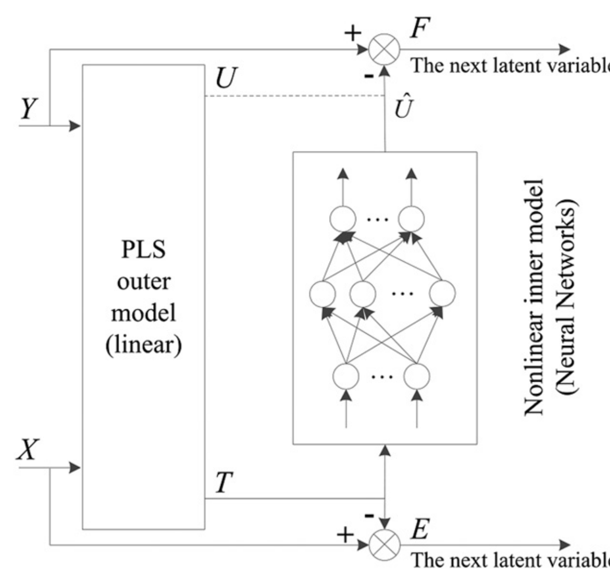

Fig. 2. The structure of the internal NPLS model.