Chemometrics and Intelligent Laboratory Systems 150 (2016) 58–64

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

The equivalence of partial least squares and principal component regression in the sufficient dimension reduction framework

You-Wu Lin [a], Bai-Chuan Deng [b], Qing-Song Xu [a,*], Yong-Huan Yun [c], Yi-Zeng Liang [c]

[a] College of Mathematics and Statistics, Central South University, Changsha 410083, PR China
[b] College of Animal Science, South China Agricultural University, Guangzhou 510642, PR China
[c] College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China

ARTICLE INFO

Article history:
Received 18 September 2015
Received in revised form 30 October 2015
Accepted 4 November 2015
Available online 1 December 2015

Keywords:
Central subspace
Sufficient dimension reduction
Principal component regression
Partial least squares

ABSTRACT

Partial least squares (PLS) and principal component regression (PCR) are two widely used techniques for dimension reduction in chemometrics. However, the relationship between PLS and PCR is not entirely understood. In this paper, we introduce the idea of sufficient dimension reduction (SDR) to chemometrics, and show that PLS and PCR are methods of SDR. Furthermore, this paper shows that these two methods are equivalent within the framework of SDR which means that there is no theoretical advantage of PLS over PCR in terms of prediction performance. The above conclusion is supported by the results of a simulated dataset and three real datasets.

1. Introduction

Over the last few years, science and technology advanced profoundly that they have changed all the fields and allow scientists to collect high-dimensional data sets such as genetic data, chemical data and spectroscopy data. A common feature of these datasets is that the number of variables (p) is much larger than that of the observations (n), which refers to the "large p, small n" problem [1,2]. Due to the "curse of dimensionality" [3], many classical statistical methods are no longer applicable. Variable selection [4–8] and dimension reduction [9–14] are effective techniques to solve the problem. And dimension reduction becomes increasingly important in high-dimensional data analysis, with the help of which, we can obtain more accurate and easier interpretation models.

There are some well-known dimension reduction methods aiming to construct linear combinations of predictors $\boldsymbol{\Gamma}^T\mathbf{x} = (\boldsymbol{\gamma}_1^T\mathbf{x}, \ldots, \boldsymbol{\gamma}_k^T\mathbf{x})^T (k < p)$ and to perform dimension reduction by replacing $\mathbf{x}$ with $\boldsymbol{\Gamma}^T\mathbf{x}$. Principal component regression (PCR) [9–11] and partial least squares (PLS) [12–14] are two kinds of dimension reduction methods. These methods are strongly promoted in chemometrics and widely applied in many kinds of fields, such as spectroscopy data [10,11,13] and microarray data analysis [16]. Other dimension reduction techniques known by statisticians but rarely used in chemometricians are continuum regression (CR) [17,18], sliced inverse regression (SIR) [19], sliced average variance estimation (SAVE) [20] and some other variations [21–23].

PCR constructs latent variables that preserve information as much as possible in $var(\mathbf{x}) = \sum_{\mathbf{x}}$. Especially, the i-th latent variable $\boldsymbol{\gamma}_i^T\mathbf{x}$ has the largest variance among all latent variables of $\mathbf{x}$ orthogonal to $\boldsymbol{\gamma}_1^T\mathbf{x}, \ldots, \boldsymbol{\gamma}_{i-1}^T\mathbf{x}$. But, there is no real evidence to suggest that components with small variances are unimportant in the regression model [26]. Since PCR only considers the structure of predictors, it is relatively easy to understand its statistical properties. PLS considers the relationship between the predictors and the response when constructing the latent variables. Since PLS components not only depend on the predictor variables but also on the response variable, the investigation of statistical properties for PLS is a particularly challenging task. Frank and Friedman showed that PLS tried to find directions with both high variance and high correlation with the response variable [27]. Helland discussed the PLS procedure under a comparatively general statistical model and indicated some connections between PCR and PLS [28,29]. Stone and Brooks constructed a unified objective function and introduced continuum regression (CR) which contains OLS and PCR as the two opposite ends of a continuum spectrum with PLS lying in between [17,18]. Rosen considered PLS from the perspective of the Gauss–Markov model and served invariant subspaces to indicate some relationships between PCR and PLS [34]. Li and Liang et al. introduced elastic component regression which was a linear combination of two criteria of PCR and PLS and showed a natural progression from PCR to PLS [35]. There are some papers that have already compared the predictive ability of these two methods. De Jong proved PLS fitted better than PCR with the same number of components, but this result did not appear to influence predictive ability [36]. Wentzell and Montoto used simulation studies of complex chemical mixtures which contained a large number of components to

[*] Corresponding author.
E-mail address: qsxu@csu.edu.cn (Q.-S. Xu).

http://dx.doi.org/10.1016/j.chemolab.2015.11.003
0169-7439/

show that PLS almost always needed fewer latent variables than PCR, but this result did not necessarily turn into superior prediction performance [37]. He and Zhou proposed a called weight-framework to show the relationship of PLS and PCR [38]. However, the relationship of PLS and PCR is not entirely clear from the literature mentioned above. Generally, there exists an overwhelming popularity of PLS over PCR among chemists; however, no obvious advantage of PLS over PCR in terms of predictive ability is observed in multivariate calibration.

Dimension reduction as a pre-process is a prominent issue today. Sufficient dimension reduction (SDR) introduced by Cook, is important in both theory and practice [21]. SDR performs dimension reduction with no loss of information and less stringent pre-specifying model structures. What's more, it is not affected by the potential structure in the dataset. These characteristics of SDR make it widely used in many kinds of high-dimensional data analysis, such as computer vision [40], biological science [41–44], and drug discovery data [46]. Although the idea of SDR is widely used in various disciplines, this idea is rarely used in chemometrics. Li first explored the link between PLS and SDR and showed the relationship between OLS and PLS in the SDR context [22]. Cook et al. used an envelope, which was a novel context for efficient estimation in multivariate analysis, to build a connection between PLS and SDR. This relation framed PLS algorithm as a Fisherian parameterization. In addition, this advance connects two different statistical methodologies, which allow for deeper comprehending of PLS algorithm and its properties [23]. Some relations of PLS and SDR have been established. In contrast, the connection between PCR and SDR is studied very little. In this article, we introduce the idea of SDR to chemometrics. Meanwhile, we establish the relationship between PCR and SDR. In particular, we show that PLS is equivalent to PCR in the SDR framework, in other words, there is no theoretical advantage of PLS over PCR in terms of prediction performance.

This paper is organized as follows. Section 2 establishes notational conventions. Section 3 briefly outlines the models of PLS and PCR, introduces the idea of SDR and discusses some relationships of PLS, PCR and SDR. Section 4 then displays the results of a simulated dataset and three real datasets to illustrate this equivalence. Finally, Section 5 summarizes the conclusions of the paper.

2. Notation

The following notations are needed in our exposition. Boldface uppercase letters (**A**, **B**) denote matrices, boldface lowercase letters (**a**, **b**) denote column vectors and lowercase italic letters (a, b) denote scalars. Let $\mathbf{x} = (x_1, x_2, \cdots, x_p)^T$ denotes a p-dimension random vector with mean $E(\mathbf{x}) = \mathbf{0}$ and $\sum_{\mathbf{x}} = \text{var}(\mathbf{x}) > \mathbf{0}$. y denotes a respond random variable with mean $E(y) = 0$ and $\boldsymbol{\sigma}_{\mathbf{x}y} = \text{cov}(\mathbf{x}, y)$. $\mathbf{X} \in R^{n \times \text{p}}$ denotes sample matrix; $\mathbf{y} \in R^{n \times 1}$ denotes sample column vector. Let $\hat{\sum}_{\mathbf{x}} = \mathbf{X}^T\mathbf{X}/(n\text{-}1)$ and $\hat{\boldsymbol{\sigma}}_{\mathbf{x}y} = \mathbf{X}^T\mathbf{y}/(n\text{-}1)$. $span(\mathbf{A})$ denotes space spanned by the columns of the $p \times q$ matrix **A**. $\langle \mathbf{a}, \mathbf{b} \rangle_{\sum} = \mathbf{a}^T \sum \mathbf{b}$ denotes the $\sum$ inner product in R^p, where $\sum$ is a symmetric, positive definite matrix in $R^{p \times p}$. When $\sum = \mathbf{I}_p$, the $p \times p$ identity matrix, this inner product changes the usual inner product. $\|\ \|_2$ denotes the Euclidean norm.

3. Theory

In this paper, we will consider the general population statistical model

$$y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of parameters and ε is an error random variable with mean zero ($\text{E}(\varepsilon) = 0$) and variance ($\text{var}(\varepsilon) = \sigma^2$).

3.1. Theory of PLS and PCR

Next we will introduce the population version of the PLS algorithm which is based on Helland's work [28,31]. For further details and references we refer to Helland's papers.

(i) Set starting residual values for **x** and y:

$$\mathbf{e}_0 = \mathbf{x},$$

$$f_0 = y.$$

For $k = 1, 2, \cdots$, do the following steps (ii)–(iv):

(ii) Scores t_k denote linear combinations of the **x** residuals from the previous procedure, $\mathbf{w}_k$ denote weights:

$$t_k = \boldsymbol{e}_{k-1}^T\mathbf{w}_k,$$

$$\mathbf{w}_k = \text{cov}(\mathbf{e}_{k-1}, f_{k-1}).$$

(iii) Conclude **x** loadings $\mathbf{p}_k$ and y loadings q_k:

$$\mathbf{p}_k = \text{cov}(\mathbf{e}_{k\text{-}1}, t_k)/\text{var}(t_k)$$

$$q_k = \text{cov}(f_{k\text{-}1}, t_k)/\text{var}(t_k)$$

(iv) Calculate new residuals:

$$\mathbf{e}_k = \mathbf{e}_{k\text{-}1} - \mathbf{p}_k t_k,$$

$$f_k = f_{k\text{-}1} - q_k t_k.$$

From (i) and (iv), bilinear representation obtained at each step k:

$$\mathbf{x} = \mathbf{p}_1 t_1 + \cdots + \mathbf{p}_k t_k + \mathbf{e}_k, \quad y = q_1 t_1 + \cdots + q_k t_k + f_k \tag{2}$$

Given a new sample **x**, scores t_k can be constructed from the new **x** value and the response y is predicted by

$$\hat{y}_{k,PLS} = q_1 t_1 + \cdots + q_k t_k \tag{3}$$

Eq. (3) can be written in the next form [28]:

$$\hat{y}_{k,PLS} = \mathbf{x}^T\boldsymbol{\beta}_{k,PLS} \tag{4}$$

where $\boldsymbol{\beta}_{k,PLS}$ is the regression vector of PLS in a population version. Helland has proved that two formulae of $\boldsymbol{\beta}_{k,PLS}$ are equivalent:

$$\boldsymbol{\beta}_{k,PLS} = \mathbf{W}_k\left(\mathbf{W}_k^T \sum\nolimits_{\mathbf{x}} \mathbf{W}_k\right)^{-1} \mathbf{W}_k^T \boldsymbol{\sigma}_{\mathbf{x}y}, \quad \text{where } \mathbf{W}_k = (\mathbf{w}_1, \cdots, \mathbf{w}_k) \tag{5}$$

$$\boldsymbol{\beta}_{k,PLS} = \mathbf{S}_k\left(\mathbf{S}_k^T \sum\nolimits_{\mathbf{x}} \mathbf{S}_k\right)^{-1} \mathbf{S}_k^T \boldsymbol{\sigma}_{\mathbf{x}y}, \quad \text{where } \mathbf{S}_k = \left(\boldsymbol{\sigma}_{\mathbf{x}y}, \sum\nolimits_{\mathbf{x}} \boldsymbol{\sigma}_{\mathbf{x}y}, \cdots, \sum\nolimits_{\mathbf{x}}^{k-1} \boldsymbol{\sigma}_{\mathbf{x}y}\right) \tag{6}$$

This result was also discussed in Cook et al. [23].

For our purpose, Eq. (6) is used for constructing a predictor of y in this paper.

By comparing with the sample version of the PLS algorithm, which was described in the references mentioned above, it is easy to see exactly the same as between the sample version and the population version, only that sample (co-)variances are replaced with population (co-)variances.

PCR is one of the most commonly statistical procedures with a wide range of applications. Due to the fact that the algorithm and theory of PCR are familiar to the chemometricians and there are lots of relevant