Chemometrics and Intelligent Laboratory Systems 145 (2015) 60–71

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Variance constrained partial least squares

Xiubao Jiang [a], Xinge You [a,*], Shujian Yu [b], Dacheng Tao [c], C.L. Philip Chen [d], Yiu-ming Cheung [e]

[a] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, Hubei, China
[b] Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA
[c] The Center for Quantum Computation & Intelligent Systems and the Faculty of Engineering and Information Technology in the University of Technology, Sydney, Australia
[d] The Faculty of Science and Technology, University of Macau, Macau, China
[e] Department of Computer Science, Hong Kong Baptist University, Hong Kong, China

ARTICLE INFO

Article history:
Received 16 December 2014
Received in revised form 20 April 2015
Accepted 21 April 2015
Available online 28 April 2015

Keywords:
Partial least squares
Near-infrared spectroscopy
Latent variable
Chemometrics

ABSTRACT

Partial least squares (PLS) regression has achieved desirable performance for modeling the relationship between a set of dependent (response) variables with another set of independent (predictor) variables, especially when the sample size is small relative to the dimension of these variables. In each iteration, PLS finds two latent variables from a set of dependent and independent variables via maximizing the product of three factors: variances of the two latent variables as well as the square of the correlation between these two latent variables. In this paper, we derived the mathematical formulation of the relationship between mean square error (MSE) and these three factors. We find that MSE is not monotonous with the product of the three factors. However, the corresponding optimization problem is difficult to solve if we extract the optimal latent variables directly based on this relationship. To address these problems, a novel multilinear regression model-variance constrained partial least squares (VCPLS) is proposed. In the proposed VCPLS, we find the latent variables via maximizing the product of the variance of latent variable from dependent variables and the square of the correlation between the two latent variables, while constraining the variance of the latent variable from independent variables must be larger than a predetermined threshold. The corresponding optimization problem can be solved computational efficiently, and the latent variables extracted by VCPLS are near-optimal. Compared with classical PLS and it is variants, VCPLS can achieve lower prediction error in the sense of MSE. The experiments are conducted on three near-infrared spectroscopy (NIR) data sets. To demonstrate the applicability of our proposed VCPLS, we also conducted experiments on another data set, which has different characteristics from NIR data. Experimental results verified the superiority of our proposed VCPLS.

1. Introduction

Numerous practical applications in chemometrics can be casted into regression problems, such as near-infrared spectroscopy based protein determination for cereal breeding and quality determination in cereal industry [1–4]. On the one hand, modern technologies often produce data sets with high feature dimensions, and some features may be highly correlated. On the other hand, data collection can be challenging and expensive when it requires manual operations or necessary instruments and materials. As a result, it is very likely that the dimension of sample features is much larger relative to the number of samples [5]. Meanwhile, the features are usually corrupted with noises in the process of measurement [6,7]. Therefore, it is essential to design reliable regression models for these data sets.

The most classical method for regression is ordinary least squares (OLS), which provides estimators via minimizing the residual squared error. Although OLS estimators have low bias, they also suffer from high variance at the same time, especially when the sample size is small. As a result, the prediction error is often larger than expected.

Principal components regression (PCR) [8], ridge regression (RR) [9], subset selection (SS) [10–12], and partial least squares (PLS) [13–18] have been proposed to improve prediction accuracy of OLS via imposing different constraints on original OLS model. RR adds the penalty term of ℓ_2-norm to the cost function, while SS is often achieved by adding a ℓ_1 norm of estimators. The ℓ_1 norm enforces many elements in the estimator to be zero. Consequently, small number of predictor variables were selected to predict the response variables. Although SS has been shown to be more effective than RR under certain conditions [10], it remains unsatisfactory for real-time applications. Different from RR and SS, PLS finds a small number of latent variables (also called scores, components) in the predictor variable space to predict the response variables. Each latent variable is the linear combination of the predictor variables. In addition, the corresponding optimization problem can be solved efficiently [14].

PLS was initially proposed in economics and chemometrics as an alternative approach to OLS in ill-conditioned linear regression problems [19]. Then, it has been successfully extended to other scientific

* Corresponding author.
E-mail address: youxg@mail.hust.edu.cn (X. You).

http://dx.doi.org/10.1016/j.chemolab.2015.04.014
0169-7439/© 2015 Elsevier B.V. All rights reserved.

areas, including bioinformatics, economics, computer vision, and medicine, etc. [20,21]. The objective of PLS is to predict a set of response variables from a set of predictor variables by means of latent variables. It is generally achieved via an iterative process: after the extraction of the score vectors (latent variables), predictor matrix and response matrix are deflated by subtracting their rank-one approximations based on score vectors [22]. Different forms of deflation schemes define several variants of PLS: PLS Mode A [19], PLS1,PLS2 [23], PLS-SB [24]. With different criteria to extract the latent vectors, there are also different forms of PLS: canonical ridge analysis (CRA) [25] , Orthonormalized PLS [26]. Apart from these, there are also some variants aiming at extending PLS to deal with two blocks of variables to model relations among a larger number of sets [27–30]. The basic PLS model assumes that the predictor variables and the response variables are linearly related. In some cases this is not true. To overcome this problem, different non-linear versions of PLS were developed: kernel PLS [31], kernel PLS SVC [32], and Reduced Kernel Orthonormalized Partial Least Squares (rKOPLS) [33].

However, it has been theoretically proved that classical PLS is not optimal [34–37]. Considering that the estimators from samples fall outside the parameter space given by the theoretical PLS algorithm, Bayes PLS [38] was proposed to improve the PLS model. Besides, most of the existing PLS and its variants extract latent variables via maximizing the covariance of the latent variables. Different criteria to extract latent variables have been proposed. In continuum regression (CR) [39], the maximization operation is conducted on the product of three factors: variance of the single response variable, square of the correlation between response variable and latent variable from independent variables, and powered variance of the latent variable. In PCovR [40], the maximization operation is conducted on simple weighted average of two factors: the percentage of variance in predictor variables, the percentage of variance in response variables explained by latent variables. Motivated by [41], PPLS [42] maximizes the correlation while constraining the form of loading coefficients by taking powers of the correlations and standard deviations. Generally speaking, the difference between above methods lies on the different methods adopted to combine the three factors: (a) variance of latent variables from predictor variables, (b) variance of latent variables from response variables, and (c) correlation of the two latent variables.

How to combine these three factors to achieve the optimal performance for regression problems? This problem has been discussed in [43] and the H-principle was proposed therein. In this paper, we address this problem based on the assumption that each predictor is corrupted with noises of the same power. We firstly explored the relationship between the regression error MSE and the three related factors. We found that: (1) when the variance of latent variable from predictor variables is larger than a threshold, MSE decreases slowly with the increase of variance; and (2) the MSE varies linearly with the increase of square of correlation. Based on these observations, we proposed a new computationally efficient method, termed VCPLS, to extract latent variables via constraining factor (b) and maximizing product of factor (a) and (c). Our proposed method can produce lower prediction errors, and experiments on several benchmark data sets validate the effectiveness of our method.

The remainder of this paper is organized as follows. In Section 2, we introduce background knowledge, where the basic motivation and formulation of PLS as well as deflation schemes of PLS are briefly described. In Section 3, we introduce our proposed VCPLS model in detail. Then, in Section 4, we briefly introduce the compared methods and our used data sets for experiments. Experiments and results analysis are conducted in Section 5. Finally, Section 6 concludes this paper.

2. Background

For convenience of theoretical analysis, in this section, we firstly briefly introduce the basic motivation and formulation of PLS in view of random variable. Then we describe the deflation schemes of PLS1 (one of the block of data consists of a single variable) and PLS2 (both blocks are multidimensional) [22], which is appropriate for predicting one block of variables using another block of variables. PLS1, PLS2 are the most frequently used PLS approaches, and hence adopted in this work.

Notation:

- Non-bold: scalar, e.g., i, j, p, q
- Bold: vector or matrix, e.g., $\mathbf{E}$, $\mathbf{F}$, $\mathbf{w}$, $\mathbf{c}$, e, f
- Uppercase: random variable, e.g., E_i, F_i
- Lowercase: specific value, e.g., $\mathbf{e}$, $\mathbf{f}$, $\boldsymbol{w}$, $\boldsymbol{c}$

2.1. Basic motivation and formulation of PLS

For multiple variable linear regression, we have to model the relation between two blocks of random variables, $\mathbf{E} = (E_1, E_2, \cdots, E_p)$ and $\mathbf{F} = (F_1, F_2, \cdots, F_q)$, where p is the number variables in block $\mathbf{E}$, q is the number of variables in block $\mathbf{F}$, $(\bullet, \bullet, \cdots, \bullet)$ represents a row vector, and "•" denotes a scalar or random variable. Meanwhile, we also assume, throughout the paper, all of the variables in $\mathbf{E}$ and $\mathbf{F}$ are zero mean.

OLS approximates $\mathbf{F}$ with $\mathbf{E}\left(\mathbb{E}\left\{\mathbf{E}^T\mathbf{E}\right\}\right)^{-1}\mathbb{E}\left\{\mathbf{E}^T\mathbf{E}\right\}$, where $\mathbb{E}$ denotes expectation operator, $\mathbf{E}^T$ denotes the transpose of $\mathbf{E}$. However, when p is large, the random variables $E_1, E_2, \cdots, E_p$ tend to be linear dependent, that is, there exists scalars $a_1, a_2, \cdots, a_n$ (not all zero) such that $a_1E_1 + a_2E_2 + \cdots + a_pE_p = 0$. In this situation, $\mathbb{E}\left\{\mathbf{E}^T\mathbf{E}\right\}$ is singular or near singular. As a result, the estimators will be failed to compute ($\mathbb{E}\left\{\mathbf{E}^T\mathbf{E}\right\}$) is singular) or with a large variance ($\mathbb{E}\left\{\mathbf{E}^T\mathbf{E}\right\}$) is near singular) which will lead to large regression error. To address this problem, PLS firstly finds latent variable T from $\mathbf{E}$ and correspondingly, U from $\mathbf{F}$ requiring that

Cond.1 T and U should be as capable as possible to explain $\mathbf{E}$ and $\mathbf{F}$, respectively;

Cond.2 The correlation of T and U should be as large as possible.

Assume $T = \mathbf{E}\boldsymbol{w}$, $U = \mathbf{F}\boldsymbol{w}$ corresponding to two latent variables, where $\boldsymbol{w}$ is a $p \times 1$ column vector with unit length, $\boldsymbol{c}$ is a $q \times 1$ column vector with unit length. Then PLS formulates these two conditions as follows

$$\underset{\boldsymbol{w},\boldsymbol{c}}{max} Cov(\mathbf{E}\boldsymbol{w}, \mathbf{F}\boldsymbol{c}) \tag{1}$$

$$s.t. \quad \begin{cases} \boldsymbol{w}^T\boldsymbol{w} = 1 \\ \boldsymbol{c}^T\boldsymbol{c} = 1 \end{cases} \tag{2}$$

where $Cov(\bullet, \bullet)$ means the covariance of two random variable.

Note that

$$Cov(T, U) = \sqrt{Var(T)}\sqrt{Var(U)}Corr(T, U) \tag{3}$$

where $Var(\bullet)$ means the variance of a random variable, and $Corr(\bullet, \bullet)$ means the correlation of two variables. It is well known that the variance of T and U can measure the ability of explaining $\mathbf{E}$ and $\mathbf{F}$ respectively [44]. Thus, from Eq. (3), we know that Eqs. (1)–(2) indeed aims at meeting the requirements described Cond.1, Cond.2.

2.2. PLS1,PLS2 deflation

After T and U were extracted, PLS then deflates $\mathbf{E}$ and $\mathbf{F}$. As two most frequently used deflation schemes of PLS, PLS1,PLS2 are more appropriate for regression [22]. The deflation scheme of PLS1 and PLS2 can be formulated as:

$$\mathbf{E} \leftarrow \mathbf{E} - T\mathbb{E}(T\mathbf{E})/\mathbb{E}T^2 \tag{4}$$

$$\mathbf{F} \leftarrow \mathbf{F} - T\mathbb{E}(T\mathbf{F})/\mathbb{E}T^2 \tag{5}$$

With the updated **E** and **F** (residual matrices), the process of extracting latent variables and deflation can be done iteratively in a number of times until the residual is small enough. This deflation scheme can guarantee the mutual orthogonality of the latent variables [23].

3. Analysis of prediction error with one predictor variable and the proposed model

Classical PLS and its variants extract the latent variables via maximizing the product of standard deviations and correlations between the latent variables, i.e., $\sqrt{Var(T)}\sqrt{Var(U)}Corr(T,U)$. However, is this the optimal criterion for latent variable extraction for regression? In this section, we firstly analyze the prediction error in linear regression when using one predictor variable corrupted by additive noise to predict one response variable. We show that the latent variable extracted by the classical PLS is not optimal. Then, based on these analysis, we propose a novel PLS model-variance constrained partial least square (VCPLS), which is computationally efficient and near-optimal in the sense of MSE.

3.1. Analysis of prediction error with one noisy predictor variable and One response variable

When the predictor is noisy-free, the approximation error in linear regression analysis has been extensively explored [23]. When the predictor variables are measured with random error, various methods have been proposed to estimate the regression coefficients [45–47]. In this section, we will focus on analyzing how the factors contribute to the MSE based on the assumption that only one predictor corrupted with noise and one response variable are available. In addition, we also assume the noise is independent of predictor and response variables.

Let $Z = T + \xi$ denote the noisy predictor, where T is the noise-free predictor (latent variable), which represents a random variable with zero mean and variance σ_T^2. ξ is the noise term, which represents another random variable independent of T with zero mean and variance σ_n^2. In addition, let Y denote the response variable that we want to predict, which represents a random variable with zero mean and variance σ_Y^2. Then the linear regression model for predicting Y with Z can be formulated as:

$$\tilde{Y} = aZ = a(T + \xi) \tag{6}$$

where a is the linear regression coefficient. In the sense of MSE, the optimal coefficient a^* is

$$a^* = \underset{a}{\operatorname{argmin}}\, \mathbb{E}\left(Y - \tilde{Y}\right)^2 \tag{7}$$

$$= \underset{a}{\operatorname{argmin}}\, \mathbb{E}(Y - aZ)^2 \tag{8}$$

$$= \frac{\mathbb{E}YZ}{\mathbb{E}Z^2} \tag{9}$$

Then the MSE of prediction for Y is

$$err = \mathbb{E}(Y - a^*Z)^2 \tag{10}$$

$$= \mathbb{E}Y^2 - \frac{(\mathbb{E}YZ)^2}{\mathbb{E}Z^2} \tag{11}$$

$$= \sigma_Y^2 - \frac{(\mathbb{E}YT)^2}{\sigma_T^2} \cdot \frac{\sigma_T^2}{\sigma_T^2 + \sigma_n^2} \tag{12}$$

$$= \sigma_Y^2 \left(1 - \left(\frac{\mathbb{E}TY}{\sigma_T \sigma_y}\right)^2 \cdot \frac{\sigma_T^2}{\sigma_T^2 + \sigma_n^2}\right) \tag{13}$$

$$= \sigma_Y^2 \left(1 - Corr^2(T, Y) \cdot \frac{\sigma_T^2}{\sigma_T^2 + \sigma_n^2}\right) \tag{14}$$

When $\sigma_n = 0$, Eq. (14) degrades to prediction error in noise-free case, which can be represented as

$$err = \sigma_Y^2 (1 - Corr^2(T, Y)) \tag{15}$$

Since our goal is to extract the optimal latent variable T from a set of candidate variables **E**, that is $T = \mathrm{E}\boldsymbol{w}$, $\boldsymbol{w}^T\boldsymbol{w} = 1$, the variance of response variable σ_Y^2 and variance of noise σ_n^2 is assumed to be fixed. According to Eqs. (14)–(15), the prediction error in noise-free case is determined by $Corr(T, Y)$, and is independent of variance of predictor σ_T^2. However, the prediction error under noisy environment dependents on both $Corr(T, Y)$ and σ_T^2. Obviously, err is not monotonous with the product of these two factors $Corr(T, Y)\sigma_x$ (equivalently $Corr^2(T, Y)\sigma_T^2$), which is adopted by classical PLS to extract the latent variable. As a result, classical PLS cannot find the optimal latent variable in the sense of MSE.

Fig. 1(a) shows the function of prediction error err and σ_T^2, $Corr^2(T, Y)$ (both σ_Y^2 and σ_n^2 are set to 1). Fig. 1(b) shows the contour of the function. To extract the optimal latent variable, we must find T that minimize (Eq. (14)). However, the corresponding optimization problem is difficult to solve. From Fig. 1, we can find that when σ_T^2 is large enough, the prediction error err changes little with different σ_T^2. Based on this observation, we propose a simple model to approximate the optimization problem in Section 3.2.

3.2. The proposed variance constrained partial least squares (VCPLS)

From Eq. (14), we find that when σ_T^2 is fixed, the prediction error err is linear with square of the correlation $Corr^2(T, Y)$. Suppose the additive noise ξ_i with each predictor E_i in **E** are independent and with the same variance σ_n^2 and zero mean. Then for arbitrary $\boldsymbol{w}$ with $\mathbf{w}^T\mathbf{w} = 1$, the noise with latent variable $T = \mathrm{E}\mathbf{w}$ is also zero mean and with variance σ_n^2. Therefore, to find the optimal latent variable $T^* = \mathrm{E}\mathbf{w}^*$ to predict multiple response variables **F**, according to Eq. (14), $\boldsymbol{w}^*$ should be:

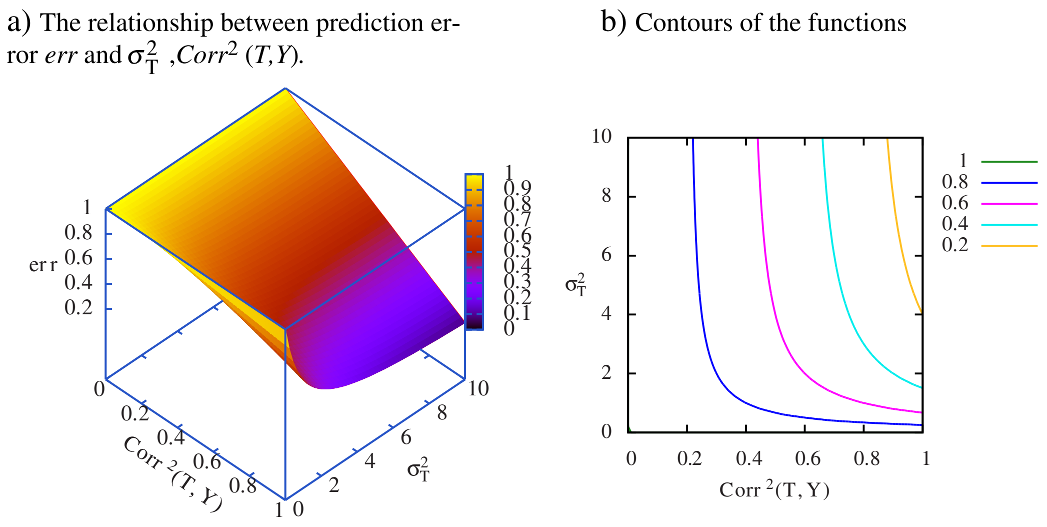
$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_{j=1}^{q} \sigma_{F_j}^2 \left(1 - Corr^2(T, F_j) \cdot \frac{\sigma_T^2}{\sigma_T^2 + \sigma_n^2}\right) \tag{16}$$

$$s.t. \quad \boldsymbol{w}^T\boldsymbol{w} = 1 \tag{17}$$

It is equivalent to

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmax}} \sum_{j=1}^{q} \sigma_{F_j}^2 \cdot Corr^2(T, F_j) \cdot \frac{\sigma_T^2}{\sigma_T^2 + \sigma_n^2} \tag{18}$$

$$= \underset{\boldsymbol{w}}{\operatorname{argmax}} \sum_{j=1}^{q} Cov^2(T, F_j)/(\sigma_T^2 + \sigma_n^2) \tag{19}$$



a) The relationship between prediction error *err* and σ_T^2 ,$Corr^2$ (T,Y).

b) Contours of the functions

Fig. 1. Function of prediction error and square of correlation $Corr^2(T,Y)$, variance σ_T^2 (both σ_Y^2 and σ_n^2 are set to 1).

$$s.t.\quad \boldsymbol{w}^T\boldsymbol{w} = 1 \tag{20}$$

However, the optimization problem (Eqs. (19) and (20)) is difficult to solve. Thus, we propose an approximation to Eqs. (19)–(20) with the following functions

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmax}} \sum_{j=1}^{q} Cov^2(T, F_j)/\sigma_T^2 \tag{21}$$

$$s.t.\quad \begin{cases} \sigma_T^2 = Var(\mathbf{E}\boldsymbol{w}) \geq \alpha \\ \boldsymbol{w}^T\boldsymbol{w} = 1 \end{cases} \tag{22}$$

where $Var(\mathbf{E}\boldsymbol{w})$ is the variance of $\mathbf{E}\boldsymbol{w}$, α is a positive constant which needs to be tuned so that the optimization problem (Eqs. (21) and (22)) can approximate problem Eqs. (19) and (20) well.

In practical applications, the distributions of $(\mathbf{E}, \mathbf{F})$ is unknown, only n observations of $(\mathbf{E}, \mathbf{F})$ are available, $(e_{i1}, e_{i2}, \cdots, e_{ip}, f_{i1}, f_{i2}, \cdots, f_{iq}), i = 1, \cdots, n$, where n is the number of observations. We denote by $\mathrm{e} = (\mathrm{e}_1, \cdots, \mathrm{e}_p)$, $\mathrm{f} = (\mathrm{f}_1, \cdots, \mathrm{f}_q)$, where $\mathrm{e}_\ell = (e_{1\ell}, \cdots, e_{n\ell})^T$, $\ell = 1, \cdots, p$, $\mathrm{f}_\ell = (f_{1\ell}, \cdots, f_{n\ell})^T$, $\ell = 1, \cdots, q$.

Then according to Eqs. (21) and (22) the proposed VCPLS can be formulated as

$$\underset{\boldsymbol{w}}{\operatorname{argmax}} \sum_{j=1}^{q} Cov^2\left(\mathrm{e}\boldsymbol{w}, f_j\right)/Var(\mathrm{e}\boldsymbol{w}) \tag{23}$$

$$s.t.\quad \begin{cases} Var(\mathrm{e}\boldsymbol{w}) \geq \alpha \\ \boldsymbol{w}^T\boldsymbol{w} = 1 \end{cases} \tag{24}$$

where $Var(\mathbf{e}\boldsymbol{w})$ is the sample variance of $\mathbf{e}\boldsymbol{w}$, $Cov(\mathbf{e}\boldsymbol{w}, f_j)$ is the sample covariance of f_j and $\mathbf{e}\boldsymbol{w}$, i.e., $Var(\mathbf{e}\boldsymbol{w}) = \boldsymbol{w}^T\mathbf{e}^T\mathbf{e}\boldsymbol{w}/n$, $Cov^2(\mathbf{e}\boldsymbol{w}, f_j) = (\boldsymbol{w}^T\mathbf{e}^T f_j)(f_j^T\mathbf{e}\boldsymbol{w}) = \boldsymbol{w}^T\mathbf{e}^T f_j f_j^T\mathbf{e}\boldsymbol{w}/n^2$.

The solution to Eqs. (23) and (24) is a scaling of the solution to the following problem (Eqs. (25) and (26))

$$\underset{\boldsymbol{w}}{\operatorname{argmax}} \sum_{j=1}^{q} Cov^2\left(\mathrm{e}\boldsymbol{w}, f_j\right) \tag{25}$$

$$s.t.\quad \begin{cases} Var(\mathrm{e}\boldsymbol{w}) = 1 \\ \boldsymbol{w}^T\boldsymbol{w} \leq 1/\alpha \end{cases} \tag{26}$$

For details please refer to Appendix A.

As $\sum_{j=1}^{q} Cov^2(\mathbf{e}\boldsymbol{w}, f_j) = \sum_{j=1}^{q} (\boldsymbol{w}^T\mathbf{e}^T f_j f_j^T\mathbf{e}\boldsymbol{w})/n^2 = \boldsymbol{w}^T\mathbf{e}^T(\sum_{j=1}^{q}(f_j f_j^T))\mathbf{e}\boldsymbol{w}/n^2 = \boldsymbol{w}^T\mathrm{e}^T\mathrm{f}\mathrm{f}^T\mathbf{e}\boldsymbol{w}/n^2$, then Eqs. (25) and (26) can be rewritten as

$$\max_{\boldsymbol{w}} \boldsymbol{w}^T\mathrm{e}^T\mathrm{f}\mathrm{f}^T\mathrm{e}\boldsymbol{w}/n^2 \tag{27}$$

$$s.t.\quad \begin{cases} \boldsymbol{w}^T\mathrm{e}^T\mathrm{e}\boldsymbol{w}/n = 1 \\ \boldsymbol{w}^T\boldsymbol{w} \leq 1/\alpha \end{cases} \tag{28}$$

which is equivalent to

$$\max_{\boldsymbol{w}} \boldsymbol{w}^T\mathrm{e}^T\mathrm{f}\mathrm{f}^T\mathrm{e}\boldsymbol{w} \tag{29}$$

$$s.t.\quad \begin{cases} \boldsymbol{w}^T\mathrm{e}^T\mathrm{e}\boldsymbol{w} = 1 \\ \boldsymbol{w}^T\boldsymbol{w} \leq 1/(n\alpha) \end{cases} \tag{30}$$

Note that Eqs. (29) and (30) can also be written as the form of two latent variables like the standard form of the classical PLS

$$\max_{\boldsymbol{w},\boldsymbol{c}} \langle \mathrm{e}\boldsymbol{w}, \mathrm{f}\boldsymbol{c} \rangle \tag{31}$$

$$s.t.\quad \begin{cases} \boldsymbol{w}^T\mathrm{e}^T\mathrm{e}\boldsymbol{w} = 1 \\ \boldsymbol{w}^T\boldsymbol{w} \leq 1/(n\alpha) \\ \boldsymbol{c}^T\boldsymbol{c} = 1 \end{cases} \tag{32}$$

For details of the equivalence of Eqs. (29)–(30) and (31)–(32) please refer to Appendix B.

According to [48] we know that for fixed $n\alpha$, there exists $\lambda = 0$ for which the solution to (Eqs. (29) and (30)) is the solution to

$$\max_{\boldsymbol{w}} \boldsymbol{w}^T\mathrm{e}^T\mathrm{f}\mathrm{f}^T\mathrm{e}\boldsymbol{w} - \lambda\boldsymbol{w}^T\boldsymbol{w} \tag{33}$$

$$s.t.\quad \boldsymbol{w}^T\mathrm{e}^T\mathrm{e}\boldsymbol{w} = 1 \tag{34}$$

which is equivalent to

$$\max_{\boldsymbol{w}} \quad \frac{\boldsymbol{w}^T\mathrm{e}^T\mathrm{f}\mathrm{f}^T\mathrm{e}\boldsymbol{w} - \lambda\boldsymbol{w}^T\boldsymbol{w}}{\boldsymbol{w}^T\mathrm{e}\boldsymbol{w}} = \frac{\boldsymbol{w}^T\left(\mathrm{e}^T\mathrm{f}\mathrm{f}^T\mathrm{e} - \lambda\mathrm{I}\right)\boldsymbol{w}}{\boldsymbol{w}^T\mathrm{e}^T\mathrm{e}\boldsymbol{w}} \tag{35}$$

where $\mathbf{I}$ is the identity matrix of size $p \times p$.

Optimization: Denote $\boldsymbol{\Gamma}_1 = \mathrm{e}^T\mathrm{f}\mathrm{f}^T\mathrm{e} - \lambda\mathbf{I}$, $\boldsymbol{\Gamma}_2 = \mathrm{e}^T\mathrm{e}$, $h(\boldsymbol{w}) = \boldsymbol{w}^T\boldsymbol{\Gamma}_1\boldsymbol{w}$, $g(\boldsymbol{w}) = \boldsymbol{w}^T\boldsymbol{\Gamma}_2\boldsymbol{w}$, then the optimization problem (Eq. (35)) can be written as

$$\max_{\boldsymbol{w}} s(\boldsymbol{w}) = \frac{h(\boldsymbol{w})}{g(\boldsymbol{w})}. \tag{36}$$

Taking derivation of $s(\boldsymbol{w})$ of $\boldsymbol{w}$ and let it to be zero, we have

$$\frac{ds(\boldsymbol{w})}{d\boldsymbol{w}} = \frac{h'(\boldsymbol{w})g(\boldsymbol{w}) - g'(\boldsymbol{w})h(\boldsymbol{w})}{g^2(\boldsymbol{w})} \tag{37}$$

$$\propto (\boldsymbol{w}^T\boldsymbol{\Gamma}_2\boldsymbol{w})\boldsymbol{\Gamma}_1\boldsymbol{w} - (\boldsymbol{w}^T\boldsymbol{\Gamma}_1\boldsymbol{w})\boldsymbol{\Gamma}_2\boldsymbol{w} \tag{38}$$

$$= 0 \tag{39}$$

As a result, we have

$$\boldsymbol{w} \quad \propto \quad \boldsymbol{\Gamma}_2^{-1}\boldsymbol{\Gamma}_1\boldsymbol{w} \tag{40}$$

Note that, from Eq. (34) we have $g(\boldsymbol{w}) = \boldsymbol{w}^T\boldsymbol{\Gamma}_2\boldsymbol{w} \neq 0$. Unfortunately, in many applications, the dimension of the feature is always larger than number of observations. As a result, $\boldsymbol{\Gamma}_2$ is always not reversible. Here we add a small disturbance $\eta\mathbf{I}$ to $\boldsymbol{\Gamma}_2$ to avoid singularity, where η is a very small positive number (we set η to 10^{-7} throughout this paper). Combined with Eq. (40), we know that $\boldsymbol{w}$ is proportional to the eigenvector of $(\boldsymbol{\Gamma}_2 + \eta\mathbf{I})^{-1}\boldsymbol{\Gamma}_1$. Assuming the corresponding eigenvalue is β, this relationship can be represented as:

$$(\boldsymbol{\Gamma}_2 + \eta\mathbf{I})^{-1}\boldsymbol{\Gamma}_1\boldsymbol{w} = \beta\boldsymbol{w} \tag{41}$$

Multiplying $(\boldsymbol{\Gamma}_2 + \eta\mathbf{I})$ on both sides of Eq. (41), we have

$$\boldsymbol{\Gamma}_1\boldsymbol{w} = \beta(\boldsymbol{\Gamma}_2 + \eta\mathbf{I})\boldsymbol{w} \tag{42}$$

which is equivalent to

$$\frac{\boldsymbol{w}^T\Gamma_1\boldsymbol{w}}{\boldsymbol{w}^T(\boldsymbol{\Gamma}_2 + \eta\mathbf{I})\boldsymbol{w}} = \beta \tag{43}$$

Therefore, to maximize the objective function (Eq. (35)), $\boldsymbol{w}$ should be the eigenvector which corresponds to the maximum eigenvalue of $(\boldsymbol{\Gamma}_2 + \eta\mathbf{I})^{-1}\boldsymbol{\Gamma}_1$.

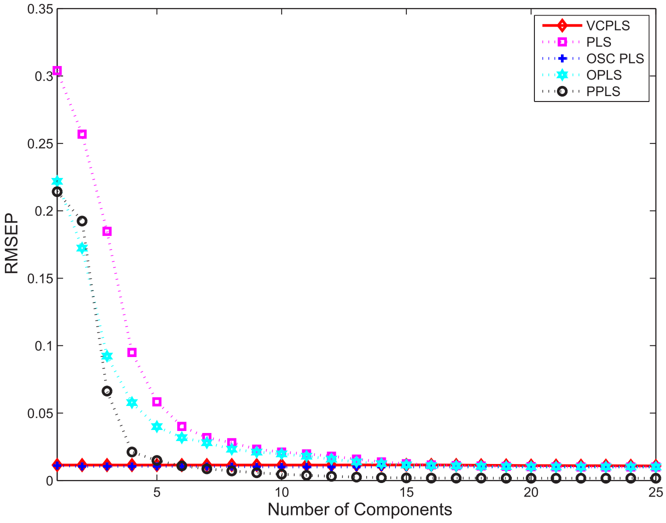


Fig. 2. Corn data set with **m5spec** spectra: RMSEP values of the moisture contents prediction of corn samples from m5spec spectra using VCPLS, PLS, OSC-PLS, OPLS and PPLS models with 1–25 components.

Using the estimated $\boldsymbol{w}$, the latent variable can be calculated as $\mathbf{t} = \mathbf{e}\boldsymbol{w}$. Then following the deflation scheme introduced in Section 2, we can update the data samples e, f. The whole algorithm of VCPLS is described in Algorithm 1.

Algorithm 1. VCPLS – Extracting the components (latent variables)

```
Input:
    Samples of predictor variables, $\mathbf{e}_0 = \mathbf{e}$;
    Samples of response variables, $\mathbf{f}_0 = \mathbf{f}$;
    Parameters, $\eta$, $\lambda$;
Output:
  Latent variables, $\mathbf{t} = \{\mathbf{t}_i, i = 1, \cdots\}$
1: Center $\mathbf{e}_0$, $\mathbf{f}_0$; Initializing $\mathrm{t} = \varnothing$, $i = 1$;
2: Using $\mathbf{e}_{i-1}$, $\mathbf{f}_{i-1}$ to calculate $\boldsymbol{w}_i$ according to Eq. (41);
3: Calculate the component, $\mathbf{t}_i = \mathbf{e}_i\boldsymbol{w}_i$, $\mathbf{t} = \mathbf{t} \cup \{\mathbf{t}_i\}$;
4: Calculate the loadings $\mathbf{p}_i = \mathbf{e}_{i-1}^T\mathbf{t}_i/(\mathbf{t}_i^T\mathbf{t}_i)$, $\mathbf{q}_i = \mathrm{f}_{i-1}^T\mathbf{t}_i/(\mathbf{t}_i^T\mathbf{t}_i)$,
5: Deflation: $\mathbf{e}_i = \mathbf{e}_{i-1} - \mathbf{t}_i\mathbf{p}^T$, $\mathbf{f}_i = \mathbf{f}_{i-1} - \mathbf{t}_i\mathbf{p}^T$,
6: $\mathrm{i} = \mathrm{i} + 1$; returns to step 2 until enough latent variables are
   extracted.
```

Algorithm 1 differs from the classical PLS only on the calculation of $\boldsymbol{w}$, both have to search the eigenvector of a $p \times p$ matrix corresponding to the maximize eigenvalue, but VCPLS requires calculating the inverse of a $p \times p$ matrix additionally.

4. Compared methods and data sets

4.1. Compared methods

As the suboptimal problem of PLS often occurs when there are dominance of irrelevant **E**-variance. The motivation of OSC-PLS [1] is to remove systematic information in **E** not correlated to modeling of **F** before extracting latent variables. Thus, small number of latent variables are desirable to predict response variables. Therefore many approaches have been proposed for removal of irrelevant **E**-information [1,2,49,41]. The motivation of OPLS [49] is the same as OSC-PLS [1], except that the OPLS method is a modification of the original NIPALS PLS algorithm, avoiding cross-validation process to choose the number of components that to be removed. While PPLS [42] adopts another strategy, that is to find the latent variables that maximize the correlation between **E** and **F** while constraining the weights $\boldsymbol{w}$ as the form of $\boldsymbol{w}(\gamma) = (s_1 \cdot |Corr(\mathbf{F}, \mathbf{E}_1)|^{\gamma/(1-\gamma)} \cdot std(\mathbf{E}_1)^{(1-\gamma)/\gamma}, \cdots, s_p \cdot |Corr(\mathbf{F}, \mathbf{E}_p)|^{\gamma/(1-\gamma)} \cdot std(\mathbf{E}_p)^{(1-\gamma)/\gamma})$. By selecting parameter γ, the optimal balance between correlation and variance maybe obtained, hence the extracted latent variable maybe more predictive.

To assess the performance of the proposed model VCPLS, we compare it with four works: the classical PLS [23], OSC-PLS [1], OPLS [49], PPLS [42]. Admittedly, there are some other latent variable extraction methods which also considered balancing correlation and variance that we did not compare in this paper, like, Continuum regression (CR) [39], Principal covariates regression (PCovR) [40]. Continuum regression uses a weighted α geometric average of the correlation and variance criterion to extract latent variables. When α varies from 0 to 1/2, to 1, CR varies from OLS to PLS, PCR. Principal covariates regression (PCovR) [40] uses the criterion of

Table 1
RMSEP of prediction of corn contents from **m5spec** spectra and the corresponding number of components. The parameter λ in VCPLS are set to 10^{-5}, 10^{-9}, 10^{-6}, 10^{-9} respectively.

	PLS	OSC-PLS	OPLS	PPLS	VCPLS
Moisture	0.0100(23)	0.0100(13)	0.0100(22)	**0.0017**(20)	0.0115(1)
Oil	0.0485(25)	**0.0484**(20)	**0.0484**(25)	0.0628(23)	**0.0484**(1)
Protein	0.1114(25)	0.1112(25)	0.1114(24)	**0.1009**(25)	0.1104(1)
Starch	**0.1865**(25)	**0.1867**(18)	**0.1865**(24)	0.2053(23)	**0.1866**(1)

Table 2
Paired *t*-test on the RMSEP for different calibration methods on the corn **m5spec** data set.

	p-value	OSC-PLS	OPLS	PPLS	VCPLS
Moisture					
	PLS	0.1847	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	OSC-PLS		0.1829	$<10^{-4}$	$<10^{-4}$
	OPLS			$<10^{-4}$	$<10^{-4}$
	PPLS				$<10^{-4}$
Oil					
	PLS	0.0043	0.0060	$<10^{-4}$	0.0074
	OSC-PLS		0.1806	$<10^{-4}$	0.1566
	OPLS			$<10^{-4}$	0.3261
	PPLS				$<10^{-4}$
Protein					
	PLS	$<10^{-4}$	0.1581	$<10^{-4}$	$<10^{-4}$
	OSC-PLS		$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	OPLS			$<10^{-4}$	$<10^{-4}$
	PPLS				$<10^{-4}$
Starch					
	PLS	0.0776	0.1169	0.0002	0.1675
	OSC-PLS		0.0775	0.0002	0.1994
	OPLS			0.0002	0.1673
	PPLS				0.0002

weighted arithmetic average of R^2_{ET} and R^2_{FT}, where R^2_{ET} is the percentage that latent variable T accounted for $\mathbf{E}$, R^2_{FT} is the percentage that latent variable explains $\mathbf{F}$. R^2_{ET} is linearly related with variance, R^2_{FT} is linearly related with correlation. When weight α varies from 0 to 1, PCovR varies from OLS to PCR. As the proposed VCPLS cannot be formulated as either weighted geometric average or arithmetic average of variance and correlation, VCPLS is therefore not a specialization of CR or PCovR. However, both CR and PCovR vary between two extreme balance of variance and correlation models: OLS and PCR. Therefore, there exists a specific weight, for which the solution of VCPLS is the same to that of CR (PCovR).

The Matlab code for an implementation of OSC is publicly available from [50]. The Matlab code for PLS and OPLS is publicly available from [51]. The Matlab code for PPLS is also available from [42].

4.2. Data sets

Near-infrared (NIR) spectroscopy has been widely used for the characterisation of solid, semi-solid, fluid and vapour samples [52]. In most of the applications, the objective is to determine the quantity of one or several contents in the samples. Usually, the number of samples is far less than the channels (dimension of NIR data). PLS is successful to complete these tasks. Meanwhile, to demonstrate the applicability of the proposed method, we also conduct experiments on housing data sets which have very different characteristics.

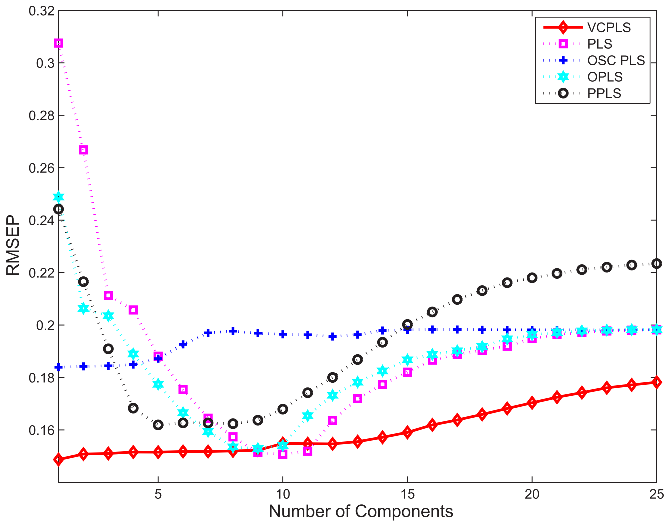


Fig. 3. Corn data set with **mp5spec** spectra: RMSEP values of the moisture contents prediction of corn samples from mp5spec spectra using VCPLS, PLS, OSC-PLS, DOSC-PLS models with 1–25 components.

Table 3
Corn data set with **mp5spec** spectra: RMSEP of prediction of corn contents from mp5spec spectra and the corresponding number of components. The parameter λ in VCPLS are set to 10^{-4}, 10^{-5}, 10^{-4}, 10^{-3} respectively.

	PLS	OSC-PLS	OPLS	PPLS	VCPLS
Moisture	0.1774 (14)	0.1956 (12)	0.1947 (19)	0.1627 (6)	**0.1487** (1)
Oil	**0.1026** (8)	0.1303 (2)	**0.1030** (7)	0.1139 (8)	0.1099 (1)
Protein	0.1578 (10)	0.1798 (1)	0.1579 (9)	0.1597 (9)	**0.1550** (1)
Starch	0.3912 (9)	0.5085 (5)	0.3909 (8)	0.4361 (8)	**0.3882** (1)

4.3. NIR spectra of corn

Three NIR spectra(m5spec, mp5spec, mp6spec) of 80 corn samples were collected from instruments m5, mp5 and mp6, respectively.[1] The wavelength range is 1100 − 2498 nm at 2-nm intervals (700 channels). The moisture, oil, protein and starch values are measured for each sample. The corn data set has been employed in [2,53,54] for regression analysis.

4.4. NIR spectra of wheat kernels

Wheat kernels (415) representing 43 different varieties or variety mixtures from two different locations in Denmark made up the calibration set, while wheat kernels (108) representing 11 different varieties from one location made up the test set.[2] The single kernel transmittance spectra were collected on an Infratec 1255 Food and Feed Analyzer (Tecator AB, Höganäs, Sweden). The wavelength range is 850-1050 nm at 2-nm intervals (100 channels). This data set has been employed in [55,53] for regression analysis.

4.5. NIR spectra of pharmaceutical tablet

The tablet data set[3] consists of a set of near infrared (NIR) transmittance spectra. The aim of the study is to determine the weight percentage of the active substance in the tablets, based on the NIR spectra recorded over the range 4000 − 14000 cm^{-1}, of which the region 7400 − 10500 cm^{-1} (corresponding to 404 predictors) was used for the development of the calibration model. The data set comprised 310 tablet samples. This data set contains 310 tablet samples, and has been employed for regression analysis in [54,56].

4.6. Housing

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University.[4] The aim is to predict housing values in suburbs of Boston from 13 predictors. There are 506 samples. This data set has been employed in [57,58] for regression analysis.

5. Results and discussion

5.1. Experimental setting and parameter determination

According to [2], one component is calculated to remove parts of the spectral data in OSC-PLS and OPLS, and the tolerance parameter is set to 10^{-3}. Before training, the predictor variables and response variables are mean centered. For our proposed VCPLS, two parameters need to be

[1] The data set is publicly available from http://www.eigenvector.com/data/index.htm.
[2] The data set is publicly available from http://www.models.life.ku.dk/wheat_kernels.
[3] The data set is publicly available from http://www.models.kvl.dk/search/node/tablet.
[4] The data set is publicly available from http://archive.ics.uci.edu/ml/datasets/Housing.

Table 4
Paired *t*-test on the RMSEP for different calibration methods on the corn mp5spec data set.

	p-value	OSC-PLS	OPLS	PPLS	VCPLS
Moisture					
	PLS	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	OSC-PLS		0.0055	$<10^{-4}$	$<10^{-4}$
	OPLS			$<10^{-4}$	$<10^{-4}$
	PPLS				$<10^{-4}$
Oil					
	PLS	$<10^{-4}$	0.0007	$<10^{-4}$	$<10^{-4}$
	OSC-PLS		$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	OPLS			$<10^{-4}$	$<10^{-4}$
	PPLS				$<10^{-4}$
Protein					
	PLS	$<10^{-4}$	0.0164	0.0513	$<10^{-4}$
	OSC-PLS		$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	OPLS			0.0660	$<10^{-4}$
	PPLS				0.0001
Starch					
	PLS	$<10^{-4}$	0.0040	$<10^{-4}$	0.0172
	OSC-PLS		$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	OPLS			$<10^{-4}$	0.0306
	PPLS				$<10^{-4}$

tuned: λ and η. The parameter η is used to add a small disturbance to a non-negative definite matrix to make the matrix positive definite, meanwhile the disturbance should small enough. For all the experiments, we have tried η from 10^{-5} to 10^{-10}, the experimental results are almost the same. The results shown in this paper corresponding to $\eta = 10^{-7}$. λ is determined according to 10-fold cross-validation from the candidate values $\{10^i, i = -10, -9, \cdots, 3\}$. For data set of corn, as there are only 80 samples totally, λ is determined from the 1 to 60 samples. For data sets of wheat kernels, λ is determined from the calibration set. For data set of pharmaceutical tablet, λ is determined from the 1 to 100 samples. Then nine-tenths of the selected samples are randomly selected to train the model under each candidate value, the remaining one-tenth are used to calculate RMSE, this process repeated 40 times. Then the candidate value with the minimum mean RMSE is determined as the optimal parameter.

Another parameter which should be determined is the optimal number of components. For the proposed VCPLS, the optimal number of components is set to q (the number of response variables). For the other four compared methods, the same cross-validation strategy and data sets which are used for determining the parameter λ are employed again to determine the optimal number of components.

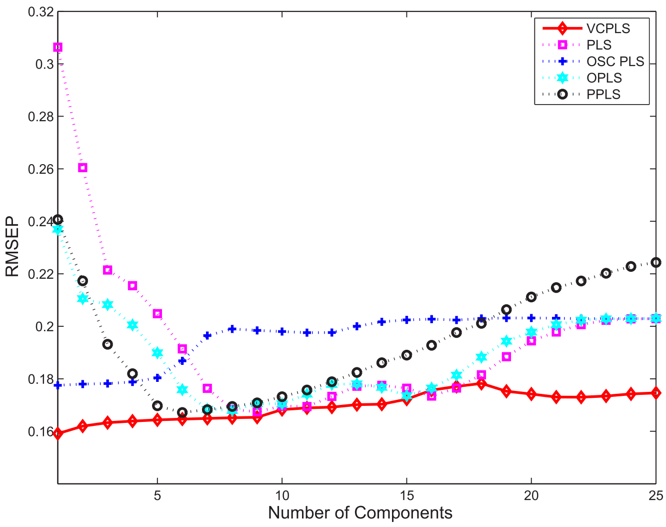


Fig. 4. Corn data set with **mp6spec** spectra: RMSEP values of the moisture contents prediction of corn samples from mp6spec spectra using VCPLS, PLS, OSC-PLS, OPLS and PPLS models with 1–25 components.

Table 5
RMSEP of prediction of corn contents from mp6spec spectra and the corresponding number of components. The parameter λ in VCPLS are set to 10^{-4}, 10^{-5}, 10^{-4}, 10^{-3} respectively.

	PLS	OSC-PLS	OPLS	PPLS	VCPLS
Moisture	0.1764 (15)	0.1976 (12)	0.1767 (14)	0.1695 (8)	**0.1591** (1)
Oil	0.1319 (22)	0.1298 (11)	0.1319 (21)	0.1141 (11)	**0.1042** (1)
Protein	0.1644 (11)	0.1896 (1)	0.1646 (10)	0.1640 (9)	**0.1583** (1)
Starch	**0.3756** (9)	0.4734 (7)	**0.3753** (8)	0.4021 (7)	**0.3758** (1)

Table 6
Paired *t*-test on the RMSEP for different calibration methods on the corn mp6spec data set.

	p-value	OSC-PLS	OPLS	PPLS	VCPLS
Moisture					
	PLS	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	OSC-PLS		$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	OPLS			$<10^{-4}$	$<10^{-4}$
	PPLS				$<10^{-4}$
Oil					
	PLS	$<10^{-4}$	0.1094	$<10^{-4}$	$<10^{-4}$
	OSC-PLS		$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	OPLS			$<10^{-4}$	$<10^{-4}$
	PPLS				$<10^{-4}$
Protein					
	PLS	$<10^{-4}$	$<10^{-4}$	0.3604	$<10^{-4}$
	OSC-PLS		$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	OPLS			0.3151	$<10^{-4}$
	PPLS				$<10^{-4}$
Starch					
	PLS	$<10^{-4}$	0.0033	$<10^{-4}$	0.4291
	OSC-PLS		$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	OPLS			$<10^{-4}$	0.3365
	PPLS				$<10^{-4}$

We compare the performance of different models in terms of the root mean square error for prediction (RMSEP) on the test set:

$$\mathrm{RMSEP} = \sqrt{\frac{1}{n \times q}\sum_{i=1}^{n}\sum_{j=1}^{q}\left(y_{ij} - \hat{y}_{ij}\right)^2} \tag{44}$$

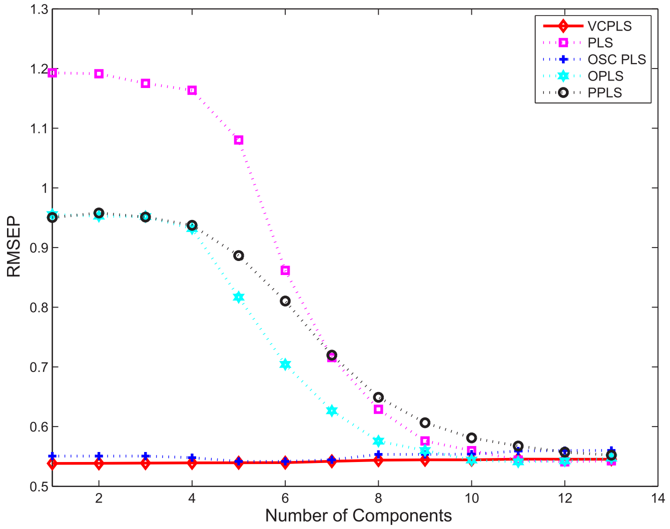


Fig. 5. NIR spectra of wheat kernels data set: RMSEP values of predicting the protein content using VCPLS, PLS, OSC-PLS, OPLS and PPLS models with 1–25 components.

Table 7
RMSEP and the corresponding number of components on the wheat kernels data set.

	PLS	OSC-PLS	OPLS	PPLS	VCPLS
Protein	0.5441 (11)	0.5507 (1)	0.5441 (10)	0.5572 (12)	**0.5383** (1)

where $\hat{y}_{ij}$ is the prediction of the ith observation's jth response variable value and n is the size of the test set.

For a fair performance evaluation of different models, the random partitioning of training and test sets is repeated I times ($I = 300$ in all the experiments), except for data set of wheat kernels, the size of the training set and test sets are equal. For wheat kernels, the samples are already partitioned as calibration set and test set, we put them together and then randomly choose training set with the same size of the original calibration set. Then the average RMSEP is used for comparison.

In addition, according to [56], paired-t test is employed to determine whether the performance of the different methods is statistically significant. Let r_i and s_i ($i = 1, \cdots, I$) be the RMSEP of two methods, then the t statistic is calculated as:

$$t = (\bar{r} - \bar{s})\sqrt{\frac{I(I-1)}{\sum_{i=1}^{I}(\hat{r}_i - \hat{s}_i)^2}} \tag{45}$$

where $\bar{r}$ and $\bar{s}$ are the means of r_i and s_i respectively. When the t statistic is calculated, the corresponding p-value can be obtained. If the p-value is lower than a given threshold, normally taken as 0.05, the performance of the two methods can be claimed as statistically significant.

5.2. Results on NIR spectra of Corn

We train the models for four contents (moisture, oil, protein, starch) separately.

Fig. 2 shows the mean RMSEP values of the prediction of the moisture content of the corn samples from m5spec spectra using VCPLS, PLS, OSC-PLS, OPLS and PPLS models with 1 — 25 components, where λ is set to 10^{-5} according to cross-validation. Table 1 shows the results of RMSEP, the corresponding number of components for each method and the parameter λ in VCPLS model. Table 2 shows the p-value of the paired t-test of the corresponding RMSEP.

Fig. 2 and Table 1 show that PPLS outperforms PLS, OSC-PLS, OPLS and VCPLS on moisture and protein contents prediction in terms of RMSEP. For oil contents prediction, OSC-PLS, OPLS and VCPLS achieve lower RMSEP than PLS and PPLS. The RMSEP of PLS seems very similar to OSC-PLS, OPLS and VCPLS, however, the p-values 0.0043, 0.0060 and 0.0074 in Table 2 demonstrate that the difference in terms of RMSEP is statistically significant. For starch contents prediction, PLS, OSC-PLS, OPLS and VCPLS achieve similar RMSEP and lower than PPLS statistically significant. It should be noted that, VCPLS can achieve the smallest RMSEP with only one component, while the other four methods normally use more components.

Fig. 3 shows the mean RMSEP values of the prediction of the moisture content of the corn samples from mp5spec spectra using VCPLS, PLS, OSC-PLS, OPLS and PPLS models with 1 — 25 components, where λ is set to 10^{-4} according to cross-validation. Table 3 shows the results of RMSEP, the corresponding number of components for each method and the parameter λ in VCPLS model. Table 4 shows the p-value of the paired t-test of the corresponding RMSEP.

Table 8
Paired t-test on the RMSEP for different calibration methods on the wheat kernels data set.

p-value	OSC-PLS	OPLS	PPLS	VCPLS
PLS	$<10^{-4}$	0.1593	$<10^{-4}$	$<10^{-4}$
OSC-PLS		$<10^{-4}$	0.0013	$<10^{-4}$
OPLS			$<10^{-4}$	$<10^{-4}$
PPLS				$<10^{-4}$

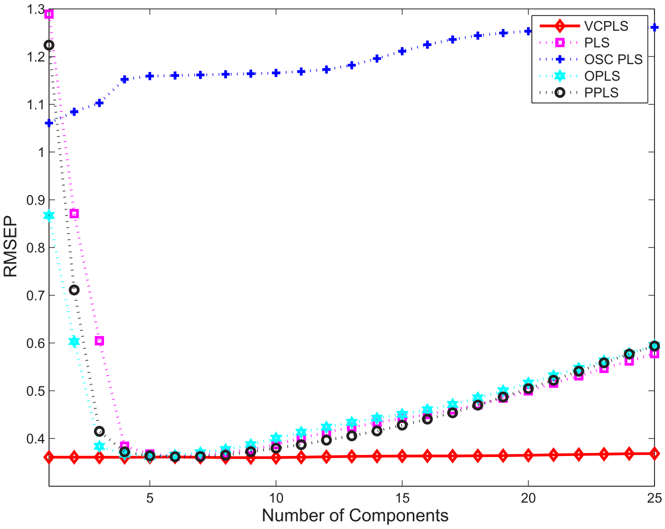


Fig. 6. NIR spectra of pharmaceutical tablet data set: RMSEP values of predicting the active ingredient using VCPLS, PLS, OSC-PLS, OPLS and PPLS models with 1–25 components.

Fig. 3 and Table 3 show that VCPLS outperforms the PLS, OSC-PLS, OPLS, PPLS on moisture, protein and starch contents prediction in terms of both RMSEP and number of components. p-values in Table 4 show that the difference in terms of RMSEP is statistically significant. For oil contents prediction, both PLS and OPLS achieve lower RMSEP compared to VCPLS, but they only use 6 and 7 more components respectively.

Fig. 4 shows the mean RMSEP values of the prediction of the moisture content of the corn samples from mp6spec spectra using VCPLS, PLS, OSC-PLS, OPLS and PPLS models with 1 — 25 components, where λ is set to 10^{-4} according to cross-validation. Table 5 shows the results of RMSEP, the corresponding number of components for each method and the parameter λ in VCPLS model. Table 6 shows the p-value of the paired t-test of the corresponding RMSEP.

Fig. 4 and Table 5 show that VCPLS outperforms the PLS, OSC-PLS, OPLS, PPLS on moisture, oil, and protein contents prediction in terms of both RMSEP and number of components. p-values in Table 6 show that the improvement is statistically significant. For starch contents prediction, PLS, OPLS and VCPLS achieve similar lower RMSEP than OSC-PLS and PPLS. In Table 6, p-values between VCPLS and PLS, OPLS are 0.4291, 0.3365 respectively, which indicate that the difference between VCPLS and PLS, OPLS is not statistically significant. However, PLS and OPLS use 8 and 7 more components respectively.

5.3. Results on NIR spectra of wheat kernels

The aim is to predict the protein content of wheat kernels from the corresponding spectra. The mean RSMEP values with 1 — 25 components are presented in Fig. 5, where λ is set to 10^{-4} according to cross-validation. Table 7 shows the results of RMSEP and the corresponding number of components for each method. Table 8 shows the p-value of the paired t-test of the corresponding RMSEP.

Table 9
RMSEP and the corresponding number of components on the tablet data set.

	PLS	OSC-PLS	OPLS	PPLS	VCPLS
Ingredient	0.3676 (5)	1.0607 (1)	0.3623 (5)	0.4150 (3)	**0.3607** (1)

Table 10
Paired *t*-test on the RMSEP for different calibration methods on the tablet data set.

p-value	OSC-PLS	OPLS	PPLS	VCPLS
PLS	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
OSC-PLS		$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
OPLS			$<10^{-4}$	$<10^{-4}$
PPLS				$<10^{-4}$

According to Fig. 5 and Table 7, the RMSEP values achieved with VCPLS are lower than those achieved with other four models. *p*-values shown in Table 8 between VCPLS and the other four methods (all smaller than 10^{-4}) demonstrate that the improvement is statistically significant. Meanwhile, VCPLS can achieve the RMSEP value with only one component.

5.4. Results on NIR spectra of pharmaceutical tablet

The aim is to predict active ingredient of tablets from the corresponding spectra. The mean RMSEP values with $1-25$ components are presented in Fig. 6, where λ is set to 10 according to cross-validation. Table 9 shows the results of RMSEP and the corresponding number of components for each method. Table 10 shows the *p*-value of the paired *t*-test of the corresponding RMSEP.

Fig. 6 and Table 9 show that the RMSEP values achieved with VCPLS are lower than those achieved with other four methods. *p*-values shown in Table 10 between VCPLS and the other four methods (all smaller than 10^{-4}) demonstrate that the improvement is statistically significant. Meanwhile, VCPLS can achieve the RMSEP value with only one component.

5.5. Results on housing

As there are only 13 predictor variables, the mean RSMEP values with $1-13$ components are presented in Fig. 7, where λ is set to 10^{-2} according to cross-validation. Table 11 shows the results of RMSEP and the corresponding number of components for each method. Table 12 shows the *p*-value of the paired *t*-test of the corresponding RMSEP.

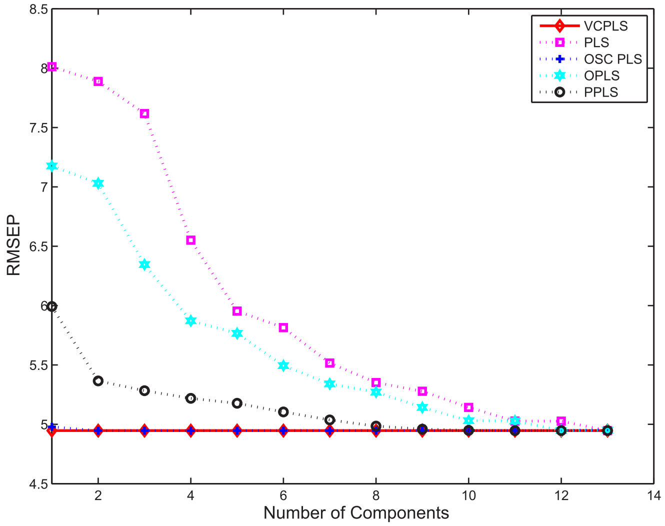


Fig. 7. Housing data set: RMSEP values of predicting housing values using VCPLS, PLS, OSC-PLS, OPLS and PPLS models with 1–13 components.

Table 11
RMSEP and the corresponding number of components on the housing data set.

PLS	OSC-PLS	OPLS	PPLS	VCPLS
4.9470633 (13)	4.9557 (13)	4.9470633 (12)	4.9473 (11)	**4.9470632** (1)

Fig. 7 and Table 11 show that the RMSEP values achieved with PLS, OPLS and VCPLS are lower than OSC-PLS and PPLS, compared with PLS and OPLS, VCPLS can achieve slightly lower RMSEP. However, *p*-values shown in Table 12 between VCPLS and PLS, OPLS are less than 0.05, which indicate that the improvement is statistically significant. *p*-value between VCPLS and PPLS is 0.1488, which indicate that the difference in terms of RMSEP between VCPLS and PPLS is not statistically significant. However, VCPLS can achieve the RMSEP value with only one component, PLS, OPLS, PPLS require far more components.

5.6. Discussion

By referring to (33)–(34), VCPLS seems to be very similar to RR [9]. However, the regularization term in VCPLS is imposed on the weights used to construct components; the regularization term in RR is imposed on the regression coefficients. If only one component is used to predict the response variable, the regression coefficient is linear with the weights. In this viewpoint, regularization with weights (VCPLS) is equivalent to regularization with regression coefficients. However, when there are multiple response variables, the regularization term with weights (hence components) is related to all response variables, there is no linear relationship with weights and regression coefficient. In addition, when more than one components (correspondingly more than one weights) are used to predict the response variables, the regression coefficients has a complicated relationship with these weights [59]. Hence, VCPLS is different with RR.

Observing the experimental results shown above, we can conclude that the proposed VCPLS always needs far less components than the other four models. We will discuss in the following part that this characteristic is reasonable.

According to our analysis in Sections 3.1 and 3.2, we know that the extracted latent variable is near optimal. Then for single response variable case, one latent variable can achieve the best performance. A simple illustration is shown in Fig. 8. Suppose Y is a single response variable, T_1 is the first extracted latent variable which is optimal, p_1 is the corresponding loading coefficient. If the second extracted latent variable T_2 can further improve the regression performance (suppose the corresponding loading coefficient is p_2), then the optimal latent variable should be proportional to $T_1p_1+T_2p_2$, which is contradict to the assumption that T_1 is the optimal latent variable. All the results shown above verified our analysis. From this point, we can get the conclusion that the criterion to extract latent variable employed by PLS, OSC-PLS, OPLS, PPLS is not optimal in the sense of minimizing RMSEP.

For multiple response variables, i.e., $\mathbf{F}=(F_1,\cdots,F_q)$, if we separately extract latent variable for each response variable, then q latent variables $T_1,\cdots,T_q$ will optimally predict $\mathbf{F}$. If we iteratively extract latent variables

Table 12
Paired *t*-test on the RMSEP for different calibration methods on the housing data set.

p-value	OSC-PLS	OPLS	PPLS	VCPLS
PLS	$<10^{-4}$	$<10^{-4}$	0.1489	0.0286
OSC-PLS		$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
OPLS			0.1490	$<10^{-4}$
PPLS				0.1488

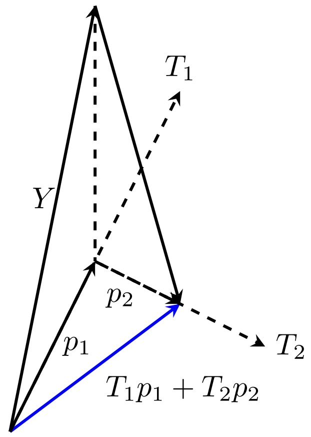


Fig. 8. Illustration of the optimal number of components for single response variable of VCPLS.

($\tilde{T}_1, \tilde{T}_2, \cdots$) for all the response variables simultaneously, then $\tilde{T}_1, \tilde{T}_2, \cdots$ will lie in the space spanned by $T_1, \cdots, T_q$. Note that the PLS2 deflation scheme makes $\tilde{T}_1, \tilde{T}_2, \cdots$ mutually orthogonal, hence no more than q latent variables (components) is enough to predict **F**. If the response variables are not highly correlated, q is the optimal number of components.

Fig. 9 shows the results of predicting four contents of corn on m5spec data. We observed that for VCPLS, 4 components can achieve the minimum RMSEP, which confirms our analysis.

From Figs. 2–4, Tables 1–5, we can find that all the methods on data set m5spec achieve smaller RMSEP and almost never over-fitting. But for data set mp5spec, mp6spec, almost all the methods suffer from over-fitting when more components are employed. The reason is probably that the measurement error of m5 spectrometer is smaller than the other two spectrometers. But for VCPLS, the optimal number of components is known, over-fitting will not occur.

OSC-PLS gives better results compared to PLS for the wheat kernel datasets, corn m5spec datasets and housing dataset. In these data all the methods will not over-fitting even with many components. Its possibly that the measurement error is very small on these data sets. So OSC indeed removes non-relevant dominance components in predictors, which is helpful for regression task.

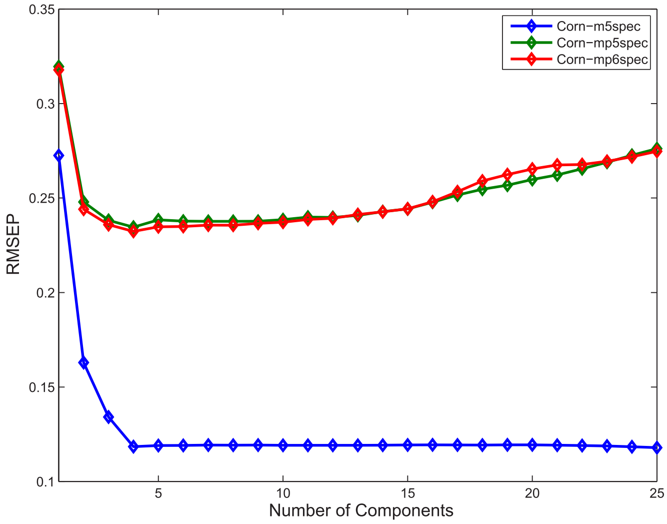


Fig. 9. Illustration of the optimal number of components for multiple response variables of VCPLS. Four contents of corn are predicted on data sets m5spec, mp5spec, mp6spec.

From Fig. 2 to 7, we can observe that the RMSEP of VCPLS seems to be independent of the number of components for some data sets. It mainly caused by two reasons:

- First, for a single response variable, VCPLS always achieves the minimum RMSEP with one component. Hence, if over-fitting does not occur with more components, the RMSEP of VCPLS will not change with the number of components.
- Second, in the deflation process of the first few iteration steps, for VCPLS, components with large correlations are removed from the original predictor variables, while for the other methods, components with large variances are removed from the original predictor variables. As discussed in [40,43], variance of the component guarantees stability, while correlation characterizes the prediction ability. Thus, with more components, RMSEP for VCPLS changes little, while instability makes RMSEP of other methods become large.

To assess the computational complexity, we repeatedly extract the first component to predict the moisture contents from m5spec 1000 times (for each trial, the training set and test set are selected as describe in 5.1). The CPU time is 119.7294 seconds and 54.7322 seconds for VCPLS and PLS respectively. The algorithms were implemented in Matlab 2013a 64 bit and were executed on a Intel(R) Core(TM) i5-2300 2.80GHz computer running under Windows 7. The VCPLS is about two times slower than PLS, however VCPLS always need far less components than PLS. Another disadvantage for the proposed VCPLS is an extra parameter λ has to be tuned before model calibration. In some applications maybe no samples for tuning this parameter are available.

6. Conclusions

In this work, we derived the mathematical formulation of the relationship between MSE, variance of latent variables and square of the correlation between latent variables. We proved that when the feature is corrupted with noise, both large correlation and large variance of features can lead to lower MSE. We further proved that when variance of features is larger than a threshold, correlation plays a more important role to improve the prediction accuracy. Based on these observations, we developed a new computationally efficient multivariate linear regression model-Variance Constrained Partial Least Squares (VCPLS). The latent variables extracted by VCPLS is near-optimal in the sense of MSE. In addition, the proposed method can easily be combined with other improvement strategies for classical PLS, such as kernel framework and sparse framework.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgment

This work is supported partially by the National Technologies R&D Program (Grand no. 2012BAK02B06), National Natural Science Foundation (NSFC) of China (Grant no. 61272203), the Ph.D. Programs Foundation of Ministry of Education of China (Grant no. 20110142110060).

Appendix A

To proof the solution of optimization problem (Eqs. (46) and (47)) is an scaling of the solution to optimization problem (Eqs. (48) and (49))

$$\underset{\boldsymbol{w}}{\operatorname{argmax}}\, g(\boldsymbol{w}) = \sum_{j=1}^{q} Cov^2\left(\mathbf{e}\boldsymbol{w}, f_j\right)/Var(\mathbf{e}\boldsymbol{w}) \tag{46}$$

$$s.t.\quad \begin{cases} Var(\mathrm{e}\boldsymbol{w})\geq\alpha \\ \boldsymbol{w}^T\boldsymbol{w}=1 \end{cases} \tag{47}$$

$$\underset{\tilde{\boldsymbol{w}}}{\operatorname{argmax}}\ \tilde{g}(\tilde{\boldsymbol{w}})=\sum_{j=1}^{q}Cov^2\left(\mathrm{e}\tilde{\boldsymbol{w}},f_j\right) \tag{48}$$

$$s.t.\quad \begin{cases} Var(\mathrm{e}\tilde{\boldsymbol{w}})=1 \\ \tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}}\leq 1/\alpha \end{cases} \tag{49}$$

Property (1). For any $\boldsymbol{w}$ satisfying Eq. (47), let $\tilde{\boldsymbol{w}}=\boldsymbol{w}/\sqrt{Var(\mathrm{e}\boldsymbol{w})}$, then

$$Var(\mathrm{e}\tilde{\boldsymbol{w}})=Var\left(\mathrm{e}\boldsymbol{w}/\sqrt{Var(\mathrm{e}\boldsymbol{w})}\right)=1 \tag{50}$$

$$\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}}=\boldsymbol{w}^T\boldsymbol{w}/Var(\mathrm{e}\boldsymbol{w})\leq 1/\alpha \tag{51}$$

$$\tilde{g}(\tilde{\boldsymbol{w}})=\sum_{j=1}^{q}Cov^2\left(\mathrm{e}\tilde{\boldsymbol{w}},f_j\right) \tag{52}$$

$$=\sum_{j=1}^{q}Cov^2\left(\mathrm{e}\boldsymbol{w},f_j\right)/Var(\mathrm{e}\boldsymbol{w}) \tag{53}$$

$$=g(\boldsymbol{w}) \tag{54}$$

Property (2). For any $\tilde{\boldsymbol{w}}$ satisfying Eq. (49), let $\boldsymbol{w}=\tilde{\boldsymbol{w}}/\sqrt{\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}}}$, then we have

$$Var(\mathrm{e}\boldsymbol{w})=Var(\mathrm{e}\tilde{\boldsymbol{w}})/\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}}\geq\alpha \tag{55}$$

$$\boldsymbol{w}^T\boldsymbol{w}=\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}}/\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}}=1 \tag{56}$$

$$g(\boldsymbol{w})=\sum_{j=1}^{q}Cov^2\left(\mathrm{e}\boldsymbol{w},f_j\right)/Var(\mathrm{e}\boldsymbol{w}) \tag{57}$$

$$=\sum_{j=1}^{q}Cov^2\left(\mathrm{e}\boldsymbol{w}/\sqrt{Var(\mathrm{e}\boldsymbol{w})},f_j\right) \tag{58}$$

$$=\sum_{j=1}^{q}Cov^2\left(\mathrm{e}\tilde{\boldsymbol{w}},f_j\right) \tag{59}$$

$$=\tilde{g}(\tilde{\boldsymbol{w}}) \tag{60}$$

Combining property **Property** (1) and **Property** (2) together, we have

- for any $\boldsymbol{w}_1$, $\boldsymbol{w}_2$ satisfying Eq. (47) and $g(\boldsymbol{w}_1)\geq g(\boldsymbol{w}_2)$, then $\tilde{\boldsymbol{w}}_1=\boldsymbol{w}_1/\sqrt{Var(\mathrm{e}\boldsymbol{w}_1)}$, $\tilde{\boldsymbol{w}}_2=\boldsymbol{w}_2/\sqrt{Var(\mathrm{e}\boldsymbol{w}_2)}$ satisfying Eq. (49) and $\tilde{g}(\tilde{\boldsymbol{w}}_1)=g(\boldsymbol{w}_1)\geq g(\boldsymbol{w}_2)=\tilde{g}(\tilde{\boldsymbol{w}}_2)$.
- for any $\tilde{\boldsymbol{w}}_1,\tilde{\boldsymbol{w}}_2$ satisfying Eq. (49) and $\tilde{g}(\tilde{\boldsymbol{w}}_1)\geq\tilde{g}(\tilde{\boldsymbol{w}}_2)$, then $\boldsymbol{w}_1=\tilde{\boldsymbol{w}}_1/\sqrt{\tilde{\boldsymbol{w}}_1^T\tilde{\boldsymbol{w}}_1}$, $\boldsymbol{w}_2=\tilde{\boldsymbol{w}}_2/\sqrt{\tilde{\boldsymbol{w}}_1^T\tilde{\boldsymbol{w}}_2}$ satisfying Eq. (47) and $g(\boldsymbol{w}_1)=\tilde{g}(\tilde{\boldsymbol{w}}_1)\geq\tilde{g}(\tilde{\boldsymbol{w}}_2)=g(\boldsymbol{w}_2)$,

Thus, if $\boldsymbol{w}_1$ is a solution of Eqs. (46) and (47), then $\tilde{\boldsymbol{w}}_1=\boldsymbol{w}_1/\sqrt{Var(\mathrm{e}\boldsymbol{w}_1)}$ is a solution of Eqs. (48) and (49) (if not, there exists $\tilde{\boldsymbol{w}}_0$, such that $\tilde{g}(\tilde{\boldsymbol{w}}_0)>\tilde{g}(\tilde{\boldsymbol{w}}_1)$, let $\boldsymbol{w}_0=\tilde{\boldsymbol{w}}_0/\sqrt{\tilde{\boldsymbol{w}}_0^T\tilde{\boldsymbol{w}}_0}$, we have $g(\boldsymbol{w}_0)>g(\boldsymbol{w}_1)$, which is contradict to $\boldsymbol{w}_1$ is a solution of Eqs. (46) and (47)). Similarly, if $\tilde{\boldsymbol{w}}_1$ is a solution of Eqs. (48) and (49), then $\boldsymbol{w}_1=\tilde{\boldsymbol{w}}_1/\sqrt{\tilde{\boldsymbol{w}}_1^T\tilde{\boldsymbol{w}}_1}$ is a solution of Eq. (46) and (47).

Appendix B

For any positive constant γ the following two optimization problems (Eqs. (61)–(62) and Eqs. (63)–(64)) are equivalent (may differ in sign).

$$\underset{\boldsymbol{w}}{max}\ \boldsymbol{w}^T\mathrm{e}^T\mathrm{f}\mathrm{f}^T\mathrm{e}\boldsymbol{w} \tag{61}$$

$$s.t.\quad \begin{cases} \boldsymbol{w}^T\mathrm{e}^T\mathrm{e}\boldsymbol{w}=1 \\ \boldsymbol{w}^T\boldsymbol{w}\leq\gamma \end{cases} \tag{62}$$

$$\underset{\boldsymbol{w},\boldsymbol{c}}{max}\langle\mathrm{e}\boldsymbol{w},\ \mathrm{f}\boldsymbol{c}\rangle=\boldsymbol{w}^T\mathrm{e}^T\,\mathrm{f}\boldsymbol{c} \tag{63}$$

$$s.t.\quad \begin{cases} \boldsymbol{w}^T\mathrm{e}^T\mathrm{e}\boldsymbol{w}=1 \\ \boldsymbol{w}^T\boldsymbol{w}\leq\gamma \\ \boldsymbol{c}^T\boldsymbol{c}=1 \end{cases} \tag{64}$$

Proof. The Lagrange function of Eqs. (63) and (64) is

$$\begin{aligned} &L(\boldsymbol{w},\boldsymbol{c},\lambda_1,\lambda_2,\lambda_3) \\ &\quad=\boldsymbol{w}^T\mathrm{e}^T\,\mathrm{f}\boldsymbol{c}-\lambda_1\left(\boldsymbol{w}^T\mathrm{e}^T\mathrm{e}\boldsymbol{w}-1\right)-\lambda_2\left(\boldsymbol{w}^T\boldsymbol{w}-\gamma\right)-\lambda_3\left(\boldsymbol{c}^T\boldsymbol{c}-1\right) \end{aligned} \tag{65}$$

Take derivative of Lagrange function L with respect to $\boldsymbol{c}$ and set it to zero,

$$\frac{\partial L}{\partial\boldsymbol{c}}=\mathrm{f}^T\mathrm{e}\boldsymbol{w}-2\lambda_3\boldsymbol{c}=0 \tag{66}$$

If $\lambda_3=0$, then $\mathrm{f}^T\mathrm{e}\boldsymbol{w}=0$, that is to say $\underset{\boldsymbol{w},\boldsymbol{c}}{max}\ \boldsymbol{w}^T\mathrm{e}^T\,\mathrm{f}\boldsymbol{c}=0$, which means **e** is orthogonal to **f**, then any $\boldsymbol{w}$, $\boldsymbol{c}$ is the solution of Eqs. (61)–(62), (63)–(64) and (61)–(62). If $\lambda_3\neq 0$, then $\boldsymbol{c}=\mathrm{f}^T\mathrm{e}\boldsymbol{w}/||\mathrm{f}^T\mathrm{e}\boldsymbol{w}||$, substitute it to Eqs. (63)–(64), then (63)–(64) can be formulated as

$$\underset{\boldsymbol{w}}{max}\ \boldsymbol{w}^T\mathrm{e}^T\mathrm{f}\mathrm{f}^T\mathrm{e}\boldsymbol{w}/\left\|\mathrm{f}^T\mathrm{e}\boldsymbol{w}\right\|=\left(\boldsymbol{w}^T\mathrm{e}^T\mathrm{f}\mathrm{f}^T\mathrm{e}\boldsymbol{w}\right)^{1/2} \tag{67}$$

$$s.t.\quad \begin{cases} \boldsymbol{w}^T\mathrm{e}^T\mathrm{e}\boldsymbol{w}=1 \\ \boldsymbol{w}^T\boldsymbol{w}\leq\gamma \end{cases} \tag{68}$$

Obviously, Eqs. (67)–(68) are equivalent to Eqs. (61)–(62). Thus, Eqs. (63)–(64) are equivalent to Eqs. (61)–(62).

References

[1] S. Wold, H. Antti, F. Lindgren, J. Öhman, Orthogonal signal correction of near-infrared spectra, Chemom. Intell. Lab. Syst. 44 (1) (1998) 175–185.

[2] J.A. Westerhuis, S. de Jong, A.K. Smilde, Direct orthogonal signal correction, Chemom. Intell. Lab. Syst. 56 (1) (2001) 13–25.

[3] D.K. Pedersen, H. Martens, J.P. Nielsen, S.B. Engelsen, Near-infrared absorption and scattering separated by extended inverted signal correction (eisc): analysis of near-infrared transmittance spectra of single wheat seeds, Appl. Spectrosc. 56 (9) (2002) 1206–1214.

[4] Q. Chen, J. Zhao, M. Liu, J. Cai, J. Liu, Determination of total polyphenols content in green tea using ft-nir spectroscopy and different PLS algorithms, J. Pharm. Biomed. Anal. 46 (3) (2008) 568–573.

[5] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[6] P. Wentzell, S. Hou, Exploratory data analysis with noisy measurements, J. Chemom. 26 (6) (2012) 264–281.

[7] P.D. Wentzell, A.C. Tarasuk, Characterization of heteroscedastic measurement noise in the absence of replicates, Anal. Chim. Acta 847 (2014) 16–28.

[8] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1) (1987) 37–52.

[9] A.E. Hoerl, R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.

[10] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Methodol. (1996) 267–288.

[11] I. Inza, P. Larrañaga, R. Etxeberria, B. Sierra, Feature subset selection by Bayesian network-based optimization, Artif. Intell. 123 (1) (2000) 157–184.
[12] H.-L. Wei, S.A. Billings, Feature subset selection and ranking for data dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 162–166.
[13] S. Wold, A. Ruhe, H. Wold, W. Dunn III, The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses, SIAM J. Sci. Stat. Comput. 5 (3) (1984) 735–743.
[14] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[15] F. Lindgren, P. Geladi, A. Berglund, M. Sjöström, S. Wold, Interactive variable selection (ivs) for PLS. part ii: Chemical applications, J. Chemom. 9 (5) (1995) 331–342.
[16] P. Shan, S. Peng, Y. Bi, L. Tang, C. Yang, Q. Xie, C. Li, Partial least squares–slice transform hybrid model for nonlinear calibration, Chemom. Intell. Lab. Syst. 138 (2014) 72–83.
[17] Y.-Q. Li, Y.-F. Liu, D.-D. Song, Y.-P. Zhou, L. Wang, S. Xu, Y.-F. Cui, Particle swarm optimization-based protocol for partial least-squares discriminant analysis: Application to ^{1}H nuclear magnetic resonance analysis of lung cancer metabonomics, Chemom. Intell. Lab. Syst. 135 (2014) 192–200.
[18] B. Li, A.J. Morris, E.B. Martin, Generalized partial least squares regression based on the penalized minimum norm projection, Chemom. Intell. Lab. Syst. 72 (1) (2004) 21–26.
[19] H. Wold, Path models with latent variables: The NIPALS approach, Acad. Press, 1975.
[20] A. Sharma, M.A. Haj, J. Choi, L.S. Davis, D.W. Jacobs, Robust pose invariant face recognition using coupled latent space discriminant analysis, Comput. Vis. Image Underst. 116 (11) (2012) 1095–1110.
[21] G. Ziegler, R. Dahnke, A. Winkler, C. Gaser, Partial least squares correlation of multivariate cognitive abilities and local brain structure in children and adolescents, NeuroImage 82 (2013) 284–294.
[22] R. Rosipal, N. Krämer, Overview and recent advances in partial least squares, Subspace, latent structure and feature selection, Springer 2006, pp. 34–51.
[23] A. Höskuldsson, PLS regression methods, J. Chemom. 2 (3) (1988) 211–228.
[24] P.D. Sampson, A.P. Streissguth, H.M. Barr, F.L. Bookstein, Neurobehavioral effects of prenatal alcohol: Part II. partial least squares analysis, Neurotoxicol. Teratol. 11 (5) (1989) 477–491.
[25] H.D. Vinod, Canonical ridge and econometrics of joint production, J. Econ. 4 (2) (1976) 147–166.
[26] K.J. Worsley, J.-B. Poline, K.J. Friston, A. Evans, Characterizing the response of PET and fMRI data using multivariate linear models, NeuroImage 6 (4) (1997) 305–319.
[27] L. Wangen, B. Kowalski, A multiblock partial least squares algorithm for investigating complex chemical systems, J. Chemom. 3 (1) (1989) 3–20.
[28] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, J. Chemom. 12 (5) (1998) 301–321.
[29] A. Smilde, R. Bro, P. Geladi, Multi-way analysis: applications in the chemical sciences, John Wiley & Sons, 2005.
[30] Q. Zhao, C.F. Caiafa, D.P. Mandic, Z.C. Chao, Y. Nagasaka, N. Fujii, L. Zhang, A. Cichocki, Higher order partial least squares (HOPLS): A generalized multilinear regression method, IEEE Trans. Pattern Anal. Mach. Intell. 35 (7) (2013) 1660–1673.
[31] R. Rosipal, L.J. Trejo, Kernel partial least squares regression in reproducing kernel hilbert space, J. Mach. Learn. Res. 2 (2002) 97–123.
[32] R. Rosipal, L.J. Trejo, B. Matthews, Kernel PLS-SVC for linear and nonlinear classification, ICML 2003, pp. 640–647.
[33] J. Arenas-Garca, K.B. Petersen, L.K. Hansen, Sparse kernel orthonormalized PLS for feature extraction in large data sets, Adv. Neural Inf. Process. Syst. 19 (2007) 33–40.
[34] L.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools, Technometrics 35 (2) (1993) 109–135.
[35] I.S. Helland, Some theoretical aspects of partial least squares regression, Chemom. Intell. Lab. Syst. 58 (2) (2001) 97–107.
[36] N.A. Butler, M.C. Denham, The peculiar shrinkage properties of partial least squares regression, J. R. Stat. Soc. Ser. B (Stat Methodol.) 62 (3) (2000) 585–593.
[37] O.C. Lingjaerde, N. Christophersen, Shrinkage structure of partial least squares, Scand. J. Stat. 27 (3) (2000) 459–473.
[38] I.S. Helland, S. Saebø, H. Tjelmeland, et al., Near optimal prediction from relevant components, Scand. J. Stat. 39 (4) (2012) 695–713.
[39] M. Stone, R.J. Brooks, Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression, J. R. Stat. Soc. Ser. B Methodol. (1990) 237–269.
[40] S. De Jong, H.A. Kiers, Principal covariates regression: part I. Theory, Chemom. Intell. Lab. Syst. 14 (1) (1992) 155–164.
[41] A. Höskuldsson, Variable and subset selection in PLS regression, Chemom. Intell. Lab. Syst. 55 (1) (2001) 23–38.
[42] U. Indahl, A twist to partial least squares regression, J. Chemom. 19 (1) (2005) 32–44.
[43] A. Höskuldsson, The H-principle in modelling with applications to chemometrics, Chemom. Intell. Lab. Syst. 14 (1) (1992) 139–153.
[44] S.R. Searle, G. Casella, C.E. McCulloch, Variance components, vol. 391, John Wiley & Sons 2009, pp. 262–264.
[45] D.S. Riggs, J.A. Guarnieri, S. Addelman, Fitting straight lines when both variables are subject to error, Life Sci. 22 (13) (1978) 1305–1360.
[46] L. Meites, H. Smit, G. Kateman, The effects of errors in measuring the independent variable in least-squares regression analysis, Anal. Chim. Acta 164 (1984) 287–291.
[47] C. Frost, S.G. Thompson, Correcting for regression dilution bias: comparison of methods for a single predictor variable, J. R. Stat. Soc. A. Stat. Soc. 163 (2) (2000) 173–189.
[48] G.H. Golub, P.C. Hansen, D.P. O'Leary, Tikhonov regularization and total least squares, SIAM J. Matrix Anal. Appl. 21 (1) (1999) 185–194.
[49] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemom. 16 (3) (2002) 119–128.
[50] B. Wise, N. Gallagher, http://www.fermentas.com/techinfo/nucleicacids/maplambda.htm.
[51] H. Li, Q. Xu, Y. Liang, libPLS: An integrated library for partial least squares regression and discriminant analysis, PeerJ PrePrints, 22014. e190v1 (source codes available at http://www.libpls.net/).
[52] Z. Xiaobo, Z. Jiewen, M.J. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta 667 (1) (2010) 14–32.
[53] T. Chen, E. Martin, Bayesian linear regression and variable selection for spectroscopic calibration, Anal. Chim. Acta 631 (1) (2009) 13–21.
[54] J. Peng, L. Li, Y.Y. Tang, Combination of activation functions in extreme learning machines for multivariate calibration, Chemom. Intell. Lab. Syst. 120 (2013) 53–58.
[55] L. Xu, Y.-P. Zhou, L.-J. Tang, H.-L. Wu, J.-H. Jiang, G.-L. Shen, R.-Q. Yu, Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration, Anal. Chim. Acta 616 (2) (2008) 138–143.
[56] T. Chen, J. Morris, E. Martin, Gaussian process regression for multivariate spectroscopic calibration, Chemom. Intell. Lab. Syst. 87 (1) (2007) 59–71.
[57] E. Frank, L. Trigg, G. Holmes, I.H. Witten, Technical note: Naive bayes for regression, Mach. Learn. 41 (1) (2000) 5–25.
[58] K. Bennett, M. Embrechts, An optimization perspective on kernel partial least squares regression, Nato Science Series sub series III computer and systems sciences, 190 2003, pp. 227–250.
[59] Q.-S. Xu, Y.-Z. Liang, H.-L. Shen, Generalized PLS regression, J. Chemom. 15 (3) (2001) 135–148.