J Math Chem (2014) 52:2197–2209
DOI 10.1007/s10910-014-0380-7

ORIGINAL PAPER

Sparse matrix transform based weight updating in partial least squares regression

Jiangtao Peng

Received: 2 April 2014 / Accepted: 20 June 2014 / Published online: 1 July 2014

Abstract Regression from high dimensional observation vectors is particularly difficult when training data is limited. Partial least squares (PLS) partly solves the high dimensional regression problem by projecting the data to latent variables space. The key issue in PLS is the computation of weight vector which describes the covariance between the responses and observations. For small-sample-size and high-dimensional regression problem, the covariance estimation is usually inaccurate and the correlated components in the predictors will distort the PLS weight. In this paper, we propose a sparse matrix transform (SMT) based PLS (SMT-PLS) method for high-dimensional spectroscopy regression. In SMT-PLS, the observation data is first decorrelated by SMT. Then, in the decorrelated data space, the PLS loading weight is computed by least squares regression. SMT technique provides an accurate data covariance estimation, which can overcome the effect of small-sample-size and benefit both the PLS weight computation and subsequent regression prediction. The proposed SMT-PLS method is compared, in terms of root mean square errors of prediction, to PLS, Power PLS and PLS with orthogonal scatter correction on four real spectroscopic data sets. Experimental results demonstrate the efficacy and effectiveness of our proposed method.

Keywords Partial least squares · Sparse matrix transform · Weight updating · High-dimensional small-sample · Spectroscopy regression

J. Peng (✉)
Faculty of Mathematics and Statistics, Hubei University, Wuhan 430062, China
e-mail: pengjt1982@126.com

1 Introduction

Considering the general linear regression, given $X^{\mathrm{T}} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]_{p\times n}$ with n samples and p variables, the response $\mathbf{y}$ is predicted by: $\mathbf{y} = X\boldsymbol{\beta}$. When X is full rank, the regression coefficients $\boldsymbol{\beta}$ can be solved by ordinary least squares (OLS) which uses the sample covariance. However, when the number of variables is large compared to the number of observations, the sample covariance is singular and OLS approach is unstable and no longer feasible. Different regularization and shrinkage methods, such as ridge regression and Lasso [1], are developed to cope with the ill-posed problems.

Partial least squares (PLS) [2] is another alternative method for addressing the high-dimensional small-sample regression problem. PLS is one of the most widely used multivariate calibration methods [3]. The intention of PLS is to summarize the high-dimensional predictor variables into a smaller set of uncorrelated components (called latent variables), which have a maximal covariance to the responses. It is followed by a regression step where the latent variables are used to predict the responses.

In the implementation of PLS, the well-known nonlinear iterative partial least squares (NIPALS) algorithm [2] is commonly used for computation of the successive PLS components. The weight vector is first computed, then the scores and loadings can be solved successively. The crux is the computation of weight vector. The PLS weight is proportional to the covariances between the responses and observations and is computed by using least squares regression. For high-dimensional small-sample spectroscopy regression, the correlated components in predictor variables will affect the least squares computation of PLS weight. In addition, the limited training samples makes the ordinary least squares regression inaccurate because the sample covariance matrix tends to distort the eigenstructure of the true covariance matrix in this case of inadequate data. Thus, it needs to eliminate the effect of correlated components and small-sample-size problem in the computation of PLS weight and covariance estimation.

Sparse matrix transform (SMT) is recently proposed to estimate the covariance matrix [4–6]. The covariance is constrained to be have an eigen-decomposition that can be represented as an SMT. The SMT is formed by a product of pairwise coordinate Givens rotations. Under this framework, the covariance can be efficiently estimated using a simple recursive local optimization procedure [4]. And the estimated covariance is always positive definite and well-conditioned. Previous results have shown that SMT method is very accurate even in small-sample-size case, and yields good results in covariance estimation and analysis of high dimensional signals [4–6].

Motivated by the superiority performance of SMT in accurate covariance estimation for small-sample and high-dimensional problem, we propose a sparse matrix transform based PLS (SMT-PLS) method for spectroscopy regression. SMT brings accurate data covariance and its eigenstructure estimation, which can be used to decorrelate the high dimensional spectroscopy data. It thus alleviates the effect of correlated variables to the least squares computation of PLS weight. Then PLS weight updating can be performed by least squares in the decorrelated data space. Finally, the obatined regression coefficient based on the decorrelated PLS model is asymptotically more efficient than the estimator based on the original PLS model.

2 The Algorithm

2.1 Sparse matrix transform

Sparse matrix transform (SMT) is originally designed to estimate the covariance matrix [4–6]. Consider a set of n samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ taking values in a p-dimensional space, and assume $\mathbf{x}_k$ has zero mean. The sample covariance is computed by $S = X^{\mathrm{T}}X/n$, where $X^{\mathrm{T}} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, and S is an unbiased estimation of the true covariance matrix R. The eigen-decomposition of R is:

$$R = E\Lambda E^{\mathrm{T}} \tag{1}$$

where E is the orthogonal eigenvector matrix and Λ is the diagonal matrix of eigenvalues. Within the maximum-likelihood framework [4], the estimates of E and Λ are

$$\hat{E} = \arg\min_{E\in\Omega} \left\{ \left| \mathrm{diag}(E^{\mathrm{T}}SE) \right| \right\} \tag{2}$$

$$\hat{\Lambda} = \mathrm{diag}(\hat{E}^{\mathrm{T}}S\hat{E}) \tag{3}$$

where Ω is the set of allowed orthogonal transforms, and $\hat{R} = \hat{E}\hat{\Lambda}\hat{E}^{\mathrm{T}}$ is the maximum-likelihood estimation of the covariance. If Ω is the set of all orthogonal matrices and S is full rank, the maximum-likelihood estimate of the covariance is given by the sample covariance: $\hat{R} = S$.

The key idea of SMT is to restrict the set Ω [5,6] such that E is represented as the product of finite (or 'sparse') K Givens rotations:

$$E = E_1 E_2 \cdots E_K \tag{4}$$

each of which is a simple rotation of angle θ_k about two axes i_k and j_k. That is, each rotation is given by a matrix of the form:

$$E_k = I + \Theta(i_k, j_k, \theta_k) \tag{5}$$

where

$$\Theta(i_k, j_k, \theta_k)_{rs} = \begin{cases} \cos(\theta_k) - 1, & \text{if } r = s = i_k \text{ or } r = s = j_k \\ \sin(\theta_k), & \text{if } r = i_k \text{ and } s = j_k \\ -\sin(\theta_k), & \text{if } r = j_k \text{ and } s = i_k \\ 0, \text{otherwise.} \end{cases}$$

The aim is to produce an estimation of the eigenvector matrix that is sparsely parametrized by a limited number of rotations. The coordinates, i_k and j_k, angle θ_k, and hence E_k and Λ_k can be iteratively determined by a greedy alternative optimization method [4–6]. At each iteration, two most correlated coordinates, i_k and j_k are first determined by minimizing the cost of (2), which results in

$$(i_k, j_k) \leftarrow \arg\min_{(i,j)} \left(1 - \frac{S_{ij}^2}{S_{ii} S_{jj}}\right) \tag{6}$$

Once i_k and j_k are determined, the Givens rotation E_k^* is given by

$$E_k^* = I + \Theta(i_k, j_k, \theta_k)$$

where $\theta_k = \frac{1}{2}\text{atan}(-2S_{i_k j_k}, S_{i_k i_k} - S_{j_k j_k})$. After K Givens rotations, the eigen-decomposition results of the covariance $\hat{E}$ and $\hat{\Lambda}$ can be obtained.

When the number of data samples is small, the sample covariance estimator is singular and will over-fit the data. While SMT estimator is usually positive definite as it regularizes the maximum-likelihood estimate (2) by constraining feasible set Ω to be the set of orthonormal transforms that can be represented as an SMT of order K. The regularization makes the estimator well-conditioned. By varying the order K of the SMT, it can reduce or increase the regularizing constraint on the covariance.

2.2 Sparse matrix transform for PLS

The well-known nonlinear iterative partial least squares (NIPALS) algorithm [2] is commonly used for computation of the successive PLS components in order to maximize the covariance structure between the predictors and the responses [7]. Let X be the $n \times p$ predictors matrix, whose rows and columns correspond to samples and spectral variables respectively, and $\mathbf{y}$ be the $n \times 1$ response vector, PLS finds unit weight vector $\mathbf{w} \in \mathcal{R}^p$ such that:

$$[\text{cov}(\mathbf{t}, \mathbf{y})]^2 = [\text{cov}(X\mathbf{w}, \mathbf{y})]^2 = \mathbf{w}^{\text{T}} X^{\text{T}} \mathbf{y} \mathbf{y}^{\text{T}} X \mathbf{w} \tag{7}$$

is maximized.

It can be shown that the desired weight vector:

$$\hat{\mathbf{w}} = X^{\text{T}} \mathbf{y} / \| X^{\text{T}} \mathbf{y} \| \tag{8}$$

Based on NIPALS, the scores and loadings can be solved successively. The weight is proportional to the covariances between $\mathbf{y}$ and the corresponding X-variables. The correlated components in X-variables will affect the computation of PLS weight. In this paper, SMT technique is used to decorrelate the observations and then the PLS weight updating is performed on the decorrelated data space.

The key idea is to use SMT to decorrelate the spectroscopy data. Next, we investigate the effects of correlated components on PLS regression. The observations X can be represented as:

$$X = T P^{\text{T}} \tag{9}$$

where score matrix $T = [\mathbf{t}_1, \ldots, \mathbf{t}_q]$ and loading matrix $P = [\mathbf{p}_1, \ldots, \mathbf{p}_q]$.

In NIPALS, the score $\mathbf{t}_i$, weight $\mathbf{w}_i$, and data X_i have the following relations:

$$\begin{aligned} \mathbf{t}_i &= X_i \mathbf{w}_i, \\ X_{i+1} &= X_i - \mathbf{t}_i \mathbf{p}_i^{\mathrm{T}} \end{aligned}$$

Based on the iteration formulation, we can obtain $\mathbf{t}_i = X\mathbf{h}_i$, where $\mathbf{h}_1 = \mathbf{w}_1$ and $\mathbf{h}_i = (I - \mathbf{w}_1\mathbf{p}_1^{\mathrm{T}}) \cdots (I - \mathbf{w}_{i-1}\mathbf{p}_{i-1}^{\mathrm{T}})$ for $i \geq 2$. Denote $H = [\mathbf{h}_1, \ldots, \mathbf{h}_q]$, then

$$T = XH \tag{10}$$

Based on Eq. (9), the linear relations can express as:

$$\mathbf{y} = X\beta + \mathbf{e} = TP^{\mathrm{T}}\beta + \mathbf{e} = T\alpha + \mathbf{e} \tag{11}$$

The least squares solution of Eq. (11) is

$$\hat{\alpha} = (T^{\mathrm{T}}T)^{-1}T^{\mathrm{T}}\mathbf{y} \tag{12}$$

Based on Eq. (10), the fitted value of $\mathbf{y}$ is

$$\mathbf{y} = T\hat{\alpha} = XH(T^{\mathrm{T}}T)^{-1}H^{\mathrm{T}}X^{\mathrm{T}}\mathbf{y} \tag{13}$$

So, the PLS regression coefficient vector $\hat{\beta}$ is

$$\hat{\beta} = H(T^{\mathrm{T}}T)^{-1}H^{\mathrm{T}}X^{\mathrm{T}}\mathbf{y} = H(T^{\mathrm{T}}T)^{-1}H^{\mathrm{T}}X^{\mathrm{T}}X\hat{\beta}_{LS} \tag{14}$$

where $\hat{\beta}_{LS}$ is the least squares solution. Denote the true regression coefficient as β_0, based on the bias-variance decomposition, the mean squared error (MSE) of $\hat{\beta}$ can be computed:

$$MSE(\hat{\beta}) = E[(\hat{\beta} - \beta_0)^{\mathrm{T}}(\hat{\beta} - \beta_0)] = \sigma^2 \sum_{i=1}^{q} \frac{\mathbf{h}_i^{\mathrm{T}}\mathbf{h}_i}{\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i} + \|E\hat{\beta} - \beta_0\|^2 \tag{15}$$

where q is the number of scores and $\sigma = \mathrm{Var}(\mathbf{e})$ is the noise variance.

Based on singularity value decomposition (SVD), $X = USV^{\mathrm{T}} \doteq TP^{\mathrm{T}}$, where $T = US$ and $P = V$. When there exists high collinearity in the data set X, some eigenvalues s_i of data X will very small and the corresponding components $\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i = s_i^2 \approx 0$, which makes the $MSE(\hat{\beta})$ in Eq. (15) very large. So, if all the components are introduced into the model (as in least squares model), the estimator will produce large variance. To tackle this problem, PCR and PLS use $k(k < q)$ principal components to construct the model, leaving out the components with very small variance so that the noise can be removed and the collinearity of the data set can be reduced [8].

Form the above analysis, we can see that the correlated components in the predictors pose a serious threat to the regression analysis. The corresponding estimator produces

large variance and hence large error. To decrease the MSE and achieve an accurate and robust prediction, the observation data should be decorrelated.

In order to decorrelate the high dimensional spectroscopic data, we first use a sparse matrix transform technique to estimate the eigenvector and eigenvalue matrices of data covariance as $\hat{E}$ and $\hat{\Lambda}$, respectively. Then, the data are whitened as follows:

$$\widetilde{X} = X\hat{E}\hat{\Lambda}^{-\frac{1}{2}} \tag{16}$$

Note that, for the decorrelated data $\widetilde{X}$, the score vector satisfies $\tilde{\mathbf{t}}_i^{\mathrm{T}}\tilde{\mathbf{t}}_i = \tilde{s}_i^2 \approx 1$. Thus, the unwanted large variance will not appear in the MSE. So, the estimator based on decorelated data is asymptotically more efficient than the estimator based on the original correlated data.

Then, we solve the PLS weight based on the decorrelated data

$$\widetilde{\mathbf{w}} = \frac{1}{n}\widetilde{X}^{\mathrm{T}}\mathbf{y} \tag{17}$$

Note that, the weight $\widetilde{\mathbf{w}}$ actually describes the relation between $\mathbf{y}$ and $\widetilde{X}$, that is $\mathbf{y} = \widetilde{X}\widetilde{\mathbf{w}}$ as $\widetilde{X}^{\mathrm{T}}\widetilde{X}/n = I$ for the whitened observations.

The PLS weight based on SMT can be expressed as

$$\mathbf{w} = \hat{E}\hat{\Lambda}^{-\frac{1}{2}}\widetilde{\mathbf{w}} \tag{18}$$

Algorithm 1 describes the sparse matrix transform based partial least squares (SMT-PLS) algorithm. In extracting each SMT-PLS component, SMT is first used to perform an eigen-decomposition of covariance matrix, and the estimated eigenvector and eigenvalue matrices are used to decorelate the observations. Then, the PLS weight vector is computed based on the decorrelated data, and the score and loading vectors are resolved successively. The above procedures are repeated until the desired components are extracted.

Algorithm 1 SMT-PLS algorithm for weight updating.

Given X, $\mathbf{y}$, and K rotations.
Let $X_0 = X$, $a = 1$.
1: Perform SMT on the observations: $\hat{R}_{X_{a-1}} = \hat{E}\hat{\Lambda}\hat{E}^{\mathrm{T}}$
2: Decorrelate the data: $\widetilde{X}_{a-1} = X_{a-1}\hat{E}\hat{\Lambda}^{-\frac{1}{2}}$
3: Solve the weight on decorrelated data: $\widetilde{\mathbf{w}}_a = \widetilde{X}_{a-1}^{\mathrm{T}}\mathbf{y}$
4: Compute the SMT-PLS weight: $\mathbf{w}_a = \hat{E}\hat{\Lambda}^{-\frac{1}{2}}\widetilde{\mathbf{w}}_a$
5: Normalize: $\mathbf{w}_a = \mathbf{w}_a/\|\mathbf{w}_a\|$.
6: Calculate the scores $\mathbf{t}_a = X\mathbf{w}_a$ and loadings $\mathbf{p}_a = X^{\mathrm{T}}\mathbf{t}_a$.
7: Deflate: $X_a = X_{a-1} - \mathbf{t}_a\mathbf{p}_a^{\mathrm{T}}$.
8: Let $a = a + 1$, and return to step 1 until A components are extracted.
Note: when $K = 0$, SMT is not performed and this algorithm is the PLS1.

3 Experimental

3.1 Data sets

Data set 1 consists of NIR spectra from 310 pharmaceutical tablet samples with a relative active substance content (%, w/w) in the range of 4.6–9.8 % [9,10]. The transmittance spectra have 404 variables collected in the range of 7,400–10,507 cm^{-1}. The 310 NIR spectra are divided into 210 calibration samples and 100 prediction samples based on Kennard–Stone (KS) algorithm [11].

Data set 2 is from the Software Shootout at the IDRC98 containing NIR spectra of 141 fescue grass powdered (dry ground) samples with specified carbon, nitrogen and sulphur contents ranging from 29.6 to 40.9, 1.1 to 6.6 and 0.3 to 1.7 %, respectively. The related chemical values are the average of the blind duplicates determined on a LECO CNS-2000 Carbon, Nitrogen and Sulphur Analyzer [9].

Data set 3 consists of 32 marzipan FTIR spectra with traditional moisture and sugar contents ranging from 7 to 19, and 33 to 68 %, respectively. The spectra in the region 6,500–650 cm^{-1} have been recorded with Perkin Elmer System 2000, equipped with the horizonal ATR Sampling Accessory (ZnSe cell) [9,12]. The 32 marzipan IR spectra are divided into 24 calibration samples and 8 prediction samples based on the KS algorithm.

Data set 4 consists of NIR transmittance spectra of meat samples [13]. The spectra have been recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850–1,050 nm. For each meat sample the data consists of a 100 channel spectrum of absorbances and the contents of moisture (water), fat and protein. The three contents, measured in percent, are determined by analytic chemistry. The data contain 172 training samples and 43 testing samples.

3.2 Model selection

The choice of latent variable number in the calibration model will be a balance between minimizing the predicted residual error sum of squares (PRESS) and limiting the model complexity. The smallest model (fewest number of latent variables) such that the PRESS for this model is not significantly greater than the minimum PRESS is adopted. We use the F-test criterion pointed out by Haaland and Thomas [14] to test the significance of incremental changes in PRESS. In this work, F-test at 95 % confidence level is employed.

The comparison of the accuracy among different models is done by using root mean square errors of prediction (RMSEP), defined by:

$$\mathrm{RMSEP} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

where y_i and $\hat{y}_i$ are the measured and estimated values of the studied property for a sample, respectively, and N is the number of samples in the prediction set.

4 Results and discussion

4.1 Pharmaceutical tablet NIR data set

4.1.1 Covariance estimation

We first show that SMT can provide accurate covariance estimation for spectroscopic data. For this purpose, we set the true covariance R as the sample covariance S computed using all the 310 Pharmaceutical tablet NIR samples, and then randomly sample M observations from the 310 samples to estimate the covariance. We perform covariance estimation using SMT method and the following regularization method:

$$\hat{R} = \alpha S + (1 - \alpha)\text{diag}(S) \tag{19}$$

The sample sizes M ranging from 20 to 80 are considered. The Kullback-Leibler (KL) distance [5] which measures the error between the estimated and true distribution is used to assess the performance of the two covariance estimation methods. Figure 1 shows the KL distances of the two estimators as a function the sample number M. The error bars indicate the standard deviation of the KL distance due to random variation in the sample selection. It clearly shows that the KL distances of SMT covariance estimation are consistently and substantially smaller than that of the regularized method (19).

4.1.2 The effect of calibration samples

To evaluate the prediction performance of the proposed SMT-PLS method in the challenging situations with high dimensionality and small-sized calibration samples,

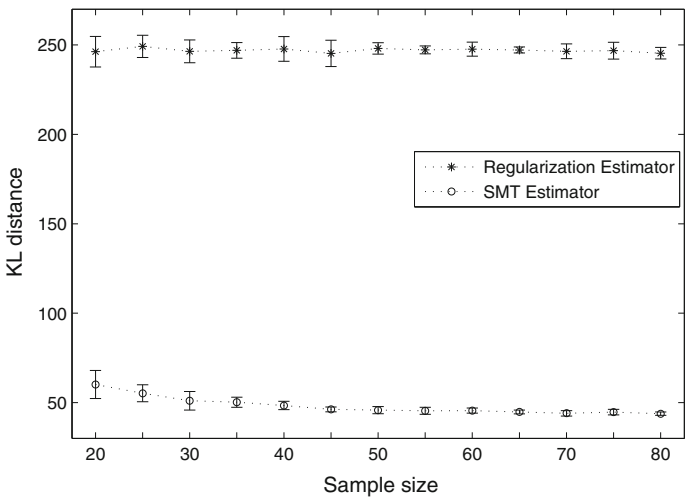

Fig. 1 Kullback–Leibler distance versus sample size

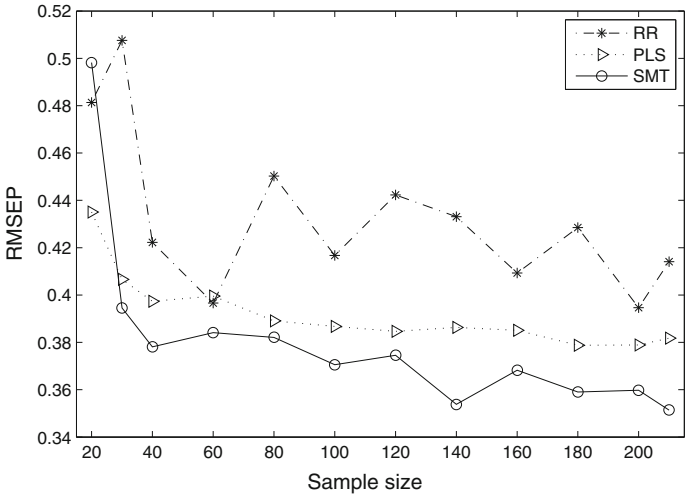

Fig. 2 RMSEP versus sample size for SMT-PLS, PLS and RR

we use parts of the original 210 calibration samples to build the model. The proposed SMT-PLS is compared with PLS and ridge regression (RR) methods. The RMSEP results as a function of the calibration sample number ranging from 20 to 210 are shown in Fig. 2, where the number of latent variables in PLS and SMT-PLS are chosen as 4 and the regularization parameter in RR is set as 10^{-4}. It can be seen that SMT-PLS provides consistently smaller RMSEPs than the other two methods when the number of training samples is not less than 30. It demonstrates that SMT-PLS can solve the high dimensional spectroscopy regression problem even with a small number of training samples.

4.1.3 The effect of SMT model orders

In the following, we investigate the effect of SMT model order K (i.e., the number of Givens rotations) on the SMT-PLS regression model. We show the 10-fold cross validation errors (RMSECV) versus different rotations K in Fig. 3, where the model order K changes from 10 to 3,000. It can be seen that the RMSECVs show an overall downtrend as the order K increases and are stable when K is not less than 500. From the changes of RMSECV results, we empirically set K to be 500.

4.1.4 Comparison of different data transform methods

The key point in SMT-PLS is the data decorrelation transform in each PLS term. In this part, we will compare SMT-PLS with modified PLS (MPLS) [15] and singular value decomposition (SVD) based data decorrelation transform PLS method (SVD-PLS). The MPLS method scales the spectroscopic data at each wavelength to have a standard deviation of 1 before each PLS term. The SVD method can also provide a diagonal structure of the covariance. In SVD-PLS, we use SVD to decorrelate the

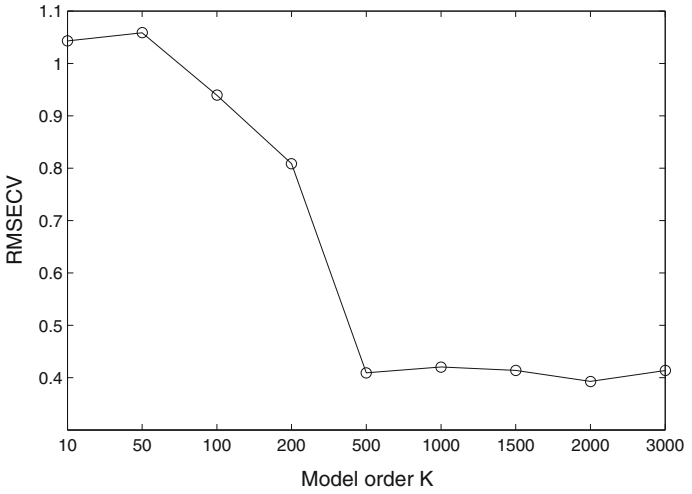

Fig. 3 RMSECV versus different model order K

Table 1 RMSEP of PLS, MPLS, SVD-PLS, SMT-PLS

	PLS	MPLS	SVD-PLS	SMT-PLS
RMSEP	0.3814	1.2553	Inf	0.3528
RMSEPpart	0.5599	1.2851	0.7152	0.5303

data before each PLS term. The models are build on the calibration set (210 samples), and the prediction results of RMSEP values are shown in Table 1. As the sample dimension (variable number) is 404, the data covariance is singular and SVD-PLS method provides meaningless result. So, we use the first 100 dimensions to build the models again, and the results is also recorded ('RMSEPpart'). From the table, it can be seen that MPLS and SVD-PLS do not improve the PLS method. Compared to SVD, SMT method can obtain positive definite and well-conditioned covariance estimator even with limited samples. Moreover, SMT estimator is more accurate than SVD even if the covariance matrix is nonsingular because SMT can be considered as a regularized SVD method [5]. Compared to PLS, SMT-PLS provides more accurate prediction. The plots of real values versus predicted values of the active substance content for PLS and SMT-PLS models are shown in Fig. 4. The plots show in a nice way that the fits and predictions of SMT-PLS are more accurate.

4.2 Fescue grass NIR data set

The 141 grass NIR spectra are divided into 100 calibration samples and 41 prediction samples based on the KS algorithm. In order to investigate the performance in small-sample-size case, we also consider the regression in the case of 41 calibration samples and 100 prediction samples, where the number of calibration samples is relatively small compared to the high dimensionality (1,050 dimensions) and also smaller than the number of prediction samples.

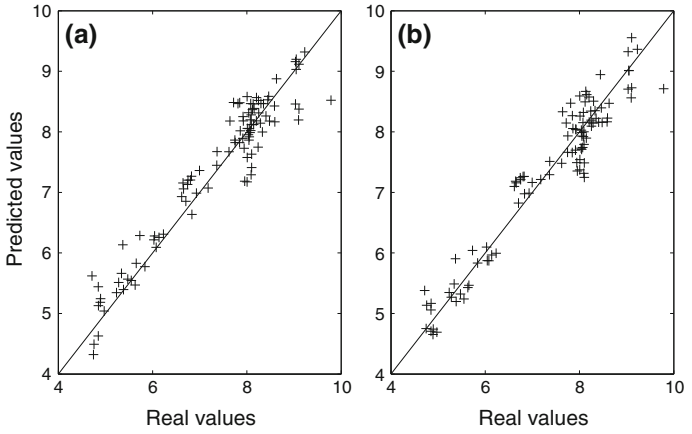

Fig. 4 Real values versus predicted values of the active substance content: **a** PLS; **b** SMT-PLS

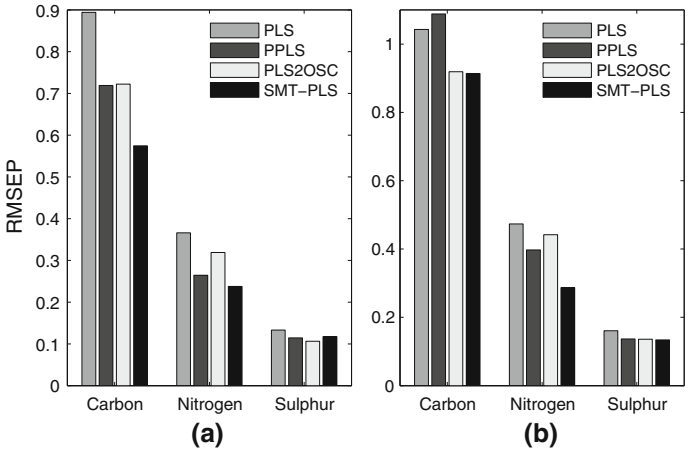

Fig. 5 RMSEP results on NIR grass data set with different number of calibration samples. **a** 100 calibration samples, **b** 41 calibration samples

As the proposed SMT-PLS focuses on the weight computation and the effect of correlated components, Power PLS (PPLS) and PLS with orthogonal scatter correction (PLS2OSC) are compared. PPLS [16] aims to increase flexibility in the computation of weights. It computes the weight vector by taking powers of correlations and standard deviations, which neutralizes the influence of dominance of irrelevant X-variance and spurious **y**-correlations. PLS2OSC [17] removes the non-correlated systematic variation, which improves the interpretation of PLS and reduces model complexity.

We build the prediction model for carbon, nitrogen, and sulphur, respectively. The optimal number of latent variables is chosen based on the 10-fold cross validation. The comparisons of predicted RMSEP results of PLS, PPLS, PLS2OSC and SMT-PLS are shown in Fig. 5, where the models built on 100 calibration samples and 41 calibration samples are considered, respectively. It can be seen that PPLS, PLS2OSC

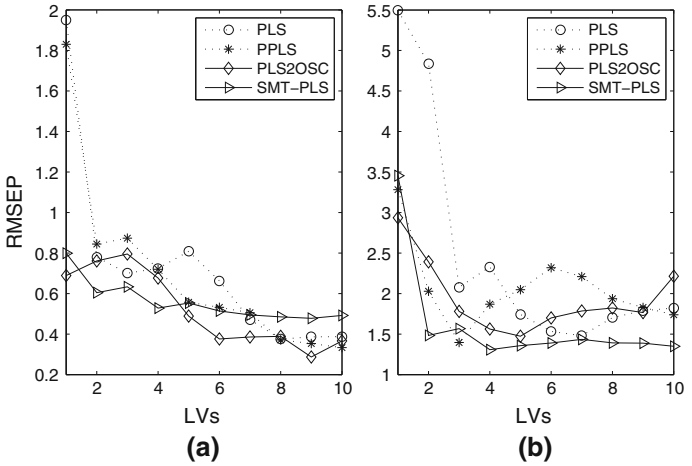

Fig. 6 RMSEP of PLS, PPLS, PLS2OSC and SMT-PLS on different latent variables for marzipan IR data set: **a** moisture; **b** sugar

Table 2 RMSEP results on Meat NIR data set

	PLS	PPLS	PLS2OSC	SMT-PLS
Moisture	2.761	2.776	2.189	1.774
Fat	3.048	2.686	2.594	2.259
Protein	0.934	0.887	0.615	0.684

and SMT-PLS improve PLS in both cases. Compared to PPLS and PLS2OSC, SMT-PLS provides better predictions for the carbon and nitrogen prediction tasks, and comparable results for sulphur content prediction. In both cases with large and small number of training samples, SMT-PLS provides better overall performance.

4.3 Marzipan IR data set

The comparisons of predicted RMSEP-values for PLS, PPLS, PLS2OSC and SMT-PLS on Marzipan IR data set with different latent variables are shown in Fig. 6. It can be seen from the figure that on sugar content, SMT-PLS outperforms PLS, PPLS and PLS2OSC for almost all the latent variables. On moisture content, SMT-PLS achieves better results when the number of latent variables is smaller than 5. Because there are only 24 calibration samples compared to 950 dimensions, the results show that SMT-PLS is suitable for high-dimensional and small-sample regression.

4.4 Meat NIR data set

The predicted results for moisture, fat, protein on the optimal latent variables are shown in Table 2. Except for the protein prediction of PLS2OSC, SMT-PLS achieves consistently better results than other three methods. The results demonstrate that the

regression on the decorrelated data can provide better results, and the PLS weight updating based on SMT indeed plays a role in the prediction.

5 Conclusions

In this paper, we have proposed a new weight updating strategy in PLS regression based on SMT. SMT technique provides accurate data covariance and its eigenstructure estimation which benefits the PLS weight computation and subsequent regression analysis. In particular, the SMT decorelation operation alleviates the effect of correlated variables to the least squares computation of PLS weight, and the obtained regression coefficient based on the decorrelated data is asymptotically more efficient than the estimator based on the original data. Experimental results demonstrate that SMT-PLS provides better predictions on different spectroscopic data sets.

Note that, SMT can be considered as a preprocessing step for decorrelating the data. When the data is decorrelated, different regularized regression procedures can be used in resolving the PLS weight, such as Lasso and shrinkage methods (soft-threshold or hard-threshold). If the sparse solution is preferred, it should pay more attention on the parameter selection, such as the Lasso parameter (regularization parameter or the number of nonzero elements), the soft or hard thresholds in shrinkage methods.

Acknowledgments This work was supported by the National Natural Science Foundation of China under Grants No. 11071058.

References

1. R. Tibshirani, J. R. Stat. Soc. B **58**, 267 (1996)
2. H. Wold, *Perspectives in Probability and Statistics* (Academic Press, London, 1975)
3. H. Martens, T. Nas, *Multivariate Calibration* (Wiley, New York, 1989)
4. G. Cao, C.A. Bouman, Adv. Neural Inf. Process. Syst. **21**, 225 (2009)
5. G. Cao, L.R. Bachega, C.A. Bouman, IEEE Trans. Image Process. **20**, 625 (2011)
6. J. Theiler, G. Cao, L.R. Bachega, C.A. Bouman, IEEE J. Sel. Topic Signal Process. **5**, 424 (2011)
7. J. Peng, S. Peng, Q. Xie, J. Wei, Anal. Chim. Acta **690**, 162 (2011)
8. Q.S. Xu, Y.Z. Liang, H.L. Shen, J. Chemom. **15**, 135 (2001)
9. http://www.models.life.ku.dk/datasets
10. M. Dyrby, S.B. Engelsen, L. Nørgaard, M. Bruhn, L. Nielsen, Appl. Spectrosc. **56**, 579 (2002)
11. R.W. Kennard, L.A. Stone, Technometrics **11**, 137 (1969)
12. J. Christensen, L. Nørgaard, H. Heimda, J.G. Pedersen, S.B. Engelsen, J. Near Infrared Spectrosc. **12**, 63 (2004)
13. http://lib.stat.cmu.edu/datasets/tecator
14. D.M. Haaland, E.V. Thomas, Anal. Chem. **60**, 1193 (1988)
15. J.S. Shenk, M.O. Westerhaus, Crop Sci. **31**, 469 (1991)
16. U. Indahl, J. Chemom. **19**, 32 (2005)
17. A. Höskuldsson, Chemom. Intell. Lab. Syst. **55**, 23 (2001)