Jointly Sparse Global SIMPLS Regression

Tzu-Yu Liu
University of California, Berkeley, CA, USA.
e-mail: tyliu@eecs.berkeley.edu

Laura Trinchera
NEOMA Business School, Mont-Saint-Aignan, France.
e-mail: laura.trinchera@neoma-bs.fr

Arthur Tenenhaus
Supélec, Gif-sur-Yvette, France.
e-mail: Arthur.Tenenhaus@supelec.fr

Dennis Wei
IBM T. J. Watson Research Center, NY, USA.
e-mail: dwei@us.ibm.com

Alfred O. Hero*
University of Michigan, Ann Arbor, MI, USA.
e-mail: hero@eecs.umich.edu

Abstract: Partial least squares (PLS) regression combines dimensionality reduction and prediction using a latent variable model. Since partial least squares regression (PLS-R) does not require matrix inversion or diagonalization, it can be applied to problems with large numbers of variables. As predictor dimension increases, variable selection becomes essential to avoid over-fitting, to provide more accurate predictors and to yield more interpretable parameters. We propose a global variable selection approach that penalizes the total number of variables across all PLS components. Put another way, the proposed global penalty encourages the selected variables to be shared among the PLS components. We formulate PLS-R with joint sparsity as a variational optimization problem with objective function equal to a novel global SIMPLS criterion plus a mixed norm sparsity penalty on the weight matrix. The mixed norm sparsity penalty is the $\ell_1$ norm of the $\ell_2$ norm on the weights corresponding to the same variable used over all the PLS components. A novel augmented Lagrangian method is proposed to solve the optimization problem and soft thresholding for sparsity occurs naturally as part of the iterative solution. Experiments show that the modified PLS-R attains better or as good performance with many fewer selected predictor variables.

Keywords and phrases: PLS, variable selection, dimension reduction, augmented Lagrangian optimization.

* to whom correspondence should be addressed.
1. Introduction

With advancing technology comes the need to extract information from increasingly high-dimensional data, whereas the number of samples is often limited. Dimension reduction techniques and models incorporating sparsity become important solution strategies. Partial least squares regression (PLS-R) combines dimensionality reduction and prediction using a latent variable model. It was first developed for regression analysis in chemometrics [1, 2], and has been successfully applied to many different areas, including sensory science and, more recently, genetics [3, 4, 5, 6, 7].

Moreover, PLS-R algorithm is designed precisely to operate with high dimensional data. Since the first proposed algorithm does not require matrix inversion nor diagonalization but deflation to find the latent components, it can be applied to problems with large numbers of variables. The latent components reduce the dimension by constructing linear combinations of the predictors, which successfully solved the collinearity problems in chemometrics [8]. However, the linear combinations are built on all the predictors. The resulting PLS-R model tends to overfit when the number of predictors increases for a fixed number of samples. Therefore, variable selection becomes essential for PLS-R in high-dimensional sample-limited problems. It not only avoids over-fitting, but also provides more accurate predictors and yields more interpretable estimates. For this reason sparse PLS-R was developed by H. Chun and S. Keles [9]. The sparse PLS-R algorithm performs variable selection and dimension reduction simultaneously using an \( \ell_1 \) type variable selection penalty. However, the \( \ell_1 \) penalty used in [9] penalizes each variable in each component independently and this can result in different sets of variables being selected for each PLS component leading to an excessively large number of variables.
In this work we propose a global criterion for PLS that changes the sequential optimization for a K component model in Statistically Inspired Modification of PLS (SIMPLS) [10] into a unified optimization formulation, which we refer to as global SIMPLS. This enables us to perform global variable selection, which penalizes the total number of variables across all PLS components. We formulate PLS-R with global sparsity as a variational optimization problem with the objective function equal to the global SIMPLS criterion plus a mixed norm sparsity penalty on the weight matrix. The mixed norm sparsity penalty is the $\ell_1$ norm of the $\ell_2$ norm on the weights corresponding to the same variable used over all the PLS components. The proposed global penalty encourages the selected variables to be shared among all the $K$ PLS components. A novel augmented Lagrangian method is proposed to solve the optimization problem, which enables us to obtain the global SIMPLS components and to perform joint variable selection simultaneously. A greedy algorithm is proposed to overcome the computation difficulties in the iterations, and soft thresholding for sparsity occurs naturally as part of the iterative solution. Experiments show that our approach to PLS regression attains better or as good performance (lower mean squared error, MSE) with many fewer selected predictor variables. These experiments include a chemometric data set, and a human viral challenge study dataset, in addition to numerical simulations.

We review the developments in PLS-R for both univariate and multivariate responses in Section 2, in which we discuss different objective functions that have been proposed for PLS-R, particularly the Statistically Inspired Modification of PLS (SIMPLS) proposed by de Jong [10]. In Section 3, we formulate the Jointly Sparse Global SIMPLS-R by proposing a new criterion that jointly optimizes over $K$ weight vectors (components) and imposing a mixed norm sparsity penalty to select variables jointly. The algorithmic implementation is discussed in Section 4 with simulation experiments presented in Section 5. The proposed Jointly Sparse Global SIMPLS Regression is applied to two applications: (1) Chemometrics in Section 6 and (2) Predictive health studies in Section 7. Section 8 concludes the paper.

2. Partial Least Squares Regression

Partial Least Squares (PLS) methods embrace a suite of data analysis techniques based on algorithms belonging to the PLS family. These algorithms consist of various extensions of the Nonlinear estimation by Iterative PArtial Least Squares (NIPALS) algorithm that was proposed by Herman Wold [11] as an alternative algorithm for implementing Principal Component Analysis (PCA) [12]. The NIPALS approach was slightly modified by Svante Wold, and Harald Martens, in order to obtain a regularized component based regression tool, known as PLS Regression (PLS-R) [1, 8].

Suppose that the data consists of $n$ samples of $p$ independent variables $X \in \mathbb{R}^{n \times p}$ and $q$ dependent variables (responses) $Y \in \mathbb{R}^{n \times q}$. In standard PLS Regression the aim is to define orthogonal latent components in $\mathbb{R}^p$, and then use
such latent components as predictors for \( Y \) in an ordinary least squares framework. The X weights used to compute the latent components can be specified by using iterative algorithms belonging to the NIPALS family or by a sequence of eigen-decompositions. The general underlying model is \( X = TP' + E \) and \( Y = TQ' + F \), where \( T \in R^{n \times K} \) is the latent component matrix, \( P \in R^{p \times K} \) and \( Q \in R^{q \times K} \) are the loading matrices, \( K \) is the number of components, \( E \) and \( F \) are the residual terms. The latent components in \( T = [ t_1 \ t_2 \ \ldots \ t_K ] \) are linear combinations of the independent variables, hence PLS can be viewed as a dimensional reduction technique, reducing the dimension from \( p \) to \( K \). The latent components should be orthogonal to each other either by construction as in NIPALS [1, 8] or via constrained optimizations as in SIMPLS [10]. This allows PLS to build a parsimonious model for high dimensional data with collinearity [8].

2.1. Univariate Response

We assume, without loss of generality, that all the variables have been centered in a pre-processing step. For univariate \( Y \), i.e \( q = 1 \), PLS Regression, also often denoted as PLS1, successively finds X weights \( R = [ r_1 \ r_2 \ \ldots \ r_K ] \) as the solution to the constrained optimization

\[
 r_k = \arg \max_r \{ r'X'_{(k-1)}Y + X'_{(k-1)}r \} \quad \text{s.t.} \quad r'r = 1. \tag{2.1}
\]

where \( X'_{(k-1)} \) is the matrix of the residuals (i.e., the deflated matrix) from the regression of the \( X \)-variables on the first \( k - 1 \) latent components, and \( X_{(0)} = X \). These weights are then used to find the latent components \( T = [ X_{(0)}r_1 \ X_{(1)}r_2 \ \ldots \ X_{(K-1)}r_K ] \). Such components can be also expressed in terms of original variables (instead of deflated variables), i.e., as \( T = XW \), where \( W \) is the matrix containing the weights to be applied to the original variables in order to exactly obtain the latent components [13].

For a fixed number of components, the response variable \( Y \) is predicted in an ordinary least squares regression model, where the latent components play the role of the exogenous variables,

\[
 \hat{Q} = \arg \min_Q \{ ||Y - TQ'||_2 \} = (T'T)^{-1}T'Y.
\]

This provides the regression coefficients \( \hat{\beta}^{PLS} = W\hat{Q}' \) for the model \( Y = X\hat{\beta}^{PLS} + F \).

Depending on the number of selected latent components the length \( ||\hat{\beta}^{PLS}||_2 \) of the vector of PLS coefficients changes. In particular, de Jong [14] had shown that the sequence of these coefficient vectors has lengths that are strictly increasing as the number of components increases. This sequence converges to the ordinary least squares coefficient vector and the maximum number of latent components obtainable equals the rank of the \( X \) matrix. Thus, by using a
number of latent components $K < \text{rank}(X)$, PLS-R performs a dimension reduction by shrinking the $X$ matrix. Hence, PLS-R is a suitable tool for problems for which the data contains many more variables $p$ than observations $n$.

The objective function in (2.1) can be interpreted as maximizing the squared covariance between $Y$ and the latent component: $corr^2(Y, X_{k-1}r_k)\text{var}(X_{k-1}r_k)$. Because the response $Y$ has been taken into account to formulate the latent matrix, PLS usually has better performance in prediction problems than principal component analysis (PCA) does. This is one of the main differences between PLS-R and PCA [15].

2.2. Multivariate Response

Similarly to univariate response PLS-R, multivariate response PLS-R selects latent components in $R^p$ and $R^q$, i.e., $t_k$ and $v_k$, such that the covariance between $t_k$ and $v_k$ is maximized. For a specific component, the sets of weights $r_k \in R^p$ and $c_k \in R^q$ are obtained by solving

$$\max \{ t^\prime v \} = \max \{ r^\prime X_{(k-1)}Y_{(k-1)}c \}$$

$$\text{s.t. } r^\prime r = c^\prime c = 1 \quad (2.2)$$

where $t_k = X_{(k-1)}r_k$, $v_k = Y_{(k-1)}c_k$, and $X_{(k-1)}$ and $Y_{(k-1)}$ are the deflated matrices associated with $X$ and $Y$. Notice that the optimal solution $c_k$ should be proportional to $Y^\prime_{(k-1)}X_{k-1}r_k$. Therefore, the optimization in (2.2) is equivalent to

$$\max \{ r^\prime X_{(k-1)}Y_{(k-1)}Y^\prime_{(k-1)}X_{k-1}r \}$$

$$\text{s.t. } r^\prime r = 1. \quad (2.3)$$

For each component, the solution to this criterion can be obtained by using a so called PLS2 algorithm. A detailed description of the iterative algorithm as presented by Höskuldsson [16] is in Algorithm 1.

In 1993 de Jong proposed a variant of the PLS2 algorithm, called Statistically Inspired Modification of PLS (SIMPLS), which calculates the PLS latent components directly as linear combinations of the original variables [10]. The SIMPLS was first developed as the solution to an optimization problem

$$w_k = \arg \max_w (w^\prime XYY^\prime Xw) \quad (2.4)$$

$$\text{s.t. } w^\prime w = 1, \quad w^\prime Xw_j = 0 \text{ for } j = 1, \ldots, k-1.$$

Ter Braak and de Jong [17] provided a detailed comparison between the objective functions for PLS2 in (2.3) and SIMPLS in (2.4) and showed that the successive weight vectors $w_k$ can be derived either from the deflated data matrices or the original variables in PLS2 and SIMPLS respectively. Let $W^+$ be the Moore-Penrose inverse of $W = [w_1 \ w_2 \ \ldots \ w_{k-1}]$. The PLS2 algorithm (Algorithm 1) is equivalent to solving the optimization

$$w_k = \arg \max_w (w^\prime XYY^\prime Xw)$$
Algorithm 1: PLS2 algorithm

for k = 1 : K do
  initialize r
  X = \textit{X}_\text{new}
  Y = \textit{Y}_\text{new}
  while solution has not converged do
    t = Xr
    c = Y'r
    Scale c to length 1
    v = Yc
    r = X'v
    Scale r to length 1
    loading vector \( \mathbf{p} = X't/(t't) \)
    deflate \( X_{\text{new}} = X - \mathbf{tp}' \)
    regression \( \mathbf{b} = Y't/(t't) \)
    deflate \( Y_{\text{new}} = Y - \mathbf{tb}' \)
    \( r_k = r \)
  end
end

s.t. \( \mathbf{w}'(I - WW^+)\mathbf{w} = 1, \mathbf{w}'X'X\mathbf{w}_i = 0 \) for \( i = 1, ..., k - 1 \).

Both NIPALS and SIMPLS have the same objective function but each is maximized under a different normalization constraint. NIPALS and SIMPLS are equivalent when Y is univariate, but provide slightly different weight vectors in multivariate scenarios. The performance depends on the nature of the data, but SIMPLS appears easier to interpret since it does not involve deflation of the data set [10]. We develop our globally sparse PLS-R based on the SIMPLS optimization formulation.

3. Sparse Partial Least Squares Regression

3.1. \( \ell_1 \) Penalized Sparse PLS Regression

One approach to sparse PLS-R is to add the \( \ell_1 \) norm of the weight vector, a sparsity inducing penalty, to (2.4). The solution for the first component would be obtained by solving

\[
\mathbf{w}_1 = \arg \max_{\mathbf{w}} (\mathbf{w}'X'Y'X\mathbf{w})
\]

s.t. \( \mathbf{w}'\mathbf{w} = 1, ||\mathbf{w}||_1 \leq \lambda. \)

The addition of the \( \ell_1 \) norm is similar to SCOTLASS (simplified component lasso technique), the sparse PCA proposed by Jolliffe [18]. However, the solution of SCOTLASS is not sufficiently sparse, and the same issue remains in (3.1). Chun and Keles [9] reformulated the problem, promoting the exact zero property by imposing the \( \ell_1 \) penalty on a surrogate of the weight vector instead of the original weight vector [9]. For the first component, they solve the following optimization by alternating between updating \( \mathbf{w} \) and updating \( \mathbf{z} \) (block
coordinate descent).
\[
\min_{w, z} \{ -\kappa w'X'Y'Xw + (1-\kappa)(z-w)'X'Y'X(z-w) + \lambda_1 ||z||_1 + \lambda_2 ||z||_2^2 \}
\]
\[
s.t. \quad w'w = 1
\]
Allen et al. proposed a general framework for regularized PLS-R [20].
\[
\max_{w, v} w'Mv - \lambda P(w) \quad s.t. \quad w'w \leq 1, \quad v'v = 1
\]
in which \( M \) is the cross-product matrix \( X'Y \), and the regularization function \( P \) is a convex penalty function. The formulation is a relaxation of SIMPLS with penalties being applied to the weight vectors, and can be viewed as a generalization of [9].

As mentioned in the Introduction, these formulations ([9, 20]) penalize the variables in each PLS component independently. This paper proposes an alternative in which variables are penalized simultaneously over all components. First, we define the global weight matrix, consisting of the \( K \) weight vectors, as
\[
W = \begin{bmatrix}
w_1 & w_2 & \cdots & w_K \\
\end{bmatrix} = \begin{bmatrix}
w'(1) & - \\
w'(2) & - \\
\vdots & \\
w'(p) & - \\
\end{bmatrix}
\]
Notice that the elements in a particular row of \( W \), i.e., \( w'_j \), are all associated with the same predictor variable \( x_j \). Therefore, rows of zeros correspond to variables that are not selected. To illustrate the drawbacks of penalizing each variable in each component independently, as in [9], suppose that each entry in \( W \) is selected independently with probability \( p_1 \). The probability that the \((j)_{th}\) variable is not selected becomes \((1 - p_1)^K\), and the probability that all the variables are selected by at least one weight vector is \([1 - (1 - p_1)^K]^p\), which increases as the number of weight vectors \( K \) increases. This suggests that for large \( K \) the local variable selection approach of [9] may not lead to an overall sparse and parsimonious PLS-R model. In such cases a group sparsity constraint can be employed to limit the overall number of selected variables.

### 3.2. Jointly Sparse Global SIMPLS Regression

The Jointly Sparse Global SIMPLS Regression variable selection problem is to find the top \( K \) weight vectors that best relate \( X \) to \( Y \), while using limited number of variables. This is a subset selection problem that is equivalent to adding a constraint on the \( \ell_0 \) norm of the vector consisting of the norms of the rows of \( W \), i.e., the number of nonzero rows in \( W \). For concreteness we use the \( \ell_2 \) norm for the rows. This leads to the optimization problem

\[
\min_{W} \{ -\kappa w'X'Y'Xw + (1-\kappa)(z-w)'X'Y'X(z-w) + \lambda_1 ||z||_1 + \lambda_2 ||z||_2^2 \}
\]
\[
s.t. \quad w'w = 1
\]
\[
\max_{w, v} w'Mv - \lambda P(w) \quad s.t. \quad w'w \leq 1, \quad v'v = 1
\]
W = \arg \min_{W} -\frac{1}{n^2} \sum_{k=1}^{K} w_k' X' Y Y' X w_k \tag{3.2}
\text{s.t.} \|\varpi\|_0 \leq t, \ w_k' w_k = 1 \ \forall \ k, \ \text{and} \ w_k' X' X w_i = 0 \ \forall \ i \neq k

in which
\varpi = \begin{bmatrix}
|w_{(1)}|_2 \\
|w_{(2)}|_2 \\
\vdots \\
|w_{(p)}|_2
\end{bmatrix}.

The objective function (3.2), which we refer to as global SIMPLS, is the sum of the objective functions (2.4) in the first $K$ iterations of SIMPLS. Instead of the sequential greedy solution in PLS2 algorithm, the proposed jointly sparse global SIMPLS Regression solves for the $K$ weight vectors simultaneously. We introduce the $\frac{1}{n^2}$ factor to the objective function to interpret it as an empirical covariance. Given the complexity of this combinatorial problem, as in standard optimization practice, we relax the $\ell_0$ norm optimization to a mixed norm structured sparsity penalty \cite{21}.

\begin{align*}
W = \arg \min_{W} -\frac{1}{n^2} \sum_{k=1}^{K} w_k' X' Y Y' X w_k + \lambda \sum_{j=1}^{p} |w_{(j)}|_2 \tag{3.3}
\text{s.t.} \ w_k' w_k = 1 \ \forall \ k, \ \text{and} \ w_k' X' X w_i = 0 \ \forall \ i \neq k
\end{align*}

The $\ell_2$ norm of each row of $W$ promotes grouping entries in $W$ that relate to the same predictor variable, whereas the $\ell_1$ norm promotes a small number of groups, as in (3.1).

4. Algorithmic Implementation for Jointly Sparse Global SIMPLS Regression

Constrained eigen-decomposition and group variable selection are each well-studied problems for which efficient algorithms have been developed. We propose to solve the optimization (3.3) by augmented Lagrangian methods, which allows one to solve (3.3) by variable splitting iterations. Augmented Lagrangian methods introduce a new variable $M$, constrained such that $M = W$, such that the row vectors $m_{(j)}$ of $M$ obey the same structural pattern as the rows of $W$: \begin{align*}
\min_{W,M} -\frac{1}{n^2} \sum_{k=1}^{K} w_k' X' Y Y' X w_k + \lambda \sum_{j=1}^{p} |m_{(j)}|_2 \\
\text{s.t.} \ w_k' w_k = 1 \ \forall \ k, \ w_k' X' X w_i = 0 \ \forall \ i \neq k, \ \text{and} \ M = W \tag{4.1}
\end{align*}
The optimization (4.1) can be solved by replacing the constrained problem by an unconstrained one with an additional penalty on the Frobenius norm of the
difference \(M - W\). This penalized optimization can be iteratively solved by an alternating direction method of multipliers (ADMM) algorithm [25, 26, 27, 28, 29, 30, 31, 32], a block coordinate descent method that alternates between optimizing over \(W\) and over \(M\) (See Algorithm 2). We initialize Algorithm 2 with \(M^{(0)}\) equal to the solution of SIMPLS, and \(D^{(0)}\) equal to the zero matrix. Setting the parameter \(\mu\) is nontrivial [31], and is hand-tuned for fastest convergence in some applications [27]. Once the algorithm converges, the final PLS regression coefficients are obtained by applying SIMPLS regression on the selected variables keeping the same number of components \(K\). The optimization over \(W\) can be further simplified to a secular equation problem, whereas the optimization over \(M\) can be shown to reduce to a soft thresholding operation. The algorithm iterates until the stopping criterion based on the norm of the residuals \(\|W^{(\tau)} - M^{(\tau)}\|_F < \epsilon\) is satisfied, for some given tolerance \(\epsilon\). As described later in the experimental comparisons section, the parameters \(\lambda\) and \(K\) are decided by cross validation.

**Algorithm 2:** Algorithm for solving the global SIMPLS with global variable selection problem using the augmented Lagrangian method.

```plaintext
set \(\tau = 0\), choose \(\mu > 0\), \(M^{(0)}, D^{(0)}\);
while stopping criterion is not satisfied do
    \(W^{(\tau+1)} = \arg\min\limits_W -\frac{1}{n^2} \sum_{k=1}^K w_k'XYY'Xw_k + \frac{\mu}{2} \|W - M^{(\tau)} - D^{(\tau)}\|_F^2\)
    s.t. \(w_k'w_k = 1 \forall k, w_k'X'Xw_i = 0 \forall i \neq k\);
    \(M^{(\tau+1)} = \arg\min\limits_M \lambda \sum_{j=1}^p \|m_{j(j)}\|_2 + \frac{\mu}{2} \|W^{(\tau+1)} - M - D^{(\tau)}\|_F^2\);
    \(D^{(\tau+1)} = D^{(\tau)} - W^{(\tau+1)} + M^{(\tau+1)}\);
    \(\tau = \tau + 1\);
```

**Optimization over \(W\)** The following optimization in Algorithm 2 is a non-convex quadratically constrained quadratic program (QCQP).

\[
W^{(\tau+1)} = \arg\min_{w_k} -\frac{1}{n^2} \sum_{k=1}^K w_k'XYY'Xw_k + \frac{\mu}{2} \|W - M^{(\tau)} - D^{(\tau)}\|_F^2
\]

s.t. \(w_k'w_k = 1 \forall k, w_k'X'Xw_i = 0 \forall i \neq k\)

We propose solving for the \(K\) vectors in \(W\) successively by a greedy approach. Let \(m_k\) and \(d_k\) be the columns of the matrices \(M\) and \(D\), and \(\omega_k = m_k + d_k\). The optimization over \(W\) becomes

\[
w_k^{(\tau+1)} = \arg\min_{w} -\frac{1}{n^2} w'XYY'Xw + \frac{\mu}{2} \|w - \omega_k\|_2^2
\]

s.t. \(w'w = 1, w'X'Xw_i = 0 \forall i < k\).  \(4.2\)
Lemma 4.1. Let $N$ be an orthonormal basis for the orthogonal complement of \{X'Xw_i\}, $i < k$. The optimization (4.2) can be solved by the method of Lagrange multipliers. The solution is $w_k = N(A - \alpha I)^{-1}b$, in which $A = \frac{1}{n^2}N'X'YY'XN$, $b = \frac{\mu_2}{2}N'w_k$ and $\alpha$ is the minimum solution that satisfies $b'(A - \alpha I)^{-2}b = 1$.

Proof. Let $w = N\tilde{w}$, then the optimization (4.2) can be written as

$$\min \tilde{w}'A\tilde{w} - 2b'\tilde{w} \text{ s.t. } \tilde{w}'\tilde{w} = 1.$$ 

Since we assume that $w$ is a linear combination of the basis vectors in $N$, the orthogonality conditions in (4.2) are automatically satisfied. Hence these conditions have been dropped in the new formulation. Then using Lagrange multipliers, we can show that the solution takes the form as stated above.

Suppose there are two solutions of $\alpha$ that satisfy $b'(A - \alpha I)^{-2}b = 1$, corresponding to two pairs of solutions to the optimization, $(\tilde{w}_1, \alpha_1)$ and $(\tilde{w}_2, \alpha_2)$. Since $\tilde{w} = (A - \alpha I)^{-1}b$,

$$A\tilde{w}_1 = \alpha_1\tilde{w}_1 + b \quad (4.3)$$

$$A\tilde{w}_2 = \alpha_2\tilde{w}_2 + b. \quad (4.4)$$

By multiplying (4.3) by $\tilde{w}_1$, and (4.4) by $\tilde{w}_2$, then subtracting the two new equations, we have

$$(\tilde{w}_1' - \tilde{w}_2')A\tilde{w}_1 - \tilde{w}_2'A\tilde{w}_2 = \alpha_1 - \alpha_2 + (\tilde{w}_1' - \tilde{w}_2')b. \quad (4.5)$$

On the other hand, by multiplying (4.3) by $\tilde{w}_2$, and (4.4) by $\tilde{w}_1$, and subtracting the two new equations, we have

$$(\tilde{w}_1' - \tilde{w}_2')b = (\alpha_1 - \alpha_2)(\tilde{w}_1'\tilde{w}_2) \quad (4.6)$$

Given (4.5) and (4.6), it can be shown that

$$(\tilde{w}_1' - \tilde{w}_2')(\tilde{w}_1 - \tilde{w}_2) = (\alpha_1 - \alpha_2)(\tilde{w}_1'\tilde{w}_2) = \frac{\alpha_1 - \alpha_2}{2}||\tilde{w}_1 - \tilde{w}_2||^2.$$ 

Hence, one should select the minimum among all the feasible $\alpha$’s. $\square$

The equation $b'(A - \alpha I)^{-2}b = 1$ is a secular equation, a well studied problem in constrained eigenvalue decomposition [23, 24]. The more general problem of least squares with a quadratic constraint was discussed in [22]. We can diagonalize the matrix $A$ as $A = UDU'$, in which $D$ is diagonal with eigenvalues $d_1, d_2, ..., d_p$ in decreasing order on the diagonal, and the columns of $U$ are the corresponding eigenvectors. Define

$$g(\alpha) = b'(A - \alpha I)^{-2}b = b'U \begin{bmatrix} \frac{1}{(d_1 - \alpha)^2} & \\ \frac{1}{(d_2 - \alpha)^2} & \\ \ddots & \\ \frac{1}{(d_p - \alpha)^2} \end{bmatrix} U' \begin{bmatrix} b \end{bmatrix}.$$
Let $\tilde{b} = U'b$, then $g(\alpha) = b'(A - \alpha I)^{-2}b = \sum_i \frac{\tilde{b}_i^2}{(d_i - \alpha)^2}$, and hence $g(\alpha) = 1$ is a secular equation. $g(\alpha)$ increases strictly as $\alpha$ increases from $-\infty$ to $d_p$, since

$$g'(\alpha) = \sum_i \frac{2\tilde{b}_i^2}{(d_i - \alpha)^3}$$

is positive for $-\infty < \alpha < d_p$. Moreover, given the limits

$$\lim_{\alpha \to -\infty} g(\alpha) = 0 \quad \text{and} \quad \lim_{\alpha \to d_p} g(\alpha) = \infty$$

we can conclude that there is exactly one solution $\alpha < d_p$ to the equation $g(\alpha) = 1$, [23]. An iterative algorithm (Algorithm 3) is used to solve $g(\alpha) = 1$ starting from a point to the left of the smallest eigenvalue $d_p$ [24].

Notice that calculating $g(\alpha) = b'(A - \alpha I)^{-2}b$ involves inverting a $(p-k+1) \times (p-k+1)$ matrix, but $A$ has rank at most $q$. We can reduce the computational burden by the use of Woodbury matrix identity,

$$(A - \alpha I)^{-1} = -\frac{1}{\alpha} (I - \frac{1}{\alpha n^2} N'X'Y(I + \frac{1}{\alpha n^2} Y'XNN'X'Y)^{-1}Y'XN).$$

The new format only requires inverting a $q \times q$ matrix, and in most applications, the number of responses $q$ is much less than the number of predictors $p$. Furthermore, $N$ is involved in $g(\alpha)$ in the form of $NN' = I - HH'$, in which $H$ is an orthonormal basis for $\{X'Xw_i\}, i < k$. $H$ can be constructed by Gram-Schmidt process as the algorithm successively finds the weight vectors $w_i's$.

**Algorithm 3:** Iteration for solving secular equation.

```plaintext
set $\tau = 0$, choose $\alpha_0 = d_p - \varepsilon_1$;
while stopping criterion is not satisfied, $|g(\alpha_\tau) - 1| > \varepsilon_2$ do
    $\alpha_{\tau+1} = \alpha_\tau + 2 \frac{g^{-1/2}(\alpha_\tau) - 1}{g^{1/2}(\alpha_\tau) g'(\alpha_\tau)}$;
```

**Optimization over $M$** The optimization over $M$ has a closed form solution. Let $\Delta = W^{(\tau+1)} - D^{(\tau)}$, and $\delta_{(j)}$ denote the $j_{th}$ row of $\Delta$, then each row of $M$ is given as $m_{(j)} = [||\delta_{(j)}|| - \frac{1}{\mu}]_+ \frac{\delta_{(j)}}{||\delta_{(j)}||}$, in which $[z]_+ = max\{z, 0\}$.

Convergence analysis of ADMM can be found in [25, 29, 30, 32]. In particular, it has been shown that ADMM converges linearly for strongly convex objective functions [30]. Although the convergence is based on strong convexity assumptions, ADMM has been widely applied in practice [26, 27], even to nonconvex problems [28]. The proposed Jointly Sparse Global SIMPLS Regression is one example of these nonconvex applications.


5. Simulation Experiments

We implement the simulation models in [33, 9]. There are four models all following $Y = X\beta + F$, in which the number of observation is $n = 100$, and the dimension is $p = 5000$. Full details of the models are given in Table 1. We compare five different methods: the standard PLS regression (denoted as PLS in the following comparison tables), PLS generalized linear regression proposed by Bastien et al. [34], $\ell_1$ penalized PLS regression [9] (denoted as $\ell_1$ SPLS), Lasso [35] and the Jointly Sparse Global SIMPLS regression (denoted as $\ell_1/\ell_2$ SPLS). All the methods select their parameters by ten fold cross-validation, except for the PLS generalized linear regression, which stops including an additional component if the new component is not significant. The parameter $\mu$ in the Jointly Sparse Global SIMPLS-R is fixed to 2000, and updated in each iteration by a scaling factor $1.01\mu$. The experiments on real data in Section 6 and 7 are using the same setting of $\mu$. Two i.i.d sets are generated for each trial: one as the training set and one as the test set. Ten trials are conducted for each model, and the averaged results are listed in Table 2.

In most of the simulations, we observe that the proposed Jointly Sparse Global SIMPLS-R performs as good or better than other methods in terms of the prediction MSE. In particular, the number of variables and the number of components chosen in Jointly Sparse Global SIMPLS-R are usually less than the $\ell_1$ penalized PLS-R. We also calculate the $R^2$ for each method on the training data to measure the variation explained. The standard PLS regression and the PLS generalized linear regression proposed by Bastien et al. [34] both have $R^2$ close to 1, but the performance in terms of MSE is not ideal in the first three models for these methods. This may suggest that these models overfit the data. In addition to the averaged performance, the p-values of one sided paired t-test suggest that Jointly Sparse Global SIMPLS-R reduces model complexity significantly from those in the standard PLS regression and the PLS generalized linear regression. Lasso achieves low complexity in terms of the number of variables, but the MSE is high compared to Jointly Sparse Global SIMPLS-R and the $\ell_1$ penalized PLS-R. The cross validation time for Jointly Sparse Global SIMPLS Regression is long, searching over a two-dimensional grid of the number of components $K$ and the regularization parameter $\lambda$ to minimize MSE. However, the performance in terms of prediction MSE improves, and the model complexity in terms of the number of variables and the number of components both decreases compared with other methods in most simulations.
to denote the indicator function. We use $I_i$ to index the $n$ samples, $j$ (1 ≤ $j$ ≤ $p$) to index the $p$ variables, and $k$ to index the hidden components. Simulation models in [33, 9]: These models were originally proposed to test supervised principal components. Model 1 and Model 2 are designed such that one latent variable dominates the multi-collinearity. Model 3 has multiple latent variables, and Model 4 has a correlation structure from an autoregressive process. In these model descriptions, $I_i$ is used to denote the indicator function. We use $i$ (1 ≤ $i$ ≤ $n$) to index the $n$ samples, $j$ (1 ≤ $j$ ≤ $p$) to index the $p$ variables, and $k$ to index the hidden components.

**Model 1**

$H_1(i) = 3I(i \leq 50) + 4I(i > 50), \ 1 \leq i \leq n$

$H_2(i) = 3.5$

$X_j = H_k + \varepsilon_j, \ \text{for} \ \ 1 \leq j \leq 50, \ \ k = 1, 2, \ (p_0, ..., p_2) = (0, 50, p)$

$\beta_j = \frac{1}{n} \ 1 \leq j \leq 50$

$\varepsilon_j$ is $N(0, I_n)$ distributed, and $F$ is $N(0, 1.5^2 I_n)$ distributed.

**Model 2**

$H_1(i) = 2.5I(i \leq 50) + 4I(i > 50)$

$H_2(i) = 3.5 + 1.5I(u_{1i} \leq 0.4)$

$H_3(i) = 3.5 + 0.5I(u_{2i} \leq 0.7)$

$H_4(i) = 3.5 - 1.5I(u_{3i} \leq 0.3)$

$H_5(i) = 3.5$

$u_{1i}, u_{2i}, u_{3i}$ are i.i.d from $Unif(0, 1)$

$X_j = H_k + \varepsilon_j, \ \text{for} \ \ 1 \leq j \leq 50, \ \ k = 1, ..., 5, \ (p_0, ..., p_5) = (0, 50, 100, 200, 300, p)$

$\beta_j = \frac{1}{n} \ 1 \leq j \leq 50$

$\varepsilon_j$ is $N(0, I_n)$ distributed, and $F$ is $N(0, I_n)$ distributed.

**Model 3**

$H_1(i) = 2.5I(i \leq 50) + 4I(i > 50)$

$H_2(i) = 2.5I(1 \leq i \leq 25, \ or \ 51 \leq i \leq 75) + 4I(26 \leq i \leq 50, \ or \ 76 \leq i \leq 100)$

$H_3(i) = 3.5 + 1.5I(u_{1i} \leq 0.4)$

$H_4(i) = 3.5 + 0.5I(u_{2i} \leq 0.7)$

$H_5(i) = 3.5 - 1.5I(u_{3i} \leq 0.3)$

$H_6(i) = 3.5$

$u_{1i}, u_{2i}, u_{3i}$ are i.i.d from $Unif(0, 1)$

$X_j = H_k + \varepsilon_j, \ \text{for} \ \ 1 \leq j \leq 50, \ \ k = 1, ..., 5, \ (p_0, ..., p_5) = (0, 25, 50, 100, 200, 300, p)$

$\beta_j = \frac{1}{n} \ 1 \leq j \leq 50$

$\varepsilon_j$ is $N(0, I_n)$ distributed, and $F$ is $N(0, I_n)$ distributed.

**Model 4**

$H_1(i) = I(i \leq 50) + 6I(i > 50)$

$H_2(i) = 3.5 + 1.5I(u_{1i} \leq 0.4)$

$H_3(i) = 3.5 + 0.5I(u_{2i} \leq 0.7)$

$H_4(i) = 3.5 - 1.5I(u_{3i} \leq 0.3)$

$H_5(i) = 3.5$

$u_{1i}, u_{2i}, u_{3i}$ are i.i.d from $Unif(0, 1)$

$X = (X^{(1)}, X^{(2)})$

$X^{(1)}$ is generated from $N(0, \Sigma_{50 \times 50})$, $\Sigma$ is from AR(1) with $\rho = 0.9$.

$X^{(2)} = H_k + \varepsilon_j, \ \text{for} \ \ 1 \leq i \leq 50, \ k = 1, ..., 5, \ (p_0, ..., p_5) = (0, 50, 100, 200, 300, p)$

$\beta_j = \frac{1}{n} \ 1 \leq j \leq 50, \ \ m = 1, ..., 6$, where

$\varepsilon_j$ is $N(0, I_n)$ distributed, and $F$ is $N(0, 1.5^2 I_n)$ distributed.
Table 2
Performance comparison table for the 4 simulation models. We compare five different methods: the standard PLS regression (PLS), PLS generalized linear regression proposed by Bastien et al. (Bastien), $\ell_1$ penalized PLS regression ($\ell_1$ SPLS), Lasso and the Jointly Sparse Global SIMPLS regression ($\ell_1/\ell_2$ SPLS). These results are obtained by averaging over 10 trails.

| Model 1 | 1. PLS | 2. Bastien | 3. $\ell_1$ SPLS | 4. Lasso | 5. $\ell_1/\ell_2$ SPLS | p-values of one sided paired t-test |
|--------|--------|------------|-----------------|---------|-----------------|----------------------------------|
| number of comp. | 1.4 | 5 | 1.9 | NA | 1.4 | 0.5 | 4.49 x 10^{-7} | 0.19 | NA |
| number of variables | 5000 | 1129.4 | 246.5 | 40.7 | 276.1 | 1.84 x 10^{-11} | 5.46 x 10^{-5} | 0.35 | 0.053 |
| MSE | 3.14 | 2.98 | 3.00 | 3.23 | 2.82 | 0.01 | 0.013 | 0.075 | 0.006 |
| $R^2$ | 0.98 | 1 | 0.71 | 0.59 | 0.83 | | | |
| Time CV | 104.17 | 0 | 43.49 | 51.59 | 11414 | |
| Time analysis | 0.89 | 121.91 | 0.05 | 0.04 | 6.49 | |
| Time prediction | 0.010 | 0.011 | 0.002 | 0.04 | 0.002 | |
| Total time | 102.37 | 121.92 | 43.54 | 51.67 | 11420 | |

| Model 2 | 1. PLS | 2. Bastien | 3. $\ell_1$ SPLS | 4. Lasso | 5. $\ell_1/\ell_2$ SPLS | p-values of one sided paired t-test |
|--------|--------|------------|-----------------|---------|-----------------|----------------------------------|
| number of comp. | 2 | 5 | 2.3 | NA | 1.1 | 0.9671 | 1.19 x 10^{-11} | 0.0184 | NA |
| number of variables | 5000 | 1158.4 | 273.4 | 15.8 | 171.7 | 1.13 x 10^{-14} | 6.91 x 10^{-9} | 0.1041 | 0.019 |
| MSE | 3.18 | 2.99 | 2.93 | 3.09 | 2.69 | 8.69 x 10^{-4} | 0.0046 | 0.0344 | 2.70 x 10^{-4} |
| $R^2$ | 0.98 | 1 | 0.79 | 0.39 | 0.75 | | | |
| Time CV | 100.51 | 0 | 43.03 | 53.97 | 11420 | |
| Time analysis | 1.28 | 122.69 | 0.06 | 0.04 | 5.72 | |
| Time prediction | 0.010 | 0.011 | 0.002 | 0.039 | 0.001 | |
| Total time | 101.80 | 122.70 | 43.09 | 54.05 | 11426 | |

| Model 3 | 1. PLS | 2. Bastien | 3. $\ell_1$ SPLS | 4. Lasso | 5. $\ell_1/\ell_2$ SPLS | p-values of one sided paired t-test |
|--------|--------|------------|-----------------|---------|-----------------|----------------------------------|
| number of comp. | 1.4 | 5 | 1.4 | NA | 1.5 | 0.4201 | 5.53 x 10^{-6} | 0.3632 | NA |
| number of variables | 5000 | 1156.4 | 89.2 | 41.3 | 60.5 | 2.32 x 10^{-19} | 1.39 x 10^{-12} | 0.1697 | 0.1726 |
| MSE | 1.82 | 1.48 | 1.27 | 1.48 | 1.25 | 1.18 x 10^{-4} | 0.004 | 0.3430 | 0.0054 |
| $R^2$ | 0.98 | 1 | 0.77 | 0.75 | 0.73 | | | |
| Time CV | 102.91 | 0 | 43.84 | 49.15 | 13295 | |
| Time analysis | 1.03 | 126.08 | 0.04 | 0.04 | 5.48 | |
| Time prediction | 0.010 | 0.011 | 0.001 | 0.039 | 0.001 | |
| Total time | 103.65 | 126.09 | 43.88 | 49.53 | 13300 | |

| Model 4 | 1. PLS | 2. Bastien | 3. $\ell_1$ SPLS | 4. Lasso | 5. $\ell_1/\ell_2$ SPLS | p-values of one sided paired t-test |
|--------|--------|------------|-----------------|---------|-----------------|----------------------------------|
| number of comp. | 2 | 5 | 2.6 | NA | 2.1 | 0.4057 | 6.82 x 10^{-5} | 0.2201 | NA |
| number of variables | 5000 | 1118.8 | 1260.8 | 9.4 | 1180.5 | 9.62 x 10^{-5} | 0.4618 | 0.3918 | 0.0485 |
| MSE | 2.15 | 2.29 | 2.41 | 2.14 | 2.36 | 0.0087 | 0.1874 | 0.3812 | 0.0056 |
| $R^2$ | 0.98 | 1 | 0.78 | 0.19 | 0.91 | | | |
| Time CV | 98.16 | 0 | 44.31 | 50.52 | 12051 | |
| Time analysis | 1.55 | 123.79 | 0.10 | 0.10 | 7.97 | |
| Time prediction | 0.010 | 0.011 | 0.007 | 0.042 | 0.004 | |
| Total time | 99.73 | 123.8 | 44.41 | 50.60 | 12059 | |
Table 3

Performance comparison table for the Octane data.

|                | 1. PLS | $\ell_1$ SPLS | $\ell_1/\ell_2$ SPLS | p-values of one sided paired t-test |
|----------------|--------|----------------|---------------------|-----------------------------------|
| number of comp. | 5.5    | 4.5            | 3.8                 | $5.888 \times 10^{-14}$            |
| number of var.  | 225    | 87.3           | 38.5                | $1.1176 \times 10^{-12}$           |
| MSE            | 0.0564 | 0.0569         | 0.0481              | $0.0032$                          |

6. Application 1: Chemometrics Study

In this section we show experimental results obtained by comparing standard PLS-R, $\ell_1$ penalized PLS-R [9] (denoted as $\ell_1$ SPLS in the performance table), and our proposed Jointly Sparse Global SIMPLS-R (denoted as $\ell_1/\ell_2$ SPLS in the performance table). All the methods have been applied on the Octane data set (see [13]). The Octane data is a real data set consisting of 39 gasoline samples for which the digitized Octane spectra have been recorded at 225 wavelengths (in nm). The aim is to predict the Octane number, a key measurement of the physical properties of gasoline, using the spectra as predictors. This is of major interest in real applications, because the conventional procedure to calculate the Octane number is time consuming and involves expensive and maintenance-intensive equipment as well as skilled labor.

The experiments are composed of 150 trials. In each trial we randomly split the 39 samples into 26 training samples and 13 test samples. The regularization parameter $\lambda$ and number of components $K$ are selected by 2-fold cross validation on the training set. The averaged results over the 150 trials are shown in Table 3. We further show the variable selection frequencies for the sparse PLS methods over the 150 trials superimposed on the octane data in Fig. 1 (B) and (C). In chemometrics, the rule of thumb is to look for variables that have large amplitudes in first derivatives with respect to wavelength. Notice that both $\ell_1$ penalized PLS-R and Jointly Sparse Global SIMPLS-R have selected variables around 1200 and 1350 nm, and the selected region in the latter case is more confined. Box and Whisker plots for comparing the MSE, number of selected variables, and number of components of these three PLS formulations are shown in Fig. 1 (A). Comparing our proposed Jointly Sparse Global SIMPLS Regression with standard PLS-R and $\ell_1$ penalized PLS-R [9], we show that Jointly Sparse Global SIMPLS-R attains better performance in terms of MSE, the number of predictors, and the number of components. Besides, the model complexity in Jointly Sparse Global SIMPLS-R is significantly lower than both standard PLS-R and $\ell_1$ penalized PLS-R, given the p-values of one sided paired t-test.
Fig 1: (A) Box and Whisker plots for comparing the MSE, number of selected variables, and number of components of three PLS formulations: standard PLS-R, $\ell_1$ penalized PLS-R ($\ell_1$ SPLS), our proposed Jointly Sparse Global SIMPLS-R ($\ell_1/\ell_2$ SPLS). (B) Variable selection frequency of $\ell_1$ SPLS superimposed on the octane data and its first derivative: The height of the surfaces represents the exact value of the data over 225 variables for the 39 samples. The color of the surface shows the selection frequency of the variables as depicted on the colorbar. (C) Variable selection frequency of $\ell_1/\ell_2$ SPLS superimposed on the octane data and its first derivative.
7. Application 2: Sparse Prediction of Disease Symptoms from Gene Expression

In this section we apply the Jointly Sparse Global SIMPLS Regression to 4 types of predictive health challenge studies involving the H3N2, the H1N1, the HRV, and the RSV viruses. In these challenge studies, publicly available from the NCBI-GEO website, serial peripheral blood samples were acquired from a population of subjects inoculated with live flu viruses [38, 39, 40, 41]. The prediction task in these experiments is to predict the symptom scores based on gene expression of 12023 genes. There were 10 symptom scores, i.e., runny nose, stuffy nose, sneezing, sore throat, earache, malaise, cough, shortness of breath, headache, and myalgia, documented over time. The symptoms are self-reported scores, ranging from 0 to 3. We linearly interpolate the gene expressions to match them with the sampling time of the symptom reports. We compare the Jointly Sparse Global SIMPLS-R with standard PLS-R and ℓ₁ penalized PLS-R by leaving one subject out as the test set, and the rest as the training set. The process is repeated until all subjects have been treated as the test set. The number of components for all methods and the regularization parameter in Jointly Sparse Global SIMPLS-R are selected by 2-fold cross validation to minimize the sum of the MSE of the responses. Since each subject has multiple samples, we perform the cross validation by splitting by subjects, i.e., no samples from the same subject will appear in both training and tuning sets. We restrict the responses to the first 3 symptoms, which are the upper respiratory symptoms, and the results are shown in Table 4. In most of the cases, the proposed Jointly Sparse Global SIMPLS-R method outperforms the standard PLS-R and ℓ₁ penalized PLS-R in terms of prediction MSE, number of components, number of genes. As can be seen in Table 4, the number of selected variables decreases significantly by applying the ℓ₁/ℓ₂ mixed norm sparsity penalty to the PLS-R objective function. Thus the proposed PLS-R method is able to construct a more parsimonious predictor relative to the other PLS-R methods having similar accuracy.

The PLS-R method can also be viewed as an exploratory data analysis tool for constructing low dimensional descriptors of the independent variables and response variables. Specifically, the general underlying matrix factorization model \( X = TP' + E \) and \( Y = TQ' + F \), with latent component \( T = XW \), provides a factor analysis model for the independent and response variables \( X \) and \( Y \). \( T \), \( P \) and \( Q \) can be interpreted in a similar manner as the singular vectors of PCA. However, different from PCA that does not account for the response variables, \( T \), \( P \) and \( Q \) contain information about both the independent variables and the response variables. The factor analysis interpretation of the underlying PLS model is that \( T \) is a latent score matrix and \( P \), \( Q \) are latent factor loading matrices that associate \( T \) with the independent variables and the response variables via the approximate matrix factorizations \( X \approx TP' \) and \( Y \approx TQ' \), respectively. The correlations between the latent component \( T \) and the sum of the 3 upper respiratory symptoms are reported in Table 5, which also shows results for classic matrix factorization methods including non-negative matrix
Table 4
Performance comparison table for the predictive health study. We apply the Jointly Sparse Global SIMPLS Regression ($\ell_1/\ell_2$ SPLS) for sparse prediction of disease symptoms from gene expression. The performance is compared with standard PLS and $\ell_1$ sparse PLS. $\ell_1/\ell_2$ SPLS achieves lower MSE with significantly fewer variables in most of the studies.

|         | 1. PLS | 2. $\ell_1$ SPLS | 3. $\ell_1/\ell_2$ SPLS (3,1) | (3,2) |
|---------|--------|------------------|-----------------------------|-------|
| **H1N1** |        |                  |                             |       |
| number of comp. | 2.8    | 2.3              | 2.4                        | 0.1163 | 0.3322 |
| number of genes | 12023  | 3624.1           | 3575.8                     | 1.4451 x 10^{-10} | 0.4842 |
| Overall MSE | 0.599  | 0.603            | 0.591                      | 0.2890 | 0.1094 |
| Runny nose MSE | 0.187  | 0.165            | 0.167                      | 0.269  |       |
| Sneezing MSE | 0.281  | 0.282            | 0.269                      |       |       |
| number of comp. | 3.2    | 2.5              | 1.9                        | 0.030  | 0.0863 |
| number of genes | 12023  | 3944.5           | 1721.5                     | 1.7547 x 10^{-9} | 0.0601 |
| Overall MSE | 0.623  | 0.622            | 0.609                      | 0.3073 | 0.2530 |
| Runny nose MSE | 0.186  | 0.174            | 0.173                      |       |       |
| Sneezing MSE | 0.277  | 0.284            | 0.272                      |       |       |
| number of comp. | 2.8    | 2.3              | 2.2                        | 0.0484 | 0.3773 |
| number of genes | 12023  | 2193.2           | 1779.1                     | 5.5038 x 10^{-13} | 0.3522 |
| Overall MSE | 0.628  | 0.607            | 0.603                      | 0.2020 | 0.4490 |
| Runny nose MSE | 0.243  | 0.226            | 0.232                      |       |       |
| Sneezing MSE | 0.324  | 0.323            | 0.314                      |       |       |
| number of comp. | 3.2    | 2.3              | 2.4                        | 0.0198 | 0.4163 |
| number of genes | 12023  | 2445.4           | 3889.8                     | 1.1584 x 10^{-9} | 0.1472 |
| Overall MSE | 0.866  | 0.920            | 0.855                      | 0.3318 | 0.0567 |
| Runny nose MSE | 0.312  | 0.327            | 0.312                      |       |       |
| Sneezing MSE | 0.412  | 0.448            | 0.397                      |       |       |
| number of comp. | 2.8    | 2.3              | 2.4                        | 0.143  | 0.145  |
| number of genes | 12023  | 3824.1           | 3575.8                     | 1.4451 x 10^{-10} | 0.4842 |
| Overall MSE | 0.599  | 0.603            | 0.591                      | 0.2890 | 0.1094 |
| Runny nose MSE | 0.187  | 0.165            | 0.167                      | 0.269  |       |
| Sneezing MSE | 0.281  | 0.282            | 0.269                      |       |       |

$p$-values of one sided paired t-test

8. Conclusion

The formulation of the global SIMPLS objective function with an added group sparsity penalty greatly reduces the number of variables used to predict the response. This suggests that when multiple components are desired, the variable selection technique should take into account the sparsity structure for the same variables among all the components. Our proposed Jointly Sparse Global SIMPLS Regression algorithm is able to achieve as good or better performance with fewer predictor variables and fewer components as compared to competing methods. It is thus useful for performing dimension reduction and variable
Matrix factorization. We use the cross-validated parameters reported in Table 4 for standard PLS-R and $\ell_1$ penalized PLS-R, and the Jointly Sparse Global SIMPLS-R to decide the number of components, and search over the grid \{1, 2, ..., 10\} to find the number of factors that achieves the highest correlation for NMF and BLU. The correlation between each factor and the sum of responses is listed for each method. The first 3 methods, PLS, $\ell_1$ SPLS, and $\ell_1/\ell_2$ SPLS are supervised matrix factorizations, whereas NMF and BLU are unsupervised. The unsupervised methods require many more factors to achieve comparable correlation.

| factor | H1N1 | H2N2 | HRV | RSV |
|--------|------|------|-----|-----|
| PLS    | 0.47 | 0.67 | 0.45 | 0.66 |
| $\ell_1$ SPLS | 0.52 | 0.73 | 0.52 | 0.70 |
| $\ell_1/\ell_2$ SPLS | 0.57 | 0.71 | 0.53 | 0.69 |
| NMF    | 0.32 | 0.62 | 0.02 | 0.41 |
| BLU    | 0.27 | 0.27 | 0.11 | 0.01 |

The correlation of each factor with the sum of upper respiratory symptoms is listed for each method and each viral infection.
selection simultaneously in applications with large dimensional data but comparatively few samples ($n < p$).

The Jointly Sparse Global SIMPLS Regression objective function is minimized using augmented Lagrangian techniques and, in particular, the ADMM algorithm. The ADMM algorithm splits the optimization into an eigen-decomposition problem and a soft-thresholding that enforces sparsity constraints. The general framework is extendable to more complicated regularization and can thus be tailored for other PLS-type applications, e.g., positivity constraints or smoothness penalties. For example, in the chemometric application, the data is smooth over the wavelengths and we can apply wavelet shrinkage on the data or include a total variation regularization to encourage smoothness. The sparsity constraints can be imposed on the wavelet coefficients if wavelet shrinkage is applied, or together with total variation regularization. The equivalence of soft wavelet shrinkage and total variation regularization was discussed in [45]. One can also consider imposing sparsity structures on the weights corresponding to the same components, adding $\ell_1$ penalty within the groups, or total variation regularization, depending on the applications. The decoupling property of the ADMM algorithm allows one to extend the Jointly Sparse Global SIMPLS Regression to these various regularizations.

References

[1] Wold, S., Martens, H., and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. *Proceedings of the Conference on Matrix Pencils. Lectures Notes in Mathematics*, 286-293.

[2] Sjöström, M., and Wold, S., and Lindberg, W., and Persson, J., and Martens, H. (1983). A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables. *Analytica Chimica Acta* **150**, 61-70.

[3] Martens, H., and Martens, M. (1999). Validation of PLS Regression models in sensory science by extended cross-validation. *PLS'99*.

[4] Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Genetics and Molecular Biology*, **7**(1), 35.

[5] Chun, H., and Keleş, S. (2009). Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*, **182**(1), 79-90.

[6] Chung, D., and Keleş, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical applications in genetics and molecular biology*, **9**(1), 17.

[7] Chun, H., Ballard, D. H., Cho, J., and Zhao, H. (2011). Identification of association between disease and multiple markers via sparse partial least squares regression. *Genetic epidemiology*, **35**(6), 479-486.

[8] Wold, S., Ruhe, A., Wold, H., and Dunn III, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to
generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3), 735-743.

[9] Chun, H., and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 3-25.

[10] de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3), 251-263.

[11] Wold, H. (1966). Nonlinear estimation by iterative least squares procedures. *Research papers in statistics*, 411-444.

[12] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6), 417.

[13] Tenenhaus, M. (1998). *La Régession PLS: théorie et pratique*. Editions Technip.

[14] de Jong, S. (1995). PLS shrinks. *Journal of Chemometrics*, 9(4), 323-326.

[15] Boulesteix, A. L., and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, 8(1), 32-44.

[16] Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, 2(3), 211-228.

[17] ter Braak, C. J., and de Jong, S. (1998). The objective function of partial least squares regression. *Journal of chemometrics*, 12(1), 41-54.

[18] Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3), 531-547.

[19] Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. *Perspectives in Probability and Statistics, In Honor of MS Bartlett*, 117-144.

[20] Allen, G. I., Peterson, C., Vannucci, M., and Maletić-Savatić, M. (2013). Regularized partial least squares with an application to NMR spectroscopy. *Statistical Analysis and Data Mining*, 6(4), 302-314.

[21] Bach, F. R. (2008). Consistency of the group Lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 9, 1179-1225.

[22] Gander, W. (1980). Least squares with a quadratic constraint. *Numerische Mathematik*, 36(3), 291-307.

[23] Gander, W., Golub, G. H., and von Matt, U. (1989). A constrained eigenvalue problem. *Linear Algebra and its applications*, 114, 815-839.

[24] Beck, A., Ben-Tal, A., and Teboulle, M. (2006). Finding a global optimal solution for a quadratically constrained fractional quadratic problem with applications to the regularized total least squares. *SIAM Journal on Matrix Analysis and Applications*, 28(2), 425-445.

[25] Eckstein, J., and Bertsekas, D. P. (1992). On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3), 293-318.

[26] Goldstein, T., and Osher, S. (2009). The split Bregman method for L1-regularized problems. *SIAM Journal on Imaging Sciences*, 2(2), 323-343.
[27] Afonso, M. V., Bioucas-Dias, J. M., and Figueiredo, M. A. (2011). An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems. *Image Processing, IEEE Transactions on*, 20(3), 681-695.

[28] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1), 1-122.

[29] Hong, M., and Luo, Z. Q. (2012). On the linear convergence of the alternating direction method of multipliers. arXiv preprint arXiv:1208.3922.

[30] Goldstein, T., O’Donoghue, B., and Setzer, S. (2012). Fast alternating direction optimization methods. CAM report, 12-35.

[31] Ramani, S., and Fessler, J. A. (2012) A splitting-based iterative algorithm for accelerated statistical X-ray CT reconstruction. *Medical Imaging, IEEE Transactions on*, 31(3), 677-688.

[32] Nien, H., and Fessler, J. A. (2014). A convergence proof of the split Bregman method for regularized least-squares problems. arXiv preprint arXiv:1402.4371.

[33] Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473).

[34] Bastien, P., Vinzi, V. E., and Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics & Data Analysis*, 48(1), 17-46.

[35] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

[36] Martens, H., and Naes, T. (1989). *Multivariate calibration*. Wiley.

[37] Masoum, S., Bouveresse, D. J. R., Vercauteren, J., Jalali-Heravi, M., and Rutledge, D. N. (2006). Discrimination of wines based on 2D NMR spectra using learning vector quantization neural networks and partial least squares discriminant analysis. *Analytica chimica acta*, 558(1), 144-149.

[38] Zaas, A. K., Chen, M., Varkey, J., Veldman, T., Hero, A. O., Lucas, J., Huang, Y., Turner, R., Gilbert, A., Lambkin-Williams, R., Øien, N. C., Nicholson, B., Kingsmore, S., Carin, L., Woods, C. W., and Ginsburg, G. S. (2009). Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans. *Cell Host and Microbe*, 6(3), 207-217.

[39] Huang, Y., Zaas, A. K., Rao, A., Dobigeon, N., Woolf, P. J., Veldman, T., Øien, N. C., McClain, M. T., Varkey, J. B., Nicholson, B., Carin, L., Kingsmore, S., Woods, C. W., Ginsburg, G. S., Hero, A. O. (2011). Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. *PLoS genetics*, 7(8), e1002234.

[40] Woods., C. W., McClain, M. T., Chen, M., Zaas, A. K., Nicholson, B. P., Varkey, J., Veldman, T., Kingsmore, S. F., Huang, Y., Lambkin-Williams, R., Gilbert, A. G., Hero, A. O., Ramsburg, E., Glickman, S., Lucas1, J. E., Carin, L., and Ginsburg, G. S. (2013). A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza H1N1.
or H3N2. *PloS one* 8.1: e52198.

[41] Zaas, A. K., Burke, T., Chen, M., McClain, M., Nicholson, B., Veldman, T., Tsalik, E. L., Fowler, V., Rivers, E. P., Otero, R., Kingsmore, S. F., Voora, D., Lucas, J., Hero, A. O., Carin, L., Woods, C. W., and Ginsburg, G. S. (2013). A Host-Based RT-PCR Gene Expression Signature to Identify Acute Respiratory Viral Infection. *Science translational medicine*, 5, 203ra126.

[42] Paatero, P., Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2), 111-126.

[43] Dobigeon, N., Moussaoui, S., Coulon, M., Tourneret, J. Y., and Hero, A. O. (2009). Joint Bayesian endmember extraction and linear unmixing for hyperspectral imagery. *Signal Processing, IEEE Transactions on*, 57(11), 4355-4368.

[44] Bazot, C., Dobigeon, N., Tourneret, J.-Y., Zaas, A. K., Ginsburg, G. S., Hero, A. O. (2013). Unsupervised Bayesian linear unmixing of gene expression microarrays. *BMC Bioinformatics*, 14(1), 99.

[45] Steidl, G., Weickert, J., Brox, T., Mrzek, P., and Welk, M. (2004). On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and SIDEs. *SIAM Journal on Numerical Analysis*, 42(2), 686-713.