Ann Oper Res (2011) 190:3–15
DOI 10.1007/s10479-009-0554-z

Enhanced predictions of wood properties using hybrid models of PCR and PLS with high-dimensional NIR spectral data

Yi Fang · Jong I. Park · Young-Seon Jeong · Myong K. Jeong · Seung H. Baek · Hyun Woo Cho

Published online: 4 June 2009

Abstract Near infrared (NIR) spectroscopy is a rapid, non-destructive technology to predict a variety of wood properties and provides great opportunities to optimize manufacturing processes through the realization of in-line assessment of forest products. In this paper, a novel multivariate regression procedure, the hybrid model of principal component regression (PCR) and partial least squares (PLS), is proposed to develop more accurate prediction models for high-dimensional NIR spectral data. To integrate the merits of PCR and PLS, both principal components defined in PCR and latent variables in PLS are utilized in hybrid models by a common iterative procedure under the constraint that they should keep orthogonal to each other. In addition, we propose the modified sequential forward floating search method, originated in feature selection for classification problems, in order to overcome difficulties of searching the vast number of possible hybrid models. The effectiveness and efficiency of hybrid models are substantiated by experiments with three real-life datasets of forest products. The proposed hybrid approach can be applied in a wide range of applications with high-dimensional spectral data.

Y. Fang
Department of Computer Science, Purdue University, West Lafayette, IN, USA

J.I. Park (✉)
Department of Industrial Engineering, Hanyang University, Seoul, Korea
e-mail: four87@psm.hanyang.ac.kr

Y.-S. Jeong · M.K. Jeong
Department of Industrial and Systems Engineering, The State University of New Jersey, Piscataway, NJ, USA

M.K. Jeong (✉)
RUTCOR (Rutgers Center for Operations Research), The State University of New Jersey, Rutgers, NJ, USA
e-mail: mjeong@rci.rutgers.edu

S.H. Baek · H.W. Cho
Department of Industrial and Systems Engineering, The University of Tennessee, Knoxville, TN, USA

Keywords Floating search · Hybrid models · Latent variables · Multivariate regression · NIR spectroscopy · Principal component analysis

1 Introduction

Real-time measurements of material properties in hardwood lumbers have great potential to be used for improving manufacturing processes. For example, rapid on-line measurements of the shrinkage of the lumbers can allow the sorting of them into more uniform groups, resulting in better use of the lumbers in the subsequent processing steps. One of the most promising ways for such real-time implementation of assessing the wood properties in manufacturing environment is near infrared (NIR) spectroscopy because it is a rapid, nondestructive, and relatively inexpensive technology. In particular, NIR spectroscopy has been shown to be an effective method to characterize a variety of material properties in the food, pharmaceutical and petroleum processing industries. For reviewing its potential for assessing the wood properties, refer to So et al. (2004).

NIR spectroscopy of wood involves measuring the surface diffuse reflection of electromagnetic radiation usually from 700 to 2500 nm. Prediction of wood properties can be stated as a multivariate regression problem in which the predictor variables are sampled NIR spectra and the response variables are the wood properties. It often results in ill-posed or ill-conditioned problems where the number of predictors greatly exceeds the number of observations. Several regression models have been developed for prediction with ill-posed data. Principal components regression (PCR) and partial least squares (PLS) are two of the most popular multivariate regression tools (Massy 1965; Wold 1966; Wentzell and Vega Montoto 2003). The relative strengths of these two approaches are often discussed and debated, but no clear conclusion has been reached. PLS is generally regarded as being superior to PCR in prediction. However, a few case studies have shown that PCR gave better prediction results than PLS did (Vigneau et al. 1996). Furthermore, no theoretical studies suggested that one method should predict better than another (Wentzell and Vega Montoto 2003). PCR and PLS have their own unique strength and weakness although they are very similar in some regards. Specifically, PCR can avoid overfitting ill-posed data by dramatically reducing the multicollinearity of the data (Massy 1965). On the other hand, the dimensionality reduction effect of PLS is less prominent, but the PLS model empirically has better discriminant ability than PCR (Kemsley 1996).

We propose hybrid models of PCR and PLS in order to combine their merits to develop more accurate regression models for predicting the wood properties based on the NIR spectral data. The key of hybrid models is that the linear transformed vector could be either a principal component (PC) or a latent variable (LV) (Fang et al. 2006). All the PCs and LVs are computed by a common iterative procedure under the constraint that they should keep orthogonal to each other. The ill-conditioned problems most likely benefit from a hybrid approach because in hybrid models, PCs can greatly decrease the multicollinearity of the data and at the same time LVs utilize the information from response variables. The optimal sequence of PCs and LVs for the hybrid model may be chosen from a vast number of candidates, which makes exhaustive search infeasible. In order to overcome this problem, the sequential forward floating search (SFFS) (Pudil et al. 1994a, 1994b) originally developed for feature selection is extended in this paper. The float search algorithm is one of the well-known heuristics, which guarantees the near optimal solution without any exhaustive search (Pudil et al. 1994a, 1994b). The modified SFFS method seems particularly effective since the optimal hybrid models with different number of components show small difference in

combination and the SFFS just searches the next optimal model in the neighborhood space of the current optimal one.

The remainder of this paper is organized as follows. In Sect. 2, we briefly review the PCR and PLS methods. In Sect. 3, hybrid models of PCR and PLS are introduced and the algorithm for building the optimal hybrid model is also given. We compare the performance of hybrid models with PCR and PLS using three real-life examples in Sect. 4. The conclusion and areas for further research are included in Sect. 5.

2 PCR and PLS regression

This section gives an overview of these two techniques in the same framework. Only single response regression problems are considered in this paper. Given a training set of data $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x}_n, y_n)$, $\mathbf{x}_i \in R^{1\times m}$, $y_i \in R$, the prediction problem is to construct some function f such that $f(\mathbf{x}_i)$ approximately equals y_i and the function generalizes well on future data. Each data point $\mathbf{x}_i$ is represented as the ith row in the data matrix $\mathbf{X}$. The ith response is denoted by y_i. Assuming that both $\mathbf{x}_i$ and y_i have been scaled to have means 0, n denotes the number of points, and m denotes the dimensionality of the data, so $\mathbf{X} \in R^{n\times m}$ and $\mathbf{y} \in R^{n\times 1}$.

2.1 Principal component regression

PCR consists of two steps. The first is to construct a linear projection mapping of the data using standard principal component analysis (PCA). PCs are usually computed by the singular value decomposition on $\mathbf{X}$, but its spirits come from the following ideas. The ith PC $\mathbf{w}_i$ can be derived by the objective function (1):

$$\max_{\mathbf{w}_i} \operatorname{var}(\mathbf{X}^{\mathrm{i}}\mathbf{w}_i) \quad \text{s.t. } \mathbf{w}_i^{\mathrm{T}}\mathbf{w}_i = 1 \tag{1}$$

where $\mathbf{X}^{\mathrm{i}}$ represents the residual after $(i-1)$ times. The optimal solution for $\mathbf{w}_i$ can be easily constructed using the first order optimality conditions (Hastie et al. 2001; Bennett and Embrechts 2003). The residual $\mathbf{X}^{\mathrm{i}}$ is updated by (2):

$$\mathbf{X}^{\mathrm{i}+1} = \mathbf{X}^{\mathrm{i}} - \mathbf{t}_i\mathbf{t}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{i}} \tag{2}$$

where $\mathbf{t}_i = \frac{\mathbf{X}^{\mathrm{i}}\mathbf{w}_i}{\|\mathbf{X}^{\mathrm{i}}\mathbf{w}_i\|}$ denotes the ith PC scores (the projected data on the ith principal component). Letting $\mathbf{X}^1 = \mathbf{X}$, then $\mathbf{w}_i$ can be calculated iteratively until the remaining components are deemed to be from noise or not to contain useful information.

The second step of PCR is to find the final regression coefficients $\mathbf{S}$ by minimizing the least squares error between the projected data $\mathbf{T}$, the score matrix whose columns are $\mathbf{t}$, and the response $\mathbf{y}$:

$$\mathbf{S} = (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{y} \tag{3}$$

In PCR, selecting a lower dimensional subspace for the mapping restricts the set of possible regression functions, thus limiting the capacity of the resulting function from overfitting the data. Therefore, PCR can perform well in ill-posed problems. Especially, when selecting the first principal components, PCR can greatly decrease the multicollinearity of the ill-posed data (Hastie et al. 2001; Bennett and Embrechts 2003).

2.2 Partial least squares regression

PLS is a supervised technique and performs a linear mapping of the data into the so called latent variables. Latent variables play, in PLS, the role that principal components play in PCR. The only difference of PLS (single response PLS model) from PCA is the objective function (4) by which the ith latent variable $\mathbf{w}_i$ is computed (note that the ith PC $\mathbf{w}_i$ is derived by the objective function (1)):

$$\max_{\mathbf{w}_i} \operatorname{cov}(\mathbf{X}^{\mathrm{i}}\mathbf{w}_i, \mathbf{y}) \quad \text{s.t. } \mathbf{w}_i^{\mathrm{T}}\mathbf{w}_i = 1 \tag{4}$$

where $\mathbf{X}^{\mathrm{i}}$ represent the residual after $(i-1)$ times. The residual $\mathbf{X}^{\mathrm{i}}$ is updated by the same way with PCA:

$$\mathbf{X}^{\mathrm{i}+1} = \mathbf{X}^{\mathrm{i}} - \mathbf{t}_i\mathbf{t}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{i}} \tag{5}$$

where $\mathbf{t}_i$ denotes the ith LV scores (the projected data on the ith latent variable). Let $\mathbf{X}^1 = \mathbf{X}$ and $\mathbf{y}^1 = \mathbf{y}$, then the final regression function can be built in the same way with PCA.

Similar to PCR development, PLS also builds a mapping of the data to a p $(<m)$ dimensional space and thus limits the capacity of the resulting function from overfitting the data. Unlike PCR, PLS utilizes both the input and the response data, $\mathbf{X}$ and $\mathbf{y}$ respectively, to form the mapping to a lower dimensional space.

3 Hybrid models of PCR and PLS

3.1 Illustration of hybrid models

Section 2 has shown that both PCR and PLS can be formulated in a similar iterative way. The difference between them only lies in the objective functions. If both objective functions are involved in a common iterative procedure, the properties of both PCR and PLS can be combined. Thus, the idea of constructing hybrid models of PCR and PLS consists of two steps. The first is to calculate PC and LV alternatively in iterative steps. In this way, the orthogonal decomposition is mixed with PCs and LVs. Based on the orthogonal decomposition, the original input data are mapped into a new subspace. The second step is to make the final regression function by minimizing the least-squares error between the projected data and the response y, as in the second step of both PCR and PLS. The key of hybrid models lies in that the projected vector in every orthogonal decomposition could be either a principal component or a latent variable. Like PCR and PLS, when the number of components in a hybrid model reaches the number of original predictor variables, the hybrid model is equivalent to the ordinary least-square (OLS) regression technique.

As shown in Fig. 1, for the 3-dimensional data (dots), PCR sequentially calculates the first three principal components PC1, PC2 and PC3. By contrast, a hybrid model of PLS and PCR may not calculate PC2 after getting PC1, but in the space orthogonal to PC1, the first latent variable LV2 may be calculated (the number 2 indicates that the computation is in the second iterative step).

It has been shown that a hybrid model of PCR and PLS is generated by a combination of PCs and LVs. Different combinations create different hybrid models. Here a sequence is used to denote a hybrid model. For example, the sequence PC1-LV2 represents the hybrid models illustrated by Fig. 1. The number following PC or LV means how many iterative steps (components) have already been calculated.

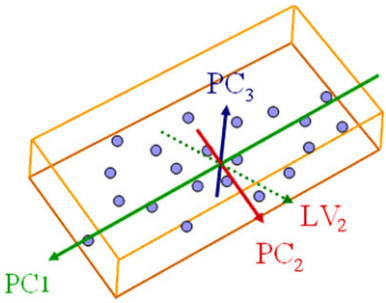

Fig. 1 A possible hybrid model with PC1 LV2

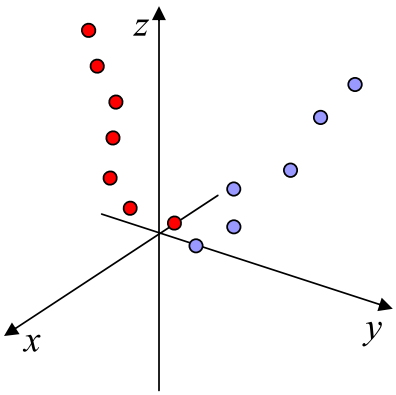

Fig. 2 Two-class data in the 3-d space

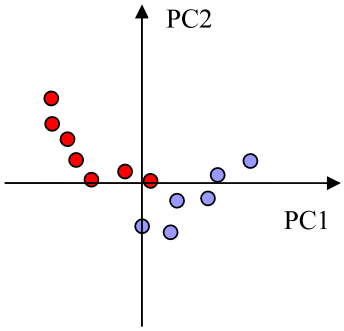

Fig. 3 The two-class data projected into 2-d PCs space

3.2 A conceptual example

Figure 2 is a two-class data (red and blue dots) simulated in 3-dimensional space. Since the data in each class seems to be distributed around a straight line, the data tend to be ill-conditioned. Now the regression function (here for a classification problem) can be constructed using 2 variables. Figure 3 shows the data in a 2-d PCs space. Although PC1 and PC2 keep most variance of the original data, neither of them is able to separate the data into 2 classes correctly. Figure 4 shows the data in a 2-d LVs space. We can see that both LV1 and LV2 have good discriminating ability. However, the projected data seem very "crowded". In this case, the projected data matrix **T** could be ill-conditioned.

Figure 5 shows the data in a 2-d hybrid model space. We can see that LV1 can make the projected data well separated and beyond that, PC2 keeps much variance of the data. It

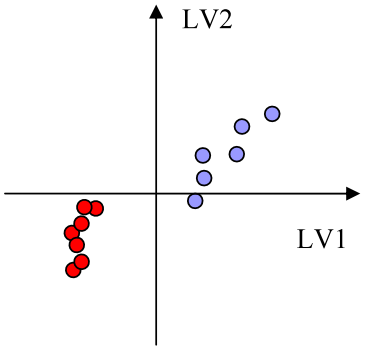

Fig. 4 The two-class data projected into 2-d LVs space

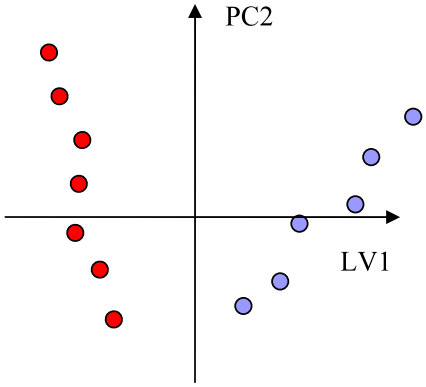

Fig. 5 The two-class data projected into a hybrid model space (LV1 PC2)

Table 1 Condition numbers for the original data and projected data

	Original	PCA	PLS	Hybrid
Condition number	31.435	1.7021	9.6286	3.2464

can be seen from these plots that hybrid models benefit from both advantages of PCA and PLS. The data projected in this way can produce a stable predictive model for a regression or classification with improved accuracy.

The condition numbers of the original and projected data in various cases are presented in Table 1. The original data have a large condition number and PCA can reduce it to the greatest extent while its discriminant ability is limited. The hybrid model seems to keep a good compromise between dimensionality reduction and discrimination by producing a relatively low condition number compared to the PLS case.

3.3 Basic algorithm

Now, we focus on the choice of the sequence of PCs and LVs. If the number of predictor variables to retain is k, there could be 2^k different hybrid models. Note that among all possible combinations of PCs and LVs, the pure PCR or PLS models are also included, such as the combination of PC1-PC2-PC3 or LV1-LV2-LV3 in the case of $k = 3$. Prediction models are created using the training data. The performance of each model can be then evaluated by comparing the root mean square error (RMSE) of prediction for the validation data. One

effective way to choose the optimal hybrid model is based on minimum cross validation (CV) error (Stone 1977). The idea behind CV is to recycle data by switching the roles of training and test samples. Based on the sample data available, different CV methods can be selected, including hold-out, k-fold, and leave-one-out cross validation (LOOCV). In this study, we employ a LOOCV method, which gives proper measure when there are limited samples. The LOOCV involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. The test sets are mutually exclusive and they effectively cover the entire data set.

In order to evaluate the predictive performance of every hybrid model, each different combination of components would be denoted by a different integral value, called the *determinant,* of a sequence. In this study, a binary numeral system is used to calculate the determinant. In a combination sequence, PC and LV are replaced by 0 and 1, respectively. Thus, every sequence can be represented by a binary number $b_k b_{k-1} b_{k-2} \ldots b_2 b_1$, where $b_i \in \{0, 1\}$. The determinant of a sequence is computed by converting the binary number to a decimal number. For example, the sequence LV1-PC2-LV3 is denoted by $(101)_2$ and thus its determinant is 5.

Below is the algorithm of constructing the optimal hybrid model of PCR and PLS. For convenience of presentation, hold-out cross validation is adopted in this algorithm flow. The only parameter of the algorithm is k, the number of components to retain.

Algorithm

Input: training input $\mathbf{X}$ and response $\mathbf{y}$, validation input $\mathbf{V}$ and response $\mathbf{z}$
Output: the optimal hybrid regression model

1. For $j = 0$ to $(2^k - 1)$ {
2. $b_k b_{k-1} b_{k-2} \ \ldots \ b_2 b_1 = (j)_2$
   /* convert j to a binary string
3. For $i = 1$ to k {
4. *If* $b_i = 0$
   Then $(\mathbf{w}_i, \mathbf{t}_i) = \mathrm{PCA}(\mathbf{X}^{\mathrm{i}})$
   /* calculate the first PC $\mathbf{w}_i$ and scores $\mathbf{t}_i$ based on the training input residual $\mathbf{X}^{\mathrm{i}}$
   $\mathbf{X}^{\mathrm{i}+1} = \mathbf{X}^{\mathrm{i}} - \mathbf{t}_i \mathbf{t}_i^{\mathrm{T}} \mathbf{X}^{\mathrm{i}}$
   /* update the training input residual
   Else $(\mathbf{w}_i, \mathbf{t}_i) = \mathrm{PLS}(\mathbf{X}^{\mathrm{i}}, \mathbf{y})$
   /* calculate the first LV $\mathbf{w}_i$ and scores $\mathbf{t}_i$ based on the training input residual $\mathbf{X}^{\mathrm{i}}$ and response $\mathbf{y}$
   $\mathbf{X}^{\mathrm{i}+1} = \mathbf{X}^{\mathrm{i}} - \mathbf{t}_i \mathbf{t}_i^{\mathrm{T}} \mathbf{X}^{\mathrm{i}}$
   /* update the training input residual
   }
5. $\mathbf{S} = (\mathbf{T}^{\mathrm{T}} \mathbf{T})^{-1} \mathbf{T}^{\mathrm{T}} \mathbf{y}$
   /* construct the least-square regression model based on the training projected data $\mathbf{T}$ and response $\mathbf{y}$
6. Calculate the test projected data $\mathbf{Q}$. Note that the test data $\mathbf{V}$ has to be projected into the subspace $\mathbf{w}$ using the same mapping as the training data did
7. $error = \|z - Qw\|$
   /* using the model calculated by step 5 to calculate the error rate on the test data
   }
8. Choose the optimal regression model with the minimum error rate

3.4 Model selection

In ill-posed problems, only few components contain most variance of data. The possible maximum number of components (denoted by D) could be small. In these cases, all the $(2^{D+1}-2)$ possible combinations of PCs and LVs are actually examined for choosing the optimal model. However, in some other cases (although they are unusual), when D is large, there are too many possible hybrid models. It can be regarded as a combinatorial optimization problem. Some heuristics can be used for solving this kind of problems, such as genetic algorithm, simulated annealing algorithm and ant colony optimization. However, these techniques fail to provide a stable solution in the sense that the solution depends on the initial values and some other parameters. In this paper, the SFFS algorithm in feature selection has been extended to choose the best hybrid model.

Floating search is originally developed for feature selection problems. It has been proven that floating search can provide a near-optimal solution to a combinatorial feature selection problem at an affordable computational cost (Jain and Zongker 1997). The model selection problem here is different from feature selection in the classification problem. In feature selection for the classification, the order of features chosen does not matter at all as long as the same features have been chosen, but in case of model selection for hybrid models, different combinations of components produce different regression models. Thus, the original SFFS algorithm for feature selection in the classification has to be modified to suit for the model selection of hybrid models. Below is the modified SFFS (MSFFS) algorithm.

Modified SFFS algorithm

Input: X and Y
Output: $B_k = \{b_k b_{k-1} b_{k-2} \ \ldots \ b_2 b_1 | b_j \in \{\mathbf{PC}_j, \mathbf{LV}_j\}\}$
Initialization:
$B_0 := \Phi$; $j = 0$; $X^j = X$; $Y^j = Y$;
Termination:
Stop when j equals the number of components required
For $j = 1 : k$ /* (If k is large, one can begin with $j = 10$ after using exhaustive search to get an optimal hybrid model B_{10}) */
Step 1 (Inclusion)
$b^+ := \arg \max_{b \in \{\mathbf{PC}_j, \mathbf{LV}_j\}} \text{LOORMSE}(B_j + b)$ /* choose either PC or LV */
$B_{j+1} := B_j + b^+$; $j := j + 1$;
Update $\mathbf{X}^j$ based on the algorithm in Sect. 3.3
Step 2 (Conditional exclusion)
$b^- := \arg \max_{b \in \{\mathbf{PC}_i, \mathbf{LV}_i\}} \text{LOORMSE}(B_j - b)$
If $\text{LOORMSE}(B_j - b^-) > \text{LOORMSE}(B_{j-1})$ then
$B_{j-1} := B_j - b^-$; $j := j - 1$;
go to Step 2
else
Update $\mathbf{X}^j$ based on the algorithm in Sect. 3
go to Step 1;
end for
(LOORMSE is a function to evaluate the LOOCV RMSE (root mean squared error) of the selected model.)

The search process for finding the best k components starts with the best sequence of the $(k-1)$ components previously found. Then, the algorithm has two steps: (1) inclusion; (2) conditional exclusion. The first step is essentially a greedy choice in the sense that the component with a minimum error rate is chosen. However, such choice may have severe drawback of being trapped into a local optimal. As a remedy, the second step looks back at the previous choice to see if there are other better combinations. The search process is repeated until no improvement is made in the model. The MSFFS may be particularly effective because the optimal models with different k seem not to have much difference in solution structures (combinations). The MSFFS searches the solution in the neighborhood space, which means that two consecutive solutions can be different in the two last components. Therefore, it is highly possible that the MSFFS can quickly find the next optimal solution based on the current one. However, an open question is that there is no theoretical bound on the computational cost of the algorithms due to their heuristic nature although the solution from the SFFS algorithm is near optimal (Pudil et al. 1994a).

In practice, for a given maximum value D, MSFFS searches among all the possible hybrid models with different k below D, which are then compared to determine the best k. This becomes possible because of the high efficiency of the SFFS algorithm. A common "best possible" value of k for hybrid models is expected to be between those of PCR and PLS. The best k for PCR and PLS can be chosen to be the number of components that gives a minimum LOOCV RMSE. Then, one way to specify D in practice could be to take the maximum of them.

4 Real-life examples

This section presents three real-life examples to evaluate the proposed model. In this section, we utilize two forest-related datasets and one public dataset.

4.1 Prediction of wood dimensional stability

Variations in wood dimensional stability (tendency to shrink and swell) can be particularly challenging for manufacturers that produce glued assemblies, such as laminated veneer lumber, plywood or parallel strand lumber. The dimensional stability of wood has known to be affected by density, extractive content, and so on. We have investigated the efficiency of NIR spectroscopy for predicting the shrinkage of woods. In this study, fifty-seven samples were obtained from mahogany, which is used in many applications, including as laminated components of musical instrument bodies. Two separate NIR spectra were collected for each wood sample (total $57 \times 2 = 114$ spectral data). The dimension of each observation is 2151 and the total number of observations is 114. The shrinkage, gravity, and extractive values for each of the samples are treated as the 3 response variables (see Taylor et al. 2008 for detailed description of the problem and data sets).

Table 2 shows the prediction errors of five methods, namely PCR, PLS, MSFFS hybrid model, optimal hybrid model, and a random selection of the sequence, for each response and the comparison results for the hybrid models using either the MSFFS or a random chosen sequence in order to provide insight into the quality of the proposed heuristic for model selection. In this experiment, each PC (or LV) was added one by one to the corresponding regression model until no improvement is made in the LOOCV RMSE in order to find the best number of components for PCR and PLS. Then, the maximum number of component D for the hybrid models took the larger one of the two. Exhaustive search was done to

Table 2 Prediction results for Mahogany data

Response	PCR		PLS		MSFFS hybrid model		Random selection hybrid model		Optimal hybrid model (HM)		Improvement of accuracy of hybrid model (%)	
	k	LOOCV RMSE	k	LOOCV RMSE	k	LOOCV RMSE	k	LOOCV RMSE	k	LOOCV RMSE	HM vs. PCR	HM vs. PLS
Shrinkage	9	0.5830	9	0.5368	9	0.5830	6	0.6284	9	0.5348	8.27%	0.37%
Gravity	9	0.4021	8	0.4222	7	0.4599	5	0.4886	9	0.4021	0.00%	4.76%
Extractives	9	0.6295	7	0.6412	9	0.6339	7	0.7120	9	0.6263	0.51%	2.32%

Table 3 Optimal hybrid models with different k for shrinkage and extractives

k	Shrinkage	Extractives
2	LV1LV2	LV1LV2
3	LV1LV2LV3	LV1LV2LV3
4	LV1LV2LV3LV4	LV1PC2LV3LV4
5	LV1LV2LV3LV4 LV5	LV1PC2PC3LV4PC5
6	LV1LV2LV3LV4 LV5LV6	LV1PC2PC3LV4PC5LV6
7	LV1LV2PC3LV4LV5LV6PC7	LV1PC2LV3LV4PC5LV6PC7
8	LV1LV2PC3LV4LV5LV6PC7LV8	LV1PC2LV3LV4PC5LV6PC7LV8
9	LV1LV2PC3LV4LV5LV6PC7LV8LV9	LV1PC2LV3LV4PC5LV6PC7LV8PC9
10	LV1LV2PC3LV4LV5LV6PC7LV8LV9LV10	LV1PC2LV3LV4PC5LV6PC7LV8PC9LV10

find an optimal hybrid model among all possible sequences of length up to D. For random selection, a combination of components was chosen uniformly among all those sequences. We also provides the relative performance of the optimal hybrid model against those of the conventional methods, which is calculated by

$$\frac{(\mathrm{RMSE}_{\mathrm{PCR/PLS}} - \mathrm{RMSE}_{\mathrm{optimal}})}{\mathrm{RMSE}_{\mathrm{PCR/PLS}}} \times 100(\%).$$

The optimal hybrid model always produces the best prediction results compared to PCR and PLS. Note that there is no further improvement in the performance by taking advantage of the hybrid model in case of predicting the gravity. Table 3 contains the optimal hybrid models with different values of k ranging from 2 to 10. It can be seen that the structures of some optimal models with different k are quite similar. For example, the optimal model with $k = 8$ for shrinkage content is the optimal model with $k = 7$ being added a LV. This interesting pattern is also found frequently in other cases for both shrinkage content and extractive content and then suggests that the MSFFS can produce near-optimal solution because the searching space in SFFS is always the neighborhood of the current optimal hybrid model. The performance of the MSFFS hybrid model slightly falls short of those of PCR and PLS, but it is much better than when the model chosen by the random sequence.

Table 4 Prediction results for gas oil dataset

Response	PCR		PLS		MSFFS hybrid model		Random selection hybrid model		Optimal hybrid model (HM)		Improvement of accuracy of hybrid model (%)	
	k	LOOCV RMSE	k	LOOCV RMSE	k	LOOCV RMSE	k	LOOCV RMSE	k	LOOCV RMSE	HM vs. PCR	HM vs. PLS
y_1	10	0.1185	6	0.0976	7	0.0976	5	0.1071	9	0.0951	16.75%	2.56%
y_2	10	0.1304	6	0.1088	10	0.1112	4	0.1253	9	0.1081	17.10%	0.64%
y_3	10	0.1961	10	0.1551	9	0.1643	8	0.1646	10	0.1551	20.91%	0.00%
y_4	8	0.1673	4	0.1738	7	0.1652	9	0.1866	8	0.1634	2.33%	5.98%

4.2 Gas oil example

The dataset in this example comes from the Wentzell group at Dalhousie University.[1] It consists of 115 samples from three subsets for which the UV spectra over 572 channels have been obtained. The dimension of each observation is 572 and the number of observations 115. The data contains 4 response variables, or the concentrations of the 4 components in each sample.

Table 4 illustrates that the optimal hybrid model produces the best prediction results compared to PCR and PLS. For some responses, the optimal model outperforms the PCR even with fewer components. Note that the MSFFS algorithm produces the near optimal solution without exhaustive search and this leads to the performance comparable to those of PCR and PLS. As expected, the MSFFS hybrid model produces better prediction results than the model chosen by a random sequence.

4.3 Chemical contents prediction in biomass

The accurate on-line prediction of ash and char contents in biomass is crucial in bio-energy manufacturing processes before any processing, such as gasification or fermentation. We collected a total of eighteen biomass samples from three different trees for different wood species (see Labbé et al. 2008 for the detailed description of experiments). Using the standard methods popular in the industry, the measurements for ash and char content were performed in triplicates for each of the eighteen biomass samples. The NIR spectra were recorded using an Analytical Spectral Devices (ASD) Field Spectrometer at 1-nm interval from 350 to 2500 nm. Three spectra per samples were recorded, and so, in total, 54 spectra (observations) were collected for the whole biomass set. The dimension of each observation is 538. The data contain two response variables, ash and char contents in each sample.

From Table 5, we can see that the optimal hybrid model is equivalent to the PLS models over all responses. Here, we have an example in which the corresponding optimal hybrid model may result in either a pure PLS or a pure PCR model. The MSFFS hybrid model gives better accuracy than PCR with equal or fewer components, but it performs slightly, worse than PLS.

[1] http://myweb.dal.ca/pdwentze/downloads.html

Table 5 Prediction results for biomass data

Response	PCR		PLS		MSFFS hybrid model		Random selection hybrid model		Optimal hybrid model (HM)		Improvement of accuracy of hybrid model (%)	
	k	LOOCV RMSE	k	LOOCV RMSE	k	LOOCV RMSE	k	LOOCV RMSE	k	LOOCV RMSE	HM vs. PCR	HM vs. PLS
Ash	10	0.0958	10	0.0594	10	0.0760	7	0.0881	10	0.0593	38.00%	0.00%
Char	10	0.2063	10	0.1637	8	0.1992	9	0.2005	10	0.1637	20.65%	0.00%

5 Conclusion and future work

A new multivariate regression procedure, the hybrid model of PCR and PLS, is proposed for developing more accurate prediction models by benefiting from both advantages of PCR and PLS. The MSFFS algorithm overcomes the difficulties of searching the vast number of possible hybrid models. The experimental results presented have shown great potential for the improvement of prediction accuracy for the ill-conditioned data.

Future research includes creating nonlinear hybrid models based on the kernel function (Schölkopf et al. 1998). Kernel PCR and kernel PLS have been proposed recently and achieve good prediction results (Rosipal and Trejo 2001). Thus, it is expected that the kernel hybrid models of PCR and PLS could function well in nonlinear cases.

Acknowledgements The part of this work was supported by the National Science Foundation grant CMMI-0644830. The authors thank the anonymous reviewers for their constructive comments.

References

Bennett, K. P., & Embrechts, M. J. (2003). An optimization perspective on partial least squares. In *Proceedings of the NATO Advanced Study Institute on learning theory and practice* (pp. 227–250). Amsterdam: IOS Press.

Fang, Y., Cho, H., & Jeong, M. K. (2006). Health monitoring of a shaft transmission system via hybrid models of PCR and PLS. In *Proceedings of SIAM conference on data mining* (pp. 554–558).

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Berlin: Springer.

Jain, A. K., & Zongker, D. (1997). Feature selection: evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 153–158.

Kemsley, E. K. (1996). Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemometrics and Intelligent Laboratory Systems*, *33*, 47–61.

Labbé, N., Lee, S.-H., Cho, H.-W., Jeong, M. K., & André, N. (2008). Enhanced discrimination and calibration of biomass NIR spectral data using nonlinear kernel methods. *Bioresource Technology*, *99*(17), 8445–8452.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, *60*, 234–256.

Pudil, P., Ferri, F. J., Novovicova, J., & Kittler, J. (1994a). Floating search methods for feature selection with nonmonotonic criterion functions. In *Proceedings of the 12th IAPR international conference on pattern recognition* (Vol. 2, pp. 279–283).

Pudil, P., Novovicova, J., & Kittler, J. (1994b). Floating search methods in feature selection. *Pattern Recognition Letters*, *15*, 1119–1125.

Rosipal, R., & Trejo, L. J. (2001). Kernel partial least squares regression in reproducing kernel Hilbert space. *Journal of Machine Learning Research*, *2*, 97–123.

Schölkopf, B., Smola, A., & Muller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, *10*, 1299–1319.

So, C.-L., Via, B. K., Groom, L. H., Schimleck, L. R., Shupe, T. F., Kelley, S. S., & Rials, T. G. (2004). Near infrared spectroscopy in the forest products industry. *Forest Products Journal*, *54*, 6–16.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika*, *64*, 29–35.

Taylor, A., Baek, S., Jeong, M. K., & Nix, G. (2008). Wood shrinkage prediction using NIR spectroscopy. *Wood and Fiber Science*, *40*(2), 301–307.

Vigneau, E., Bertrand, D., & Qannari, M. (1996). Application of latent root regression for calibration in near-infrared spectroscopy. Comparison with principal component regression and partial least squares. *Chemometrics and Intelligent Laboratory Systems*, *35*, 231–238.

Wentzell, P. D., & Vega Montoto, L. (2003). Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemometrics and Intelligent Laboratory Systems*, *65*, 257–279.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In *Proceedings of international symposium on multivariate analysis* (pp. 391–420). New York: Academic Press.