12-2006

Latent Variable Models with Applications to Spectral Data Analysis

Yi Fang
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a thesis written by Yi Fang entitled "Latent Variable Models with Applications to Spectral Data Analysis." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Industrial Engineering.

Myongkee Jeong, Major Professor

We have read this thesis and recommend its acceptance:

Denise Jackson, Adam Taylor

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a thesis written by Yi Fang entitled "Latent Variable Models with Applications to Spectral Data Analysis". I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Industrial Engineering.

Myongkee Jeong

Major Professor

We have read this thesis
and recommend its acceptance:

Denise Jackson

Adam Taylor

Accepted for the Council:

Linda Painter

Interim Dean of Graduate Studies

(Original signatures are on file with official student records.)

Latent Variable Models with Applications

to

Spectral Data Analysis

A Thesis

Presented for the

Master of Science Degree

The University of Tennessee, Knoxville

Yi Fang

December 2006

DEDICATION

This Thesis is dedicated to my parents

ACKNOWLEDGEMENT

I have thoroughly enjoyed and benefited from my career as a graduate student in the University Of Tennessee due to the strategic guidance of my advisor, Dr. Myong Kee Jeong. Dr. Jeong encouraged me to explore various areas, and ultimately provided priceless support in my chosen research topic. Under his patient supervision, I have discovered the area that interests me most and my research interests have been greatly intensified.

The other two members of my thesis committee also took many hours during this endeavor, providing valuable insights and feedback for my work. Dr. Adam Taylor in the department of Forestry, Wildlife, and Fisheries provided a nice application area for my work and unselfishly shared his expertise in this field with me. Moreover, he gave me detailed suggestions on writing my thesis and made me understand how to write professionally. I also gratefully thank Dr. Denise Jackson in the department of Industrial and Information Engineering for coming on board to help advise me to complete this thesis work.

Finally, I would like to thank all the other people lending a helping hand to me on this work. I pray that The Almighty God will bless you all.

ABSTRACT

Recent technological advances in automatic data acquisition have created an ever increasing need to extract meaningful information from huge amount of data. Multivariate predictive models have become important statistical tools in solving modern engineering problems. The purpose of this thesis is to develop novel predictive methods based on latent variable models and validate these methods by applying them into spectral data analysis.

In this thesis, hybrid models of principal components regression (PCR) and partial least squares regression (PLS) is proposed. The basic idea of hybrid models is to develop more accurate prediction techniques by combining the merits of PCR and PLS. In the hybrid models, both principal components in PCR and latent variables in PLS are involved in the common regression process.

Another major contribution of this work is to propose the robust probabilistic multivariate calibration model (RPMC) to overcome the drawback of Gaussian assumption in most latent variable models. The RPMC was designed to be robust to outliers by adopting a Student-t distribution instead of the Gaussian distribution. An efficient Expectation-Maximization algorithm was derived for parameter estimation in the RPMC. It can also be shown that some popular latent variables such as probabilistic PCA (PPCA) and supervised probabilistic PCA (SPPCA) are special cases of the RPMC.

Both the predictive models developed in this thesis were assessed on the real-life spectral data datasets. The hybrid models were applied into the shaft misalignment prediction problem and the RPMC are tested on the near-infrared (NIR) dataset. For the classification problem on the NIR data, the fusion of the regularized discriminant analysis (RDA) and principal components analysis (PCA) was also proposed. The experimental results have shown the effectiveness and efficiency of the proposed methods.

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

Recent technological advances in automatic data acquisition have created an ever increasing need to extract meaningful information from huge amounts of data. Researchers face larger data sets with more variables and more observations. Traditional statistical methods fail to work in many cases mostly because of the increase in the number of variables compared to the number of observations. These small sample size problems (SSS), in which the number of variables exceeds the number of observations, present many mathematical challenges. One of the notorious problems is so-called "curse of dimensionality" where performance degrades exponentially as a function of dimensionality (Hastie, 2001). These big challenges are bound to give rise to new theoretical developments.

Latent variables, also called hidden variables, are variables that are not directly observed but are rather inferred from other variables that are directly measured. One advantage of using latent variables is that it reduces the dimensionality of the data. A large number of observable variables can be simplified by latent variables to represent an underlying concept, making it easier to understand the data. In this sense, they serve the same function as theories do in science. At the same time, latent variables are link observable data in the real world to symbolic data in the modeled world. In the sense of dimensionality reduction, latent variable models include any rank-reduced techniques such as principal components regression (PCR) and partial least squares regression

(PLS). In the statistics community, there also exists a narrower definition of latent variable models (Everitt,1984).

In this thesis, hybrid models of PCR and PLS are proposed. Discussions and debate often arise as to the relative merits of these two approaches when applied to data come from real industrial applications, but neither method is clearly superior. PLS is generally regarded as being superior to PCR in prediction. However, a few cases have shown that PCR can give better prediction results than PLS. What is more, there were no theoretical studies which suggest that one method should predict better than the other (Wentzell et al, 2003). PCR and PLS have their unique strength and weakness although they are very similar in some ways. The goal of this work is to combine the strengths of PCR and PLS in order to develop more accurate prediction techniques.

Moreover, based on the specific definition of latent variable models in the statistics community, a robust probabilistic multivariate calibration model (RPMC) was developed in this thesis. Most latent variable models exploit the Gaussian assumption about the noise because the nice analytical property of Gaussian distributions often yields tractable algorithms for linear Gaussian models. A major limitation of them, however, is their sensitivity to outliers. This is easily understood by recalling linear Gaussian regression models in which maximization of likelihood function is equivalent to finding the least-squares solution, whose lack of robustness is well known. The RPMC is designed to be robust to outliers by adopting a Student-t distribution instead of the Gaussian distribution.

In addition, the fusion of the regularized discriminant analysis (RDA) and principal components analysis (PCA), denoted by RDA+PCA, is presented for dealing with near-infrared data. To the best of our knowledge, no previous work has been done in exploring RDA+PCA on NIR applications.

This thesis work consists of three major parts: hybrid models of PCR and PLS, the robust probabilistic multivariate calibration (RPMC) model and the fusion approach of RDA and PCA. In each part, a case study is demonstrated. The thesis is organized as follows: Chapter 2 develops hybrid models of PCR and PLS, and gives an example of its application to shaft misalignment prediction. Chapter 3 describes several latent variable models and introduces the robust probabilistic multivariate calibration (RPMC) model, and also gives an example of its application to NIR spectral data analysis. Chapter 4 presents the fusion of RDA and PCA for the classification problem, also a NIR application. Chapter 5 concludes the thesis with proposing possible future research directions.

2 Hybrid models of PCR and PLS

In a data-intensive world, multivariate regression models have become important statistical tools in solving modern engineering problems. For many years, researchers have sought to develop better regression techniques. Countless methods have evolved in an attempt to improve on existing methods. Principal components regression (PCR) (Chatterjee, 1977) and partial least squares (PLS) (Wold, 1966) are two of the most popular multivariate regression tools. 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 can give better prediction results than PLS. Furthermore, theoretical studies do not suggest that one method should predict better than the other (Wentzell, 2003). PCR and PLS have their own unique strength and weakness although they are very similar in some regards.

Fang *et al.* (Fang, 2005) proposed hybrid models of PCR and PLS to combine the merits of PCR and PLS in order to develop more accurate regression models. The key of hybrid models is that the linear transformed vector could be either a principal component (PC) or a latent variable (LV). In this chapter, Fang *et al.*'s idea is extended by exploring some properties of hybrid models, introducing conceptual examples, and applying the hybrid models to sample datasets. One of the challenging issues in hybrid models is that the optimal hybrid model may be chosen by comparing a vast number of candidates, which makes exhaustive search infeasible. In order to overcome this problem, I propose the

modified sequential forward floating search (SFFS) (Pudil, 1994) to choose the best hybrid model. The modified SFFS method seems particularly effective because the optimal hybrid models with different number of components show small differences in combination and the SFFS just searches for the next optimal model in the neighborhood space of the current optimal one. Small sample size 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 information from the response variable.

2.1 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 chapter. Given a calibration set of data $(\mathbf{x_1}, y_1), \cdots, (\mathbf{x}_n, y_n)$, $\mathbf{x}_i \in R^{1\times m}$, $y_i \in R$, 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 i^{th} row in the data matrix $\mathbf{X}$. The i^{th} 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 $\mathbf{y} \in R^{n\times 1}$.

2.1.1 Principal component regression

Principal component regression comprises two steps. The first is to construct a linear projection mapping of the data using standard principal component analysis. PCs are usually computed by the singular value decomposition on $\mathbf{X}$. The i^{th} PC $\mathbf{w_i}$ can be derived by the objective function (1):

$$\max_{\mathbf{w_i}} \operatorname{var}\left(\mathbf{X^i w_i}\right) \quad s.t \quad \mathbf{w_i}^T \mathbf{w_i} = 1 \tag{1}$$

where $\mathbf{X^i}$ represents the residual after (i-1) times. The residual $\mathbf{X^i}$ is updated by (2):

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

where $\mathbf{t_i}$ denotes the i^{th} PC scores (the projected data on the i^{th} principal component). Letting $\mathbf{X^1} = \mathbf{X}$ and $\mathbf{y^1} = \mathbf{y}$, 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 $\mathbf{T}$ and the response $\mathbf{y}$.

$$\mathbf{S} = \left(\mathbf{T^T T}\right)^{-1} \mathbf{T^T y}$$

(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 ill-posed data (Bennett, 2003).

2.1.2 Partial least squares regression

PLS is a supervised technique and performs a linear mapping between score vectors on latent variables. The only difference of PLS1 (single response PLS model) from PCA is the objective function (4) by which the i^{th} latent variable $\mathbf{w_i}$ is computed (note that the i^{th} PC $\mathbf{w_i}$ is derived by the objective function (1)):

$$\max_{\mathbf{w_i}} \operatorname{cov}\left(\mathbf{X^i w_i}, \mathbf{y}\right) \quad s.t. \ \mathbf{w_i}^T \mathbf{w_i} = 1 \tag{4}$$

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

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

where $\mathbf{t_i}$ denotes the i^{th} LV scores (the projected data on the i^{th} 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 (Geladi, , 1986).

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 PCA, PLS utilizes both the input and the response data, $\mathbf{X}$ and $\mathbf{y}$ respectively, to form the mapping to a lower dimensional space.

2.2 Hybrid models of PCR and PLS

2.2.1 Illustration of hybrid models

Section 2.1 has shown that both PCR and PLS1 can be formulated in a similar iterative way. The difference between them is in the objective functions. If the various 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 decompositions are mixed with PCs and LVs. Based on these orthogonal decompositions, the original input data is 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. 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 Figure 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 instead be calculated (the number 2 indicates that the computation is in the second iterative step).

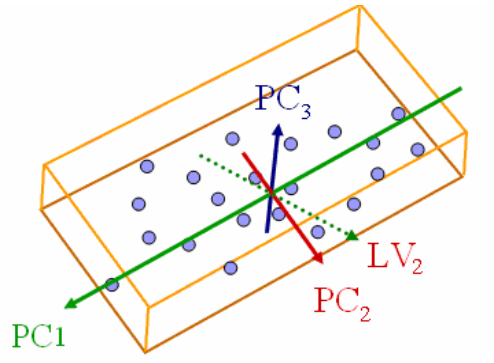

Figure 1 A possible hybrid model with the sequence PC1LV2

Thus 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 Figure 1. The number following PC or LV means how many iterative steps (components) have already been calculated.

2.2.2 A conceptual example

Figure 2 is a 2-class dataset (red and blue dots) in 3-dimensional space. Because the data in each class seem to be distributed around a straight line, the data suffers the small sample problem (SSS) effect. Now the regression function (here for a classification problem) can be constructed using 2 variables. Figure 3 shows the data in a 2-d PC space. Although PC1 and PC2 keep most of the information from the original data, neither of them is able to separate the data into 2 classes correctly.

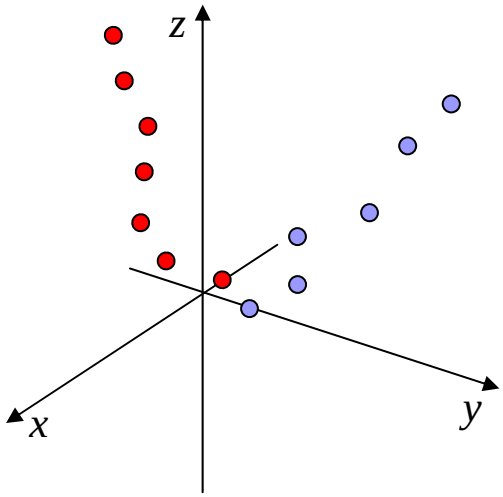

Figure 2 Two-class data in the 3-d space

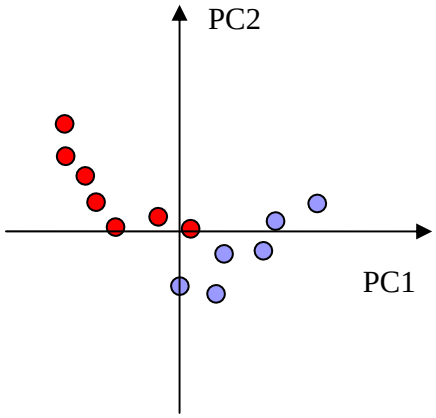

Figure 3 The Two-class data projected into 2-d PCs space

Figure 4 shows the data in a 2-d LVs space. One can see that both LV1 and LV2 have good discriminating ability. However, the projected data seem very "crowded". Figure 5 shows the data in a 2-d hybrid model space. One can see that LV1 can make the projected data well separated and beyond that, PC2 keeps much of the variance in the original data. It 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 better predictive model for a regression or classification problems.

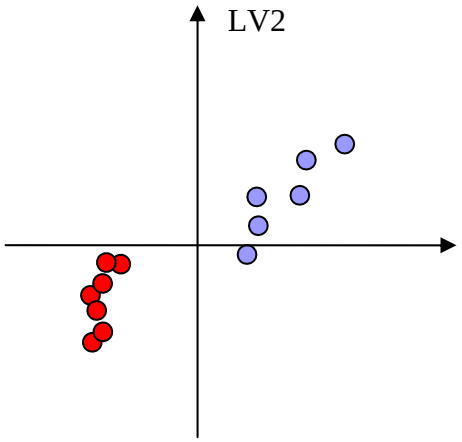

Figure 4 The Two-class data projected into 2-d LVs space

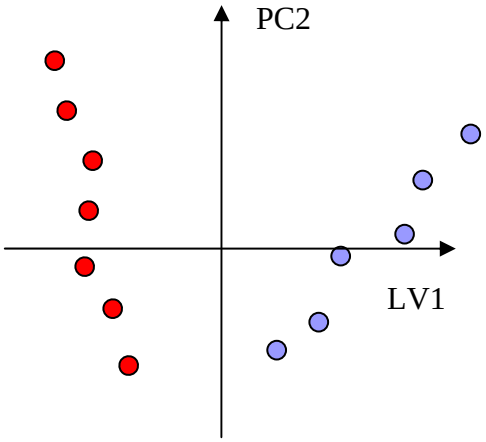

Figure 5 The 2-class data projected into a hybrid model space (LV1PC2)

2.2.3 Basic algorithm

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 case of k=3. 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 calibration and validation samples. Based on the sample data available, different CV methods can be selected, including hold-out, k-fold, and leave-one-out (LOOCV) methods.

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 $b_k b_{k-1} b_{k-2} \cdots b_2 b_1$, 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 6.

Below is the algorithm for 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.

Input: calibration 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. $b_k b_{k-1} b_{k-2} \cdots b_2 b_1 = (j)_2$

   /* convert *j* to a binary string
3. For $i=1$ to k {
4. If $b_i = 0$

   Then $(\mathbf{w}_i, \mathbf{t}_i) = \text{PCA}(\mathbf{X}^i)$

/* calculate the first PC $\mathbf{w_i}$ and scores $\mathbf{t}_i$ based on the calibration input residual $\mathbf{X^i}$

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

/* update the calibration input residual

Else $(\mathbf{w_i}, \mathbf{t}_i) = \text{PLS}(\mathbf{X^i}, \mathbf{y})$

/* calculate the first LV $\mathbf{w_i}$ and scores $\mathbf{t}_i$ based on the calibration input residual $\mathbf{X^i}$ and response $\mathbf{y}$

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

/* update the calibration input residual

}

5. $\mathbf{S} = \left(\mathbf{T}^\mathbf{T}\mathbf{T}\right)^{-1}\mathbf{T}^\mathbf{T}\mathbf{y}$

/* construct the least-square regression model based on the calibration projected data $\mathbf{T}$ and response $\mathbf{y}$

6. Calculate the validation projected data $\mathbf{Q}$. Note that the validation data $\mathbf{V}$ has to be projected into the subspace $\mathbf{w}$ using the same mapping as the calibration data did

7. $error = \|z - Qw\|$

/* using the model calculated by step 5 to calculate the cross-validation error on the validation data

}

8. Choose the optimal regression model with the minimum cross-validation error

2.2.4 Model selection

In SSS problems, few components contain most of the information in the data. The possible maximum value of k (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 random optimization techniques are currently popular for solving this kind of problem, such as genetic algorithm, simulated annealing algorithm and ant colony optimization (Engelbrecht, 2002). However, sometimes they fail to provide a stable solution. In this chapter, the SFFS algorithm in feature selection has been extended to choose the best hybrid model. Floating search was 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, 1997). The model selection problem here is different from feature selection in classification problems. In feature selection for classification, the order of features chosen does not matter as long as the same features have been chosen, but in 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 model selection of hybrid models. The modified SFFS algorithm is shown as follows.

Input:

$\mathbf{Y} = \{ y_j \mid y_j \in \{\mathbf{PCj}, \mathbf{LVj}\}, j = 1,..., k \}$

Output:

$B_k = \{ b_k b_{k-1} b_{k-2} \cdots b_2 b_1 \mid b_j \in \mathbf{Y} \}$

Initialization:

$B_0 := \Phi$; j =0; (In practice one can begin with j=10 after using exhaustive search to get an optimal hybrid model B_{10})

Termination:

Stop when j equals the number of components required

Step 1(Inclusion)

$b^+ := \arg\max_{b \in y_{j+1}} LOORMSE(B_j + b)$

$B_{j+1} := B_j + b^+$; j:=j+1;

Step 2 (Conditional exclusion)

$b^- := \arg\max_{b \in Y} LOORMSE(B_j - b)$

If $LOORMSE(B_j - b^-) > LOORMSE(B_{j-1})$ then

$B_{j-1} := B_j - b^-$; j:=j-1;

go to Step 2

else

go to Step1;

(LOORMSE is a function to evaluate the LOOCV RMSE of the selected model.)

The modified SFFS seems particularly effective because the optimal models with different *k* seem not to have much difference in solution structures (combinations). The modified SFFS searches for the solution in the neighborhood space. Therefore, it is highly possible that the modified SFFS can easily find the next optimal solution based on the current one.

In practice, similar to the approach mentioned in the exhaustive search, the value of *k* is not specified while the maximum value *D* is specified. All the possible optimal hybrid models with different *k* below *D* will be compared. 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 that for PCR and that for PLS. Therefore, one way to specify *D* in practice could be to take the maximum of them.

2.2.5 Case study: shaft misalignment prediction

Having shown the effectiveness of the hybrid models on a conceptual example, the proposed approach can be tested on motor shaft misalignment data accrued from a real industrial mechanical process. A shaft transmission system is one of the most fundamental parts of rotary machinery. Shaft misalignment measures how far apart the two centerlines are away from each other. Such shift in centers can be in parallel position, when the centerlines of the two shafts are parallel with each other, but at a constant distance apart, in angular position, when the centerlines are at an angle to each other, or a combination of these positions (Piotrowski, 1995), as shown in Figure 6. Recent studies

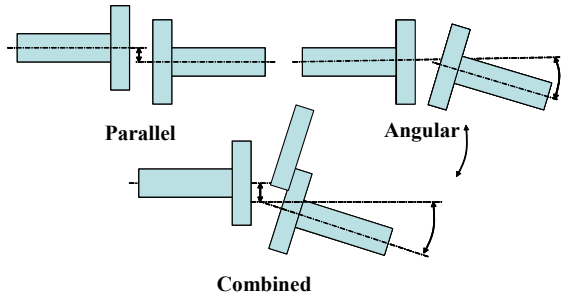

Figure 6 An illustration of parallel, angular, and combined misalignment conditions (Omitaomu, 2006)

indicate that there is a measurable change in the input power frequency spectrum of an electric motor for different shaft alignment conditions (Hines, 1999). Thus, the purpose of this research is to develop the optimal hybrid model to predict electric motor shaft alignment conditions based on the motor's power frequency spectrum. The problem can be stated as a multivariate regression problem in which the number of predictors greatly exceeds the number of observations.

Data was obtained from Oak Ridge National Laboratory Advanced Motor. A detailed description on the data collection was included in Omitaomu's work (Omitaomu, 2006). The time domain waveform data for the motor is shown in Figure 7. There are some differences in the power time domain for each alignment condition, such as different frequency components and some very small differences in magnitude. Advanced signal processing techniques were then used to transform the raw voltage and current data into the power time waveform. Therefore, for this analysis the input data is the power

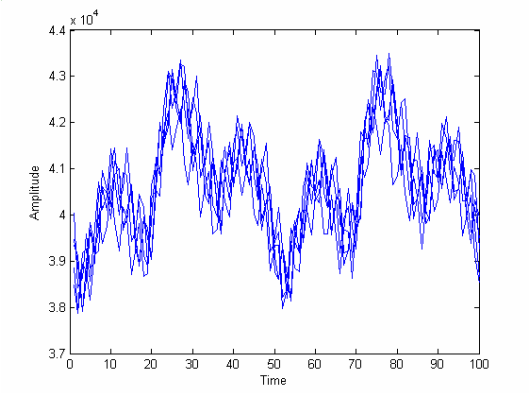

Figure 7 Part of raw data in time domain

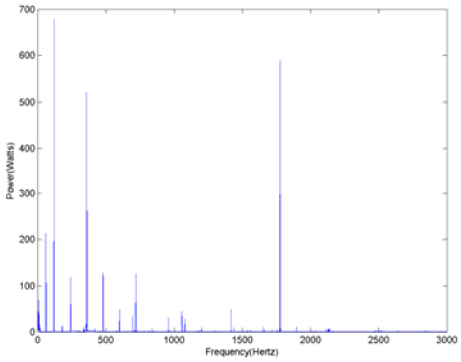

Figure 8 Typical FFT for an alignment condition

frequency spectrum and the response data is the misalignment condition. Figure 8 is a plot of the frequency spectrum input data for one of the misalignment conditions. The response data ranges from 0 to 50 mils (1mil= 2.540×10^{-5} m) for the parallel offset and from 0 to 15 mils for the angular offset. The sizes of the entire input set and response set are 50×3000 and 50×2 respectively. The condition number of the input data is 6.1246e+093, so it is a typical ill-posed problem. The objective then is to use the optimal hybrid model of PCR and PLS to predict the misalignment condition based on the power frequency spectrum. Although the response data has two variables, parallel misalignment and angular misalignment, they were predicted independently in this experiment.

PCR and PLS algorithms were used to determine the parameter *K*, the number of predictors. Because only 50 observations were available, leave-one-out cross validation (LOOCV) was used for accurately evaluating the prediction performance of regression models.Both PCR and PLS achieve minimum LOOCV MSEs with the number of predictors at 9 (see Figure 9 and Figure 10). Furthermore, the first 9 PCs can explain the 99.64% variance of the input data. For the angular response, similar results were produced. Therefore, the number of predictor variables to retain in hybrid models was specified as 9. Although the original input data has a large number of variables (i.e. 3000), most data information is actually contained in many fewer components (i.e. 9). Thus, the proposed algorithm with exponential complexity appears to be computationally efficient.

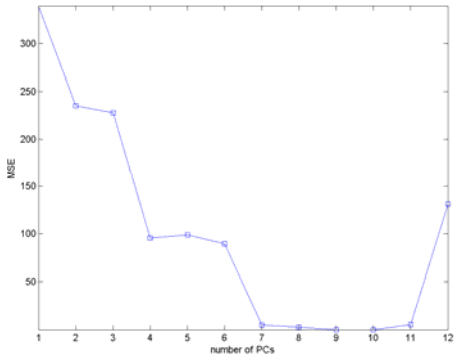

Figure 9 LOOCV MSEs with PCR for parallel misalignment condition

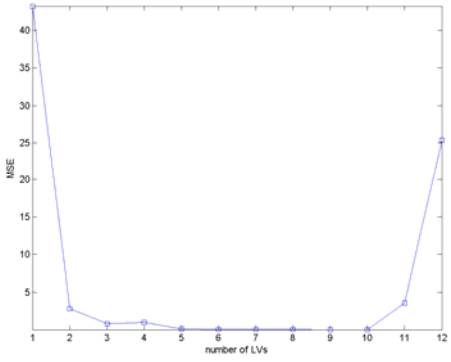

Figure 10 LOOCV MSEs with PLS for parallel misalignment condition

With *K* specified as 9, the proposed algorithm was performed on the misalignment data. Figure 11 and Figure 12 are the plots of LOOCV MSEs with different hybrid models for parallel and angular conditions, respectively. In the plots, the optimal regression models with minimum MSE are represented by points with red circles around them. The optimal model for the parallel condition is that with a determinant of 75 equivalent to binary 001001011. Therefore the corresponding combination sequence of PCs and LVs is PC1-PC2-LV3-PC4-PC5-LV6-PC7-LV8-LV9. The optimal prediction model for angular condition is that with a determinant of 99 whose corresponding combination sequence is PC1-PC2-LV3-LV4-PC5-PC6-PC7-LV8-LV9 (001100011). Note that in the plots, the points whose determinants are equal to 0 is the PCR model and determinant 511 is the PLS model. It can be seen that the optimal hybrid models outperform PCR and PLS in both parallel and angular misalignment prediction conditions.

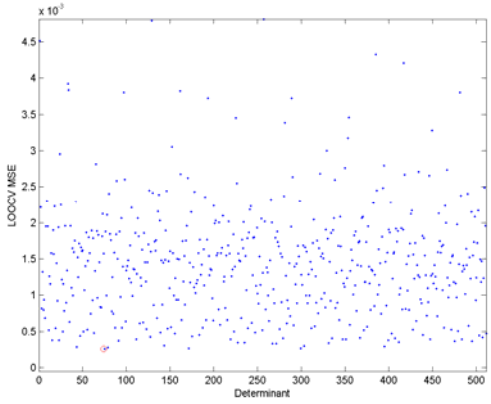

Figure 11 LOOCV MSEs with different hybrid models (*K*=9) for parallel conditions

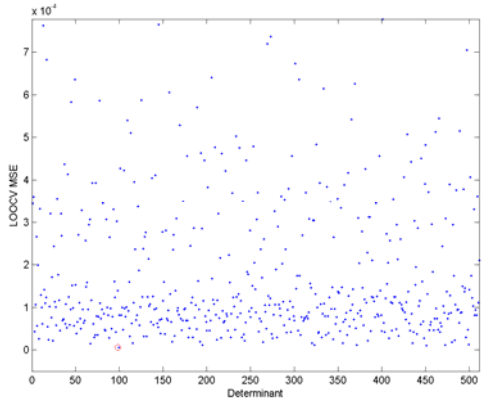

Figure 12 LOOCV MSEs with different hybrid models (*K*=9) for angular conditions

Table 1 contains all the experimental results when *K* ranges from 1 to 10. These experimental results validate that the optimal hybrid models achieve the best prediction results when *K* is 9. Table 1 also shows that the optimal hybrid model predicts more accurately than PCR and PLS when *K* is greater than 3. This suggests that the proposed approach may be particularly useful for complex prediction tasks that need more predictors. In addition, the MSEs for angular offset are much smaller than the MSEs for parallel offset, which implies that modeling the parallel offset is more difficult, at least for the given calibration data.

Table 1 The LOOCV error rates with PCR, PLS and the optimal hybrid model

K	**PCR**		**PLS**		**Optimal Hybrid Model**	
	Parallel	Angular	Parallel	Angular	Parallel	Angular
10	6.3259e-2	1.2564e-3	8.2317e-2	5.6314e-4	3.6479e-3	8.7461e-5
9	**4.5123e-3**	**3.4315e-4**	**1.4657e-3**	**7.4153e-5**	**2.5559e-4**	**5.5223e-6**
8	1.2684	3.5749e-2	1.0361e-2	1.0129e-4	9.3247e-4	1.2479e-5
7	4.3695	1.9654	9.8621e-2	9.1476e-4	5.2947e-3	9.5514e-5
6	91.237	2.2143	0.1579	6.3471e-3	1.2568e-2	3.2694e-4
5	98.865	9.9176	0.3214	1.0874e-2	0.0974	6.3247e-3
4	96.364	27.3695	1.8694	6.3727e-2	1.6987	0.8591e-2
3	227.68	25.9873	1.5697	0.1458	20.317	2.3697
2	235.41	25.416	3.1843	1.5147	41.585	8.1211

2.3 Conclusion

In this chapter, a new multivariate regression tool has been proposed. It aims to develop more accurate prediction models by benefiting from the advantages of both PCR and PLS. The results from the case study suggest the potential for improvement of prediction accuracy for SSS problems. Future research will include creating nonlinear hybrid models because many engineering problems have nonlinear properties. A possible solution is based on the Kernel Trick (Scholkopf, 1998), which has been proven as an efficient approach to deal with nonlinear problems. Kernel PCR and Kernel PLS have been proposed recently and can achieve good prediction results (Rosipal, 2001). Thus, it is expected that Kernel hybrid models of PCR and PLS could work well on some nonlinear cases.

3 Robust Probabilistic Multivariate Calibration

In this chapter, a robust probabilistic multivariate calibration (RPMC) model is proposed based on the specific definition of latent variable models in the statistics community (Everitt, 1984). The RPMC was intended to overcome the lack of robustness of the linear Gaussian models by adopting the Student-t distribution as the distribution of noises and latent variables instead of Gaussian distribution. It turns out that the RPMC includes some latent variable models as special cases, such as probabilistic PCA (PPCA) and supervised probabilistic PCA (SPPCA).

3.1 Latent variable models

A latent variable model is a statistical model that investigates the dependence of a set of observed variables on a set of latent variables (Everitt, 1984). The most well-known latent variable model is factor analysis, which was initially developed by psychologists. Recently, it has been found that many popular multivariate statistical techniques are closely related to latent variable models. These include vector quantization, independent component analysis models (ICA), Kalman filter models and hidden Markov models (HMMs) (Roweis, 1999). The general latent variable model has the following form:

$$p(\mathbf{x}) = \int p(\mathbf{x} \mid \boldsymbol{\theta}) h(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

where $\mathbf{x} = [x_1, \ldots, x_M]^T$ represents the observable variables and $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_P]^T$ the latent variables. The number of latent variables, P, is usually much less than the number of observable variables, M. In essence all latent variable models assume that $\mathbf{x}$ have a joint

probability distribution conditional on $\boldsymbol{\theta}$, denoted by $p(\mathbf{x}|\boldsymbol{\theta})$. Based on some assumptions, we can infer the density functions, *p* and *h* from the known or assumed density of $\mathbf{x}$ to discover how the manifest variables depend on the latent variables. The key assumption of latent variable models is that of conditional independence, i.e., the observable variables are independent of one another given the values of latent variables. In other words, the observed interdependence among the observable variables totally comes from their common dependence on the latent variables; once the latent variables are fixed, the behavior of the observable variables is essentially random. Mathematically, this can be expressed as:

$$p(\mathbf{x}) = \int h(\boldsymbol{\theta}) \prod_{i=1}^{M} p(x_i | \boldsymbol{\theta}) d\boldsymbol{\theta}$$

3.2 Probabilistic PCA

Principal component analysis (PCA) is a widely used statistical tool in chemometrics. While PCA originates from the analysis of data variances, it was recently expressed as the maximum likelihood solution for a generative latent variable model, which is called Probabilistic PCA or PPCA (Tipping *et al*, 1999b):

$$\mathbf{x} = \mathbf{W}_x \mathbf{t} + \mu_x + \varepsilon_x$$

where $\mathbf{t} \in \Re^P$ are latent variables, $\mathbf{W}_x$ is a $M \times P$ matrix called factor loadings, and ε_x defines a noise process. Additionally, we have parameters μ_x which allow non-zero means for the data. In this probabilistic model, latent variables $\mathbf{t}$ are conventionally

assumed as a standard Gaussian distribution, i.e., $\mathbf{t} \sim N(\mathbf{0}, \mathbf{I})$, and ε_x takes an isotropic Gaussian form as $\varepsilon_x \sim N(\mathbf{0}, \sigma_x^2 \mathbf{I})$.

The maximum likelihood solution of $\mathbf{W}_x$ is given as

$$\mathbf{W}_x = \mathbf{U}_P (\mathrm{E}_P - \sigma_x^2 \mathbf{I}_P)^{\frac{1}{2}} \mathbf{R} \tag{6}$$

where $\mathbf{U}_P$ is the matrix of the P principal eigenvectors of the sample covariance matrix $S_x = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \mu_x)(\mathbf{x}_i - \mu_x)^T$, $\mathbf{E}_P \in \mathbb{R}^{P \times P}$ is the diagonal matrix of the corresponding eigenvalues, $I_P \in \mathbb{R}^{P \times P}$ is the P-dimensional unit matrix, and $\mathbf{R}$ is an arbitrary $P \times P$ orthogonal matrix.

It can be shown that PCA is a limiting case of PPCA as $\sigma_x^2 \to 0$. In other words, PCA is recovered when the covariance of the noise becomes infinitesimally small and equal in all directions. This probabilistic formulation provides additional advantages over conventional PCA as discussed in Bishop's work (Tipping, 1999b), such as a principled way of handling missing values, a fast EM learning procedure and the availability of a Bayesian treatment.

3.3 Supervised PPCA

As each data observation is not only associated with input $\mathbf{x}$, but also with output $\mathbf{y} = [y_1, \ldots, y_K]^T \in \mathbb{R}^K$, unsupervised learning such as PCA or PPCA may be not able to project the data into useful subspaces. Many supervised learning methods have been proposed to make use of output information, such as principal component regression (PCR), partial least square (PLS) and linear discriminant analysis (LDA). Based on latent variable models, supervised probabilistic PCA (SPPCA) was recently introduced (Yu, 2006). Like PPCA, SPPCA utilizes the key point of latent variable models that all the observations are conditionally independent given the latent variables. In SPPCA, the observed data $(\mathbf{x}, \mathbf{y})$ is generated from a latent variable model as:

$$\mathbf{x} = \mathbf{W}_x \mathbf{t} + \mu_x + \varepsilon_x$$

$$\mathbf{y} = \mathbf{W}_y \mathbf{t} + \mu_y + \varepsilon_y$$

Again a unit isotropic Gaussian distribution is assumed for the P–dimensional latent vectors $\mathbf{t}$ and for the error terms ε_x and ε_y, i.e., $\mathbf{t} \sim N(\mathbf{0}, \mathbf{I})$, $\varepsilon_x \sim N(\mathbf{0}, \sigma_x^2 \mathbf{I})$, $\varepsilon_y \sim N(\mathbf{0}, \sigma_y^2 \mathbf{I})$. It is shown that the maximum likelihood solution of $\mathbf{W}_x$ and $\mathbf{W}_y$ are given by

$$\begin{aligned} \mathbf{W}_x &= \sigma_x \mathbf{U}_M (\mathrm{E}_P - \mathbf{I}_P)^{\frac{1}{2}} \mathbf{R} \\ \mathbf{W}_y &= \sigma_y \mathbf{U}_K (\mathrm{E}_P - \mathbf{I}_P)^{\frac{1}{2}} \mathbf{R} \end{aligned} \tag{7}$$

where $\mathbf{U}_M (\mathbf{U}_K)$ contains the first M (or last K) rows of eigenvectors of the normalized sample covariance matrix $\mathbf{S}$ for centered observations $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$,

$$\mathbf{S} = \begin{pmatrix} \frac{1}{\sigma_x^2}\mathbf{S}_{xx} & \frac{1}{\sigma_x\sigma_y}\mathbf{S}_{xy} \\ \frac{1}{\sigma_y\sigma_x}\mathbf{S}_{yx} & \frac{1}{\sigma_y^2}\mathbf{S}_{yy} \end{pmatrix},$$

$\mathbf{E}_P \in \mathbb{R}^{P\times P}$ is the diagonal matrix of the corresponding eigenvalues, $I_P \in \mathbb{R}^{P\times P}$ is the P-dimensional unit matrix, and $\mathbf{R}$ is an arbitrary $P\times P$ orthogonal matrix. The projected latent variable $\mathbf{t}^*$ for centered new input $\mathbf{x}^*$ is given by

$$\mathbf{t}^* = \frac{1}{\sigma_x}\mathbf{R}^T\left(\mathbf{E}_P - \mathbf{I}_P\right)^{\frac{1}{2}}\left[\mathbf{U}_M^T\mathbf{U}_M + \left(\mathbf{E}_P - \mathbf{I}_P\right)^{-1}\right]^{-1}\mathbf{U}_M^T\mathbf{x}^*$$

It is easy to check that (7) degrades to (6) when $K = 0$. In other words, PPCA is a special case of SPPCA. When $K > 0$, SPPCA explains not only intra-covariance of inputs $\mathbf{S}_x$ and intra-covariance of output $\mathbf{S}_y$, but also the inter-covariance between input and output, $\mathbf{S}_{xy}$ and $\mathbf{S}_{yx}$. In contrast to SPPCA, PCA only explains the covariance of inputs, and PLS finds the maximal covariance between inputs and outputs, but ignores the intra covariance of either inputs or outputs.

3.4 Student-t distribution

Both PPCA and SPPCA take advantage of the Gaussian assumption about noise based on the fact that the convolution of two independent Gaussian distributed quantities is also Gaussian distributed. This nice analytical property of Gaussian distributions often yields tractable algorithms for linear Gaussian models. A major limitation of them, however, is their sensitiveness to outliers. This is easily understood by recalling the linear Gaussian

regression models in which the maximization of likelihood function is equivalent to finding the least-squares solution, whose lack of robustness is well known (Svensen, 2004).

Several approaches have been proposed to address the limitation of Gaussian models. Most methods rely on robust estimation, particularly M-estimation (Huber, 1981). M-estimation assumes that the data are heavy tailed distributed instead of normally distributed. Consequently, maximum likelihood solutions are more robust to outliers.

In this section, a robust multivariate calibration approach is developed based on latent variable models whose components have a Student distribution, also known as t-distribution,

$$S(\mathbf{x} \mid \mu, \Lambda, \nu) = \frac{\Gamma(\nu/2 + d/2) \mid \mathbf{\Lambda} \mid^{1/2}}{\Gamma(\nu/2)(\nu\pi)^{d/2}} \left(1 + \frac{\Delta^2}{\nu}\right)^{-(\nu+d)/2}$$

where

$\Gamma(z) = \int_0^\infty y^{z-1} e^{-y} dy$ and $\Delta^2 = (x-\mu)^T \Lambda (x-\mu)$ is the squared Mahalanobis distance.

The t-distributions have heavier tails compared to the exponentially decaying tails of a Gaussian (see Figure 13). They are commonly used in robust regression (Langer, 1989). Previous work has been done to increase the robustness of PPCA and PCCA by replacing Gaussian distributions with t-distributions (Cedric, 2006, Bach, 2005). They have also

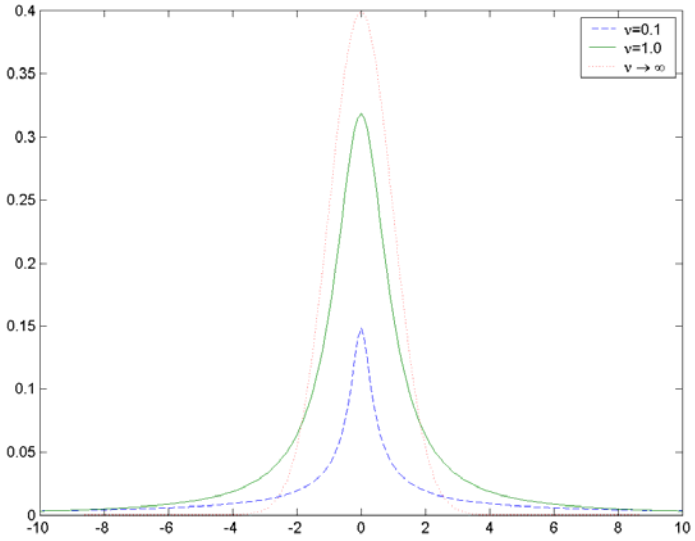

Figure 13 Illustration of heavy tail Student-t distribution

been shown to be effective in the computer vision and mixture modeling (Svensen, 2004).

3.5 Robust probabilistic multivariate calibration model

Instead of taking a Gaussian distribution on the latent variables in SPPCA, I choose a Student-t distribution because I assume that outliers in the original data space will also be outliers in the latent variable space. In addition, I assume that the noise is also drawn from a Student-t distribution. Consequently, this leads to the following probabilistic model,

$$p(\mathbf{t}_i) = S(\mathbf{t}_i \mid \mathbf{0}, \mathbf{I}_P, \nu)$$
$$p(\mathbf{x}_i) = S(\mathbf{x}_i \mid \mathbf{W}_x \mathbf{t}_i + \mu_x, \sigma_x^{-2}\mathbf{I}_M, \nu)$$
$$p(\mathbf{y}_i) = S(\mathbf{y}_i \mid \mathbf{W}_y \mathbf{t}_i + \mu_y, \sigma_y^{-2}\mathbf{I}_K, \nu)$$

In contrast to the Gaussian, there is no close form solution for maximizing likelihood under this model. However, there exists an alternative representation of a t-distribution in terms of latent variable models. In particular, we can write it as an infinite mixture of scaled Gaussians (Liu, 1995),

$$S(\mathbf{x} \mid \mu, \Lambda, \nu) = \int_0^\infty N(\mathbf{x} \mid \mu, \boldsymbol{\theta}\Lambda) G(\boldsymbol{\theta} \mid \nu/2, \nu/2) d\boldsymbol{\theta} \tag{8}$$

Based on the latent variable model definition (8) of the Student-t distribution, we can define an tractable robust probabilistic multivariate model:

$$p(\boldsymbol{\theta}_i) = G(\boldsymbol{\theta}_i \mid \frac{\nu}{2}, \frac{\nu}{2})$$
$$p(\mathbf{t}_i \mid \boldsymbol{\theta}_i) = N(\mathbf{t}_i \mid \mathbf{0}, \boldsymbol{\theta}_i \mathbf{I}_P)$$
$$p(\mathbf{z_i} \mid \mathbf{t_i}, \boldsymbol{\theta_i}) = N(\mathbf{z_i} \mid \mathbf{W t_i} + \mathbf{u}, \boldsymbol{\theta_i} \boldsymbol{\Phi}^{-1})$$

where

$$\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i), \mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \Phi = \begin{pmatrix} \sigma_x^2 \mathbf{I} & 0 \\ 0 & \sigma_y^2 \mathbf{I} \end{pmatrix}, \mathbf{W} = \begin{pmatrix} \mathbf{W}_x \\ \mathbf{W}_y \end{pmatrix}$$

The log likelihood of the whole observation is

$$L = \sum_{i=1}^{N} \log \int p(\mathbf{x}_i \mid \mathbf{t}_i) p(\mathbf{y}_i \mid \mathbf{t}_i) p(\mathbf{t}_i) d\mathbf{t}_i$$

When both input and output are observed, the posteriori distribution of given as

$$p(\boldsymbol{\theta}_i \mid \mathbf{z}_i) \propto p(\mathbf{z}_i \mid \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i) = G(\boldsymbol{\theta}_i \mid \frac{M + K + \nu}{2}, \frac{(\mathbf{z}_i - \mu)^T \mathbf{A} (\mathbf{z}_i - \mu) + \nu}{2})$$

where $\mathbf{A}^{-1} \equiv \mathbf{W}\mathbf{W}^T + \boldsymbol{\Phi}$

$$p(\mathbf{t}_i \mid \mathbf{z}_i, \boldsymbol{\theta}_i) \propto p(\mathbf{z}_i \mid \mathbf{t}_i, \boldsymbol{\theta}_i) p(\mathbf{t}_i \mid \boldsymbol{\theta}_i) = N(\mathbf{t}_i \mid \mathbf{B}^{-1} \mathbf{W}^T \boldsymbol{\Phi}^{-1} (\mathbf{z}_i - \mu), \boldsymbol{\theta}_i \mathbf{B})$$

where $\mathbf{B} \equiv \mathbf{W}^T \boldsymbol{\Phi}^{-1} \mathbf{W} + \mathbf{I}_P$

The sufficient statistics needed to update the parameters in the M-step are then given by

$$\langle \boldsymbol{\theta}_i \rangle = \frac{M + K + \nu}{(\mathbf{z}_i - \mu)^T \mathbf{A} (\mathbf{z}_i - \mu) + \nu}$$

$$\langle \log \boldsymbol{\theta}_i \rangle = \psi(\frac{M + K + \nu}{2}) - \log(\frac{(\mathbf{z}_i - \mu)^T \mathbf{A} (\mathbf{z}_i - \mu) + \nu}{2})$$

where $\psi(.)$ denotes the digamma function.

$$\langle \mathbf{t}_i \rangle = \mathbf{B}^{-1} \mathbf{W}^T \boldsymbol{\Phi}^{-1} (\mathbf{z}_i - \mu)$$

$$\langle \boldsymbol{\theta}_i \mathbf{t}_i \mathbf{t}_i^T \rangle = \mathbf{B}^{-1} + \langle \boldsymbol{\theta}_i \rangle \langle \mathbf{t}_i \rangle \langle \mathbf{t}_i \rangle^T$$

Then, in the M-step, the mean vector is updated by

$$\mu = \frac{\sum_{i=1}^{N} \langle \boldsymbol{\theta}_i \rangle (\mathbf{z}_i - \mathbf{W} \langle \mathbf{t}_i \rangle)}{\sum_{i=1}^{N} \langle \boldsymbol{\theta}_i \rangle}$$

The factor loading matrices are updated by

$$\mathbf{W} = (\sum_{i=1}^{N} \langle \boldsymbol{\theta}_i \rangle (\mathbf{z}_i - \mu) \langle \mathbf{t}_i \rangle^T)(\sum_{i=1}^{N} \langle \boldsymbol{\theta}_i \mathbf{t}_i \mathbf{t}_i^T \rangle)^{-1}$$

The variance matrices are updated by

$$\sigma_x^2 = \frac{1}{N \times M} \sum_{i=1}^{N} \{ \langle \boldsymbol{\theta}_i \rangle \| \mathbf{x}_i - \mu \|^2 - 2 \langle \boldsymbol{\theta}_i \rangle (\mathbf{x}_i - \mu)^T \mathbf{W}_x \langle \mathbf{t}_i \rangle + \mathbf{tr} \{ \langle \boldsymbol{\theta}_i \mathbf{t}_i \mathbf{t}_i^T \rangle \mathbf{W}_x{}^T \mathbf{W}_x \} \}$$

$$\sigma_y^2 = \frac{1}{N \times K} \sum_{i=1}^{N} \{ \langle \boldsymbol{\theta}_i \rangle \| \mathbf{y}_i - \mu \|^2 - 2 \langle \boldsymbol{\theta}_i \rangle (\mathbf{y}_i - \mu)^T \mathbf{W}_y \langle \mathbf{t}_i \rangle + \mathbf{tr} \{ \langle \boldsymbol{\theta}_i \mathbf{t}_i \mathbf{t}_i^T \rangle \mathbf{W}_y{}^T \mathbf{W}_y \} \}$$

Finally, the ML solution of ν is calculated by solving the following equation:

$$1 + \log(\frac{\nu}{2}) - \psi(\frac{\nu}{2}) + \frac{1}{N} \sum_{i=1}^{N} \{ \langle \log \boldsymbol{\theta}_i \rangle - \langle \boldsymbol{\theta}_i \rangle \} = 0$$

3.6 Case study: prediction of green moisture content and density of solid wood

In this experiment, RPMC was assessed as a possible method for predicting the moisture content and density of solid red oak wood (response variables). The data was collected by the near infrared (NIR) spectrometer. A detailed description of the data collection is contained in Defo *et al.* (In press). The dataset used in this analysis were collected in a similar manner; however, the actual wood samples used were different and a different near infrared spectrometer was used (A. Taylor pers. comm.).

Figure 14 shows the NIR data of the 45 samples with 275 different wavelengths. The total samples were divided into a calibration dataset with 30 samples and a validation dataset with 15 samples. Table 2 summarizes the mean error rates for PLS, SPPCA and RPMC on the calibration and validation datasets respectively. The number of components was chosen as 10.

The experimental results in Table 2 show that RPMC and SPPCA predicted much better than conventional PLS in both responses. RPMC performed slightly better than SPPCA in both cases, which indicates that the dataset may contain certain outliers.

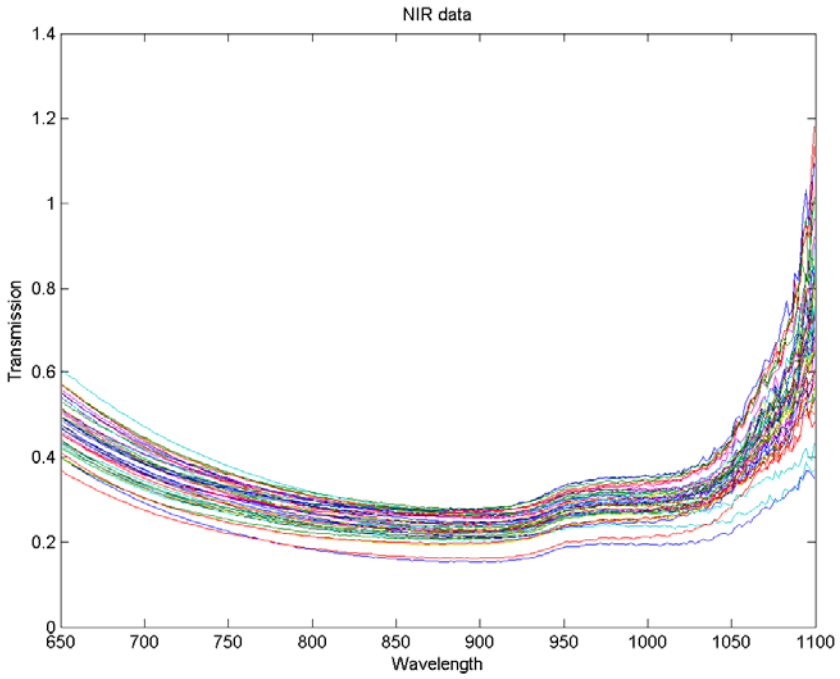

Figure 14 NIR data profile of 45 samples

Table 2 Mean square errors for prediction of moisture content (y1) and basic density content (y2) by PLS , SPPCA and RPMC

		PLS	SPPCA	RPMC
y1	Calibration	31.6504	22.9949	23.0228
	Validation	75.2048	33.9409	19.7674
y2	Calibration	2887.1	2814.3	2859.8
	Validation	8038.5	3778.7	3615.4

3.7 Conclusions

In this chapter, a robust probabilistic multivariate calibration (RPMC) model was developed. The RPMC was intended to overcome the lack of robustness of the linear Gaussian models by adopting the Student-t distribution as the distribution of noises and latent variables instead of the Gaussian distribution. In contrast to the linear Gaussian models, there is no close form solution for the linear Student-t model. Based on the latent variable model Student-t distribution, an efficient EM algorithm was derived for parameter estimation. The experimental results show the promise of RPMC for NIR–based quantification problems.

4 The Fusion Approach of RDA and PCA

In this section, the fusion approach of the regularized discriminant analysis (RDA) and the PCA is presented. The proposed method was motivated by the idea to overcome the curse of dimensionality using PCA. As introduced in the section 2.1, PCA is a powerful statistical tool to reduce the dimensionality of the high-dimensional data. After PCA is performed on the original data, conventional techniques are expected to work well on the reduced- rank data. In addition, a classification problem based on NIR spectral data is also demonstrated. To the best of our knowledge, no previous research has been conducted applying RDA+PCA into NIR spectral data analysis.

4.1 Regularized discriminant analysis

Linear discriminant analysis (LDA) and the related Fisher's linear discriminant (Fisher, 1936) are used in statistics to find the linear combination of features which best separate two or more classes of object or event. It searches for a set of projection vectors onto which the data points of the same class are close to each other, while requiring data points of different class to be far from each other. Quadratic discriminant analysis (QDA) is closely related with LDA, while its decision boundaries are approximated by quadratic equations.

Regularized discriminant analysis (Friedman, 1989), a compromise between LDA and

QDA, allows one to shrink the separate covariance as LDA toward a common covariance as in LDA. The regularized covariance matrices have the form

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$$

where $\hat{\Sigma}$ is the pooled covariance matrix as used in LDA. Here $\alpha \in [0,1]$ allows a continuum of models between LDA and QDA, and need to be specified. When $\alpha = 1$, RDA is equivalent to LDA, and when $\alpha = 0$, RDA is equivalent to QDA.

4.2 RDA+PCA

Unfortunately, in many classification tasks, $\hat{\Sigma}$ is typically singular, due to the fact that the number of the samples is much smaller than the dimension of the sample space, i.e., the so-called Small Sample Size (SSS) problem. In this chapter, we intend to overcome this problem by integrating PCA into RDA. The approach, called RDA+PCA, is the method that performs feature extraction in two sequential stages: in PCA+RDA, the NIR samples are projected to a PCA subspace in the first stage, and then RDA is applied secondly; in RDA, the NIR samples are firstly projected to the range space of between-class matrix $\mathbf{S}_B$ through QR-decomposition, followed by RDA in the second stage.

4.3 Case study: wood preservative identification

In this experiment, the RDA+PCA approach was applied to a classification problem. The task was to correctly differentiate preservative treatments on wood using NIR signal data.

A description on the data collection technique is included in Defo *et al*. (In press). However, the dataset provided for this analysis has not been previously published (A. Taylor pers. comm.). 80 samples were prepared and the data are shown in Figure 15. Among them, 20 were treated with a borates wood preservative (from the 21st sample to the 40th sample), denoted as the "target", while other 60 were treated with different borate-based products, denoted by "non-target". The whole dataset is divided into two parts, i.e., a calibration set and a validation set. The predictive model was created with the calibration set and subsequently the model was verified with the validation dataset. The dataset was evenly divided in the way illustrated in Figure 16, where k is the number chosen in each data section for calibration purposes.

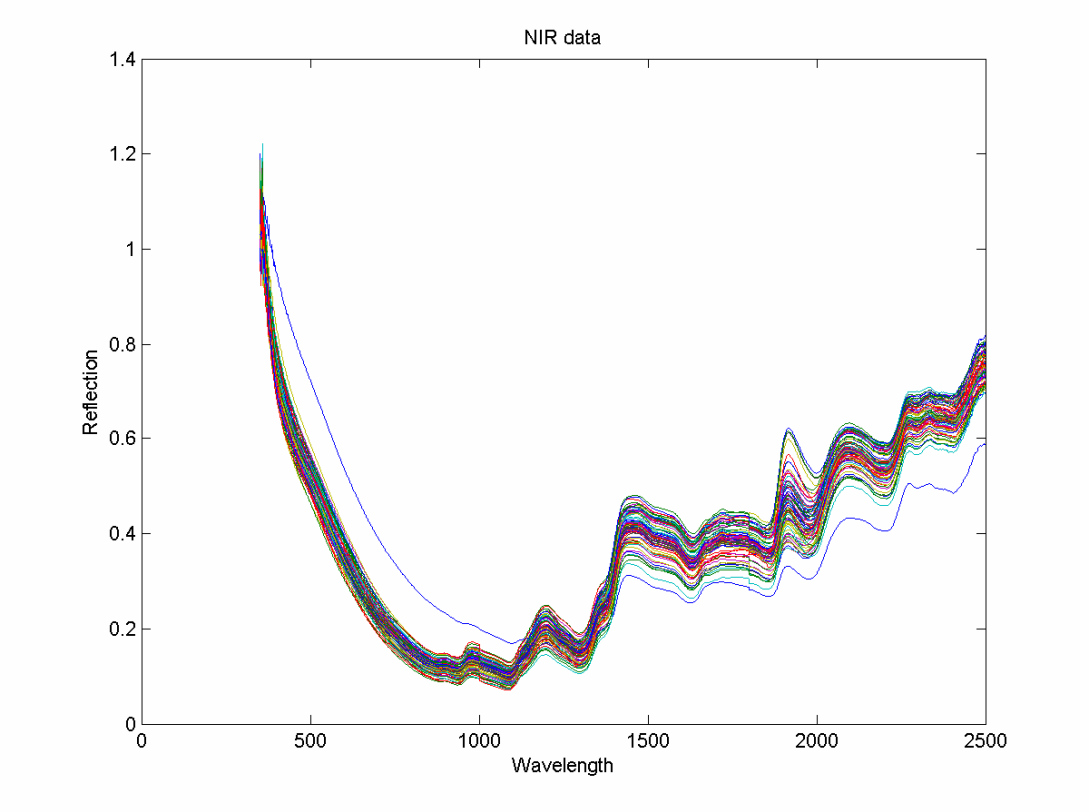

Figure 15 NIR data profile of 80 samples

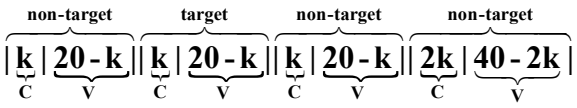

Figure 16 Dataset partition with variable k

4.3.1 Classification results

RDA+PCA was applied to the dataset with two different experimental settings:

1. *k* is fixed as 15 and α ranging from 0.1 to 0.9;

2. α is fixed as 0.9 and *k* is changed from 13 to 19.

The experimental results are shown in Figure 17 and Figure 18, respectively. When α=0.8 or 0.9, the classification error rate was 0. As discussed in the section 4.1, when α is close to 0, RDA is going towards QDA; when α is close to 1, RDA is going towards LDA. The experimental results with a large α value suggest that the RDA is almost equivalent to LDA. When the sample size (*k)* of the calibration set was incrementally increased, the sample size of validation set was correspondently decreased. Consequently, we can see from Figure 18 that the prediction results were improved, which indicates RDA+PCA could achieve good performance when sufficient calibration data are available.

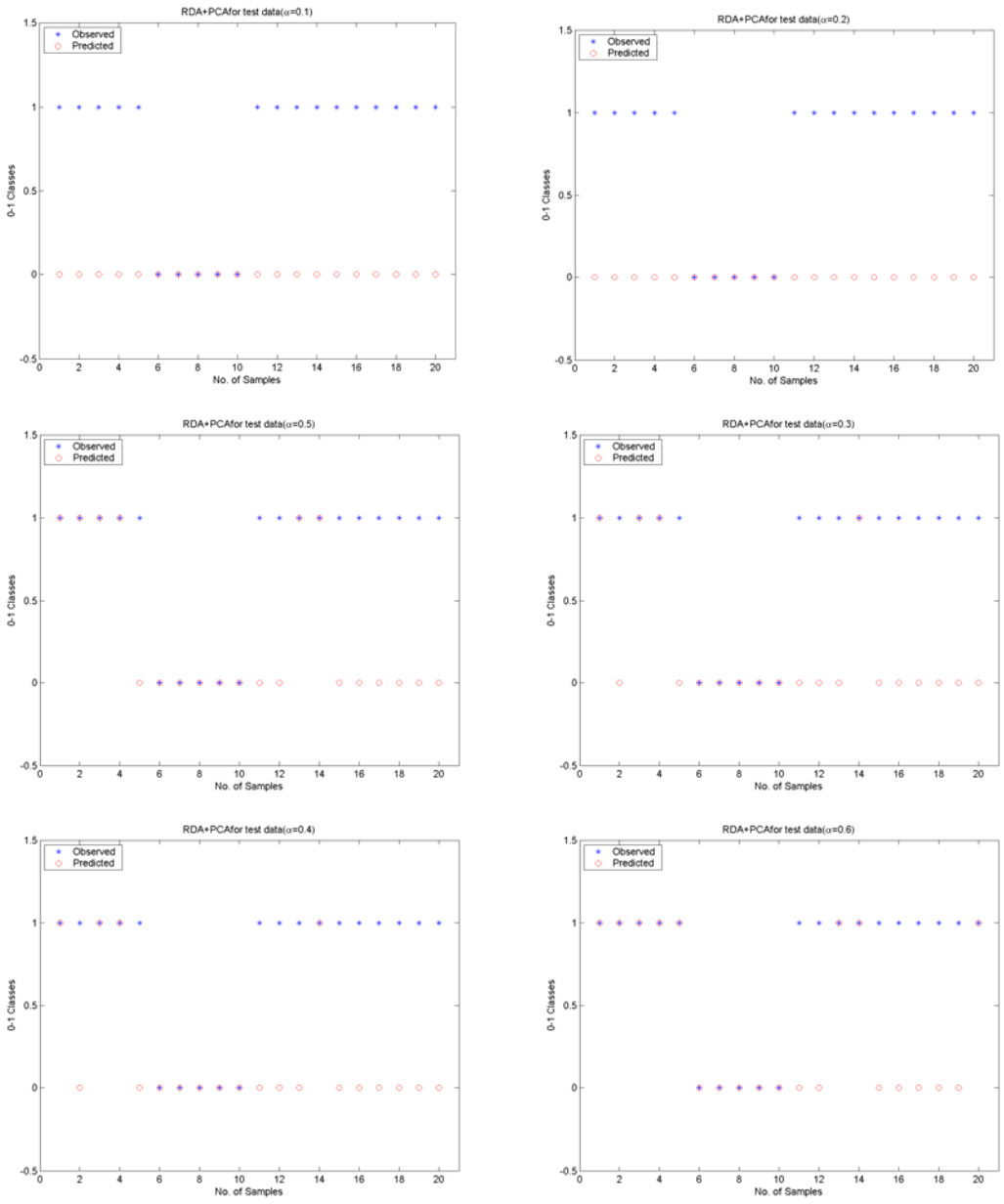

Figure 17 "Target" classification by RDA+PCA on the validation dataset with different α (α =0.1, …, 0.9) and fixed *k* (*k*=15). 0="Target", 1="Non-target".

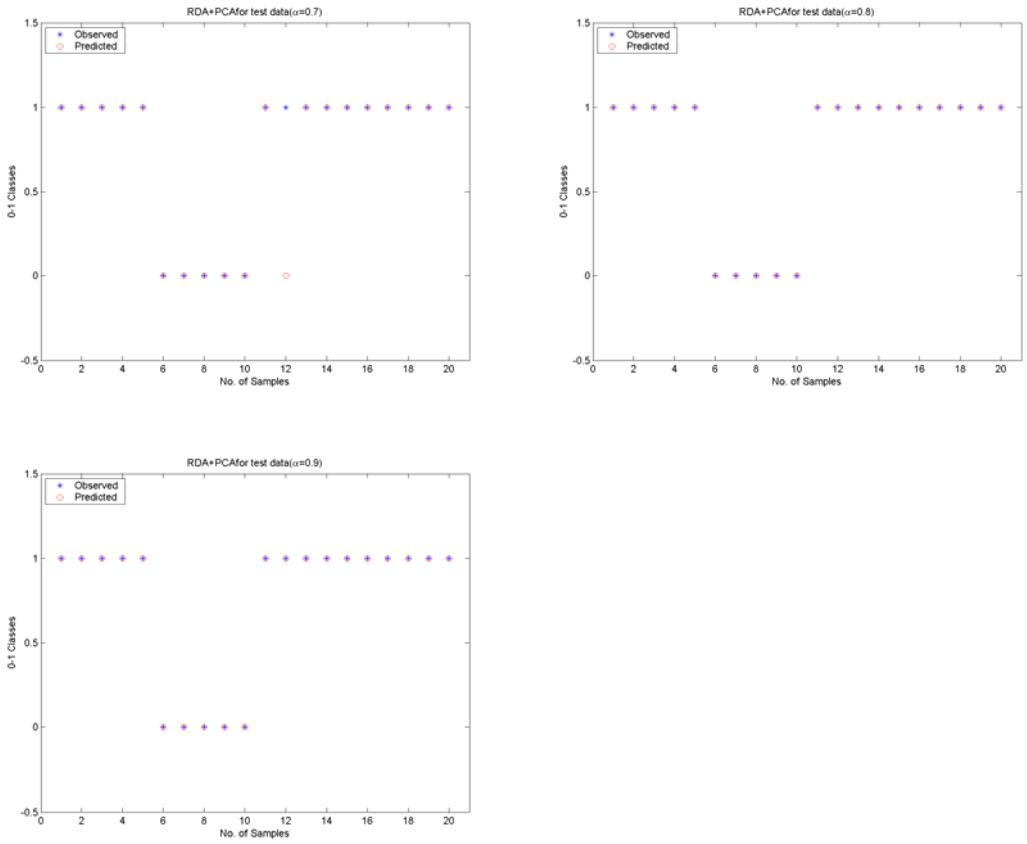

Figure 17 continued

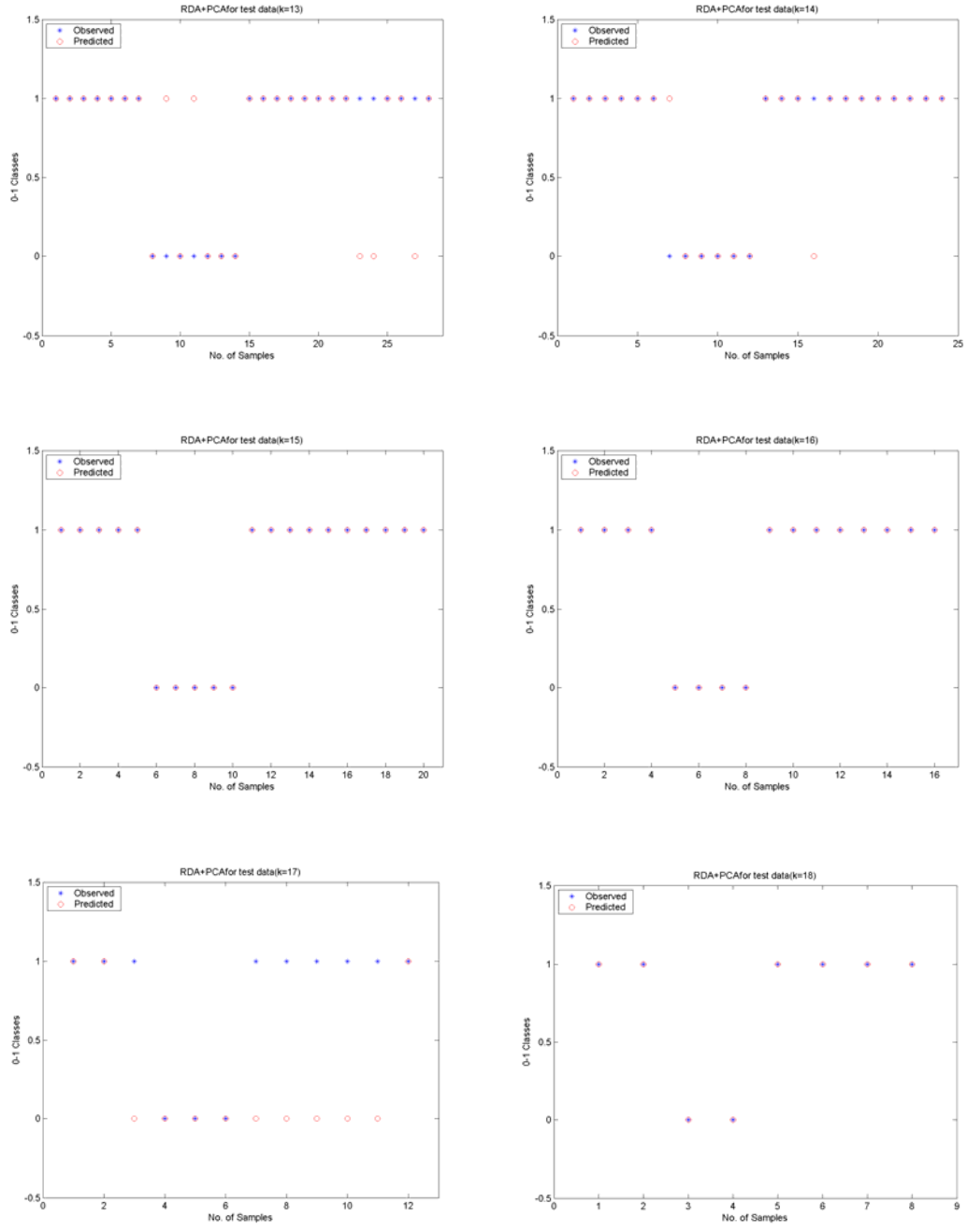

Figure 18 "Target" classification by RDA+PCA on the validation dataset with different *k* (k=13,…,19) and fixed α (α =0.9). 0="Target", 1="Non-target".

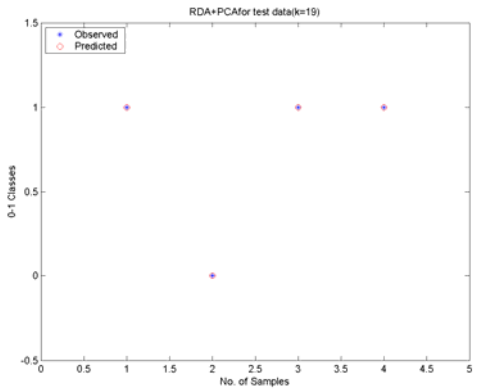

Figure 18 continued

Table 3 contains the misclassification results on the validation dataset obtained by conventional PLS. In contrast to the RDA+PCA, the PLS+PCA cannot achieve perfect discrimination results on the data, no matter how much training data is available.

4.4 Conclusion

In this chapter, the fusion approach of RDA and PCA is presented. The proposed approach achieved highly accurate classification results on the illustrated example, which shows potential to be a good solution to the SSS problem. By recalling the procedures of PCR, PLS and hybrid models discussed in chapter 2, one can see similarity between those approaches and RDA+PCA. Both of them consist of two steps: the first is to simplify the original data by certain dimensionality reduction techniques; the second is to perform regression or classification analysis based on the reduced-rank data. Following this line of research, it is possible to propose other multivariate predictive models for SSS problems.

Table 3 The number of misclassifications by conventional PLS with different K

K	13	14	15	16	17	18	19
Misclassification	7	6	5	4	3	2	1

5 Conclusions

In this thesis, certain multivariate predictive approaches based on latent variable models were proposed to attack the small sample size (SSS) problem, one of the most challenging problems in the data mining area. The proposed methods were validated by the case studies with spectral data, which belong to the class of SSS problems.

The proposed approaches include hybrid models of PCR and PLS and the fusion approach of RDA and PCA built on the generalized definition of latent variable models. They have much in common in that both of them consist of two steps as discussed in Section 4.4: dimensionality reduction and then conventional regression/classification procedures. Their major difference exists in the dimensionality reduction component, where hybrid models perform a mix of PCA and PLS and the latter only projects the data by pure PCA. The hybrid models of PCR and PLS show satisfactory performance when applied to a regression problem. The RDA+PCA was assessed on a classification problem. Although either of RDA or PCA is widely used in chemometrics, the fusion of them is rarely explored in this area. The case study showed that the proposed approach can achieve 100% prediction accuracy if sufficient data are available. In contrast, the conventional PLS approach showed poor performance on the dataset.

By the narrower definition of latent variable models in the statistics community, the robust probabilistic multivariate calibration (RPMC) model was developed in this thesis. The RPMC was intended to overcome the lack of robustness of the linear Gaussian

models by adopting the Student-t distribution as the distribution of noises and latent variables instead of the Gaussian distribution. It turns out that the RPMC includes some latent variable models as special cases, such as probabilistic PCA (PPCA) and supervised probabilistic PCA (SPPCA).

All the proposed methods in this thesis are basically linear predictive models, which could work particularly well on data that have inherently linear or near-linear properties. Many engineering problems present highly nonlinear patterns, or even chaos. Therefore, much emphasis of future research could be placed on building nonlinear models. A possible solution is based on the Kernel Trick (Scholkopf et al, 1998), which has been proven an efficient approach to deal with nonlinear problems. Kernel PCR and Kernel PLS have been proposed recently and achieve good prediction results (Rosipal, 2001). Thus, it is expected that the Kernel hybrid models of PCR and PLS could work well on some nonlinear cases. Furthermore, a Kernel version of the RPMC is also possible because one important observation in the dual solution is that all the calculation involving input data X can be done via inner product, e.g., in the Gram matrix $\mathbf{K}$ we have $\mathbf{K}_{ij} = \mathbf{x}_i^{'}\mathbf{x}_j$. This motivates us to consider non-linear RPMC where we first map the data into a new feature space (via, e.g., basis functions), and then perform PCA in that space with a proper definition of inner product.

References

1. Archambeau, C. (2005). Probabilistic models in noisy environments and their application to a visual prosthesis for the blind. Doctoral dissertation, Universit´ecatholique de Louvain, Belgium.
2. Archambeau, C., Delannay, N., Verleysen, M. (2006). Robust Probabilistic Projections In W. W. Cohen and A. Moore (Eds.), Proceedings of the 23rd International Conference on Machine Learning (ICML), pp. 33-40. ACM.
3. Bach, F. R., and Jordan, M. I. (2005). A probabilistic interpretation of canonical correlation analysis. Technical Report 688. Department of Statistics, University of California, Berkeley.
4. Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2004). Prediction by supervised principal components. Technical report. Department of Statistics, University of Stanford.
5. Bennett, K. P. and Embrechts, M. J. (2003). An optimization perspective on partial least squares, In Proceedings of the NATO Advanced Study Institute on Learning Theory and Practice, IOS Press Amsterdam, 190, 227-250.
6. Chatterjee, S. and Price, B. (1977). Regression Analysis by Examples. Wiley and Sons, New York and Toronto.
7. De la Torre, F., and Black, M. J. (2001). Robust principal component analysis for computer vision. Int. Conf. on Computer Vision, 362-369.
8. Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391–407.

9. Defo, M., Taylor, A.M., and Bond, B. (In press). Determination of moisture content and density of fresh-sawn red oak lumber by near infrared spectroscopy. Forest Product Journal. (Accepted)
10. Duda, R.O., Hart, P.E., and Stork, D.G. (2000). Pattern Classification. Wiley. Educational Psychology, 24, 417–441.
11. Engelbrecht, Andries. (2002). Computational Intelligence: An Introduction. John Wiley and Sons Ltd.
12. Everitt, B.S. (1984). An Introduction to Latent Variable Models. London: Chapman & Hall.
13. Evgeniou, T. and Pontil, M. (2004). Regularized multi-task learning. In Proceedings SIGKDD.
14. Fang, Y., Cho, H., and 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, 553-557.
15. Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. Journal of Machine Learning Research, 5(Jan):73–99.
16. Hastie, T., Tibshirani, R., and Friedman, J. (2001). The Elements of Statistical Learning. Springer Verlag.
17. Hotelling, H. (1936). Relations between two sets of variables. Biometrika, 28:321-377

18. Jain, A.K. and Zongker, D. (1997). Feature Selection: Evaluation, Application, and Small Sample Performance, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, pp. 153-158
19. Jolliffe, I. T. (2002). Principal Component Analysis. Springer Verlag.
20. Kim, H. and Ghahramani, Z. (2003). The EM-EP algorithm for Gaussian process classification. In Proceedings of the Workshop on Probabilistic Graphical Models for Classification at ECML.
21. Kim, H., Howland, P., and Park, H. (2005). Dimension reduction in text classification with support vector machines. Journal of Machine Learning Research, 6:37-53.
22. Lewis, D. D., Yang, Y., Rose, T., and Li, F. (2004). RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research, 5:361-397.
23. Liu, C., and Rubin, D. B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. Statistica Sinica, 5, 19-39.
24. Miller, R. G. (1986). Beyond Anova: Basics of Applied Statistics. John Wiley.
25. Minka, T. P. (2001). A family of algorithms for approximate Bayesian inference. PhD thesis, Massachusetts Institute of Technology.
26. Neal, R. M., and Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.), Learning in graphical models, Kluwer, 355-368.

27. Peel, D., and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. Statistics and Computing, 10, 339-348.
28. Pudil, P., Novovi˘cov´a, J., and Kittler, J. (1994). Floating search methods in feature selection. Pattern Recognition Letters, 15(11), 1119-1125.
29. Rosipal, R., and Trejo, L.J. (2001). Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. Journal of Machine Learning Research. 2, 97.
30. Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear Gaussian models. Neural Computaion, 11(2):305-345.
31. Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science, pages 2323-2326.
32. Roweis, S. T. (1998). EM algorithms for PCA and SPCA. Advances in Neural Information Processing Systems.
33. Scholkopf, B. and Smola, A. J. (2002). Learning with Kernels. MIT Press.
34. Scholkopf, B., Smola, A., and Muller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10, 1299-1319.
35. Stone, M. (1977). Asymptotics for and against cross-validation, Biometrika, 64, 29-35.
36. Svensen, M. and Bishop, C. M. (2004). Robust Bayesian mixture modeling. Neurocomputing 64, 235–252.
37. Tipping, M. E., and Bishop, C. M. (1999a). Mixtures of probabilistic principal component analyzers. Neural Computation, 11, 443-482.

38. Tipping, M. E., and Bishop, C. M. (1999b). Probabilistic principal component analysis. Journal of the Royal Statistical Society B, 61, 611-622.
39. Wentzell, Peter D., and Vega Montoto, Lorenzo. (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.
40. Wold, H. (1975). Soft modeling by latent variables; the nonlinear iterative partial least squares approach. Perspectives in Probability and Statistics, Chapters in Honour of M.S. Bartlett.
41. Ye, J., Janardan, R., Li, Q., and Park, H. (2004). Feature extraction via generalized uncorrelated linear discriminant analysis. In Proceedings of the Twenty-First International Conference on Machine Learning.

VITA

Yi Fang was born in Wuhan, Hubei, P. R. China. He had completed all his education from elementary school to Bachelor's degree in the same city. In the summer of 2005, he was admitted by the University of Tennessee, Knoxville as a graduate student, and awarded the Hilton-Smith Graduate Fellowship.