Journal of Systems Science and Information
Apr., 2024, Vol. 12, No. 2, pp. 212–228
DOI: 10.21078/JSSI-2023-0144

Principal Model Analysis Based on Bagging PLS and PCA and Its Application in Financial Statement Fraud

Xiao LIANG
School of Economics and Management, Beijing University of Technology, Beijing 100124*, China*
E-mail: *liangxiao*22@*emails.bjut.edu.cn*

Qiwei XIE
School of Economics and Management, Beijing University of Technology, Beijing 100124*, China*
E-mail: *qiwei.xie@bjut.edu.cn*

Chunyan LUO
School of Economics and Management, Beijing University of Technology, Beijing 100124*, China*
E-mail: *luochunyan@bjut.edu.cn*

Liang TANG
Shanghai Tongtu Semiconductor Technology Co., Ltd, Shanghai 210203*, China*
E-mail: *liang.tang@hotmail.com*

Yi SUN*
School of Finance, Anhui University of Finance & Economics, Bengbu 233000*, China*
E-mail: *sunyi@aufe.edu.cn*

Abstract Motivated by the Bagging Partial Least Squares (Bagging PLS) and Principal Component Analysis (PCA) algorithms, a novel approach known as Principal Model Analysis (PMA) method is introduced in this paper. In the proposed PMA algorithm, the PCA and the Bagging PLS are combined. In this method, multiple PLS models are trained on sub-training sets, derived from the training set using the random sampling with replacement approach. The regression coefficients of all the sub-PLS models are fused in a joint regression coefficient matrix. The final projection direction is then estimated by performing the PCA on the joint regression coefficient matrix. Subsequently, the proposed PMA method is compared with other traditional dimension reduction methods, such as PLS, Bagging PLS, Linear discriminant analysis (LDA) and PLS-LDA. Experimental results on six public datasets demonstrate that our proposed method consistently outperforms other approaches in terms of classification performance and exhibits greater stability. Additionally, it is employed in the application of financial statement fraud identification. PMA and other five algorithms are utilized to financial statement fraud which concerned by the academic community, and the results indicate that the classification of PMA surpassed that of the other methods.

Keywords principal model analysis; partial least squares; principal component analysis; dimension reduction; ensemble learning; financial statement fraud detection

Received September 20, 2023, accepted March 1, 2024
Supported by the Beijing Municipal Social Science Foundation (SZ202210005004) and Beijing Natural Science Foundation (9242004)
* Corresponding author

1 Introduction

For qualitative analysis, abundant information is provided by high-dimensional datasets. However, not all measured variables contribute significantly to qualitative models[1]. In addition, a situation is required by traditional statistical methods where the number of variables is smaller than the number of samples. Otherwise, it will lead to the curse of dimensionality[2]. In order to solve these problems, it is necessary to perform dimensionality reduction on the dataset prior to qualitative analysis. Dimension reduction methods such as PCA[3–6], LDA[7] and PLS[8–10] are frequently employed for this purpose. These methods eliminate the statistical redundancy and noise among the components of high-dimensional vector data, obtaining a lower-dimensional representation without significant loss of information.

In unsupervised data analysis, PCA is recognized as a valuable dimension reduction tool. Its primary objective is to decrease the dimensionality of datasets with numerous interrelated variables while retaining as much as possible of the variation present in the dataset[11, 12]. However, PCA is solely applicable to unsupervised datasets. After adding the sample labels, analyzing the dataset requires employing supervised methods. LDA is widely recognized as a supervised method for feature extraction and dimension reduction, achieving maximum discrimination by maximizing the ratio of between-class and within-class distance[13]. An intrinsic limitation of classical LDA is the so-called small sample size problem[7], various methods have been proposed to solve this concern[14–16]. One of the most effective approaches is subspace LDA, which applies an intermediate dimension reduction stage before LDA. Among all the subspace LDA methods, significant attention has been given to PCA plus LDA (PCA-LDA[17]) and PLS plus LDA (PLS-LDA[18]). Other approaches use the algorithms based on PLS as a dimension reduction.

The capability to overcome both dimensionality and collinearity problems is possessed by the PLS algorithm[19, 20], Simultaneously, it has demonstrated exceptional performance for solving the problem of small sample size[21, 22]. Despite being originally a supervised method, PLS has been extended to the unsupervised field. Unsupervised partial least squares (UPLS) method was proposed by Jin to summarize the structure of unlabeled samples, which utilizes the mean vector as PLS label vector to solve the regression matrix, and use the PCA method to select the principal models[23]. However, PLS remains confronted with several challenges, including the need to extract more meaningful information, enhance the model's robustness, and more effectively eliminate redundancy and noise. One solution to these problems is found in ensemble learning which is derived from the field of machine learning[24], and can be applied to tackle both classification and regression problems. In this study, the primary focus lies in dimensionality reduction and classification. Compared with the single model, ensemble models, including Boosting[25, 26], Bagging[27] and stacked regression[28, 29], have demonstrated heightened robustness and accuracy when compared to single models[30], and have found successful application in the last several years. In order to overcome the over-fitting problem, the concept of Boosting was introduced by Zhang, et al. through the amalgamation of a collection of PLS models, each with only one PLS component, and called it Boosting PLS[26]. On the basis of Boosting PLS, several researchers have modified and applied it for spectroscopic quantitative analysis[31, 32]. Numerous training sets are generated from the original data set by using the Bagging strategy[33, 34]. Bagging PLS trains a model from each of those training sets, and

the final model can be obtained by averaging the coefficient B from each sub-model. From overcoming the disadvantages of Moving-window PLS (MWPLS) and interval PLS (iPLS), a stack-based PLS method using Monte Carlo Cross-validation was presented by Xu, et al[35]. Ni, et al. have proposed two new stacked PLS which can establish PLS models based on all intervals of a set of spectra to take advantage of the information from the whole spectrum by incorporating parallel models in a way to emphasize intervals highly related to the target property[29].

Following the creation of the PLS sub-models, a range of ensemble algorithms for the fusion of the final model are available, mainly including average weighting, cross-validation error weighting and minimum square error weighting rule and so on. In this paper, for adopting Bagging model training method, the data set is divided into a number of sub-training sets. The PLS models are then employed on these sub-training sets. Subsequently, the coefficients B of all the PLS sub-models becoming an asymmetric positive semi-definite matrix BB^{T}, are fused in a joint matrix. Finally, using the PCA, an eigenvalue decomposition by taking the largest variance model or final projection model is performed. This proposed method is termed as the Principal Model Analysis (PMA). In the subsequent sections, the relationship between the model parameters (the number of latent variables, models and remained dimensions) and the classification accuracy is discussed. It is indicated by the theory and experimental evidence that the robustness and generalization ability of the PLS algorithm are enhanced by PMA. Furthermore, valuable insights into the application of the PLS algorithm in semi-supervised dimensionality reduction are offered by PMA.

Financial statement fraud is categorized as falling under financial fraud. It refers to the management that deliberately hide the true financial situation and modify the financial statements of the behavior to obtain a certain benefit, such as tax evasion to obtain a higher salary[36, 37]. A variety of investors, employees, as well as governments, are being plagued by financial fraud, which eventually leads to serious economic and social problems such as bankruptcy[38–40]. Besides, with the advancement of technology, fraud is increasingly being made more sophisticated and systematic. Therefore, significant emphasis has been placed on research related to financial fraud, which holds great practical significance. As the rapidly development of modern financial technologies such as machine learning methods[41, 42], a growing trend is observed in the construction of prediction models. Additionally, given that fraud prediction involves a binary classification problem, machine learning methods are particularly well-suited for their application.

In previous studies, some classification methods were incorporated into the research. Due to the excellent strength of variable extraction, PLS and PCA were frequently employed for the selection of fraud indicators. In fraud prediction, the combination of partial least squares regression (PLS) with SVM (support vector machine) was utilized by Li[43], confirming more significant effects of this method than other traditional data analysis methods due to the components extracted through PLS. Recently, Bagging method has also been employed in order to solve the problem of small sample. The Bagging classifier was used by Zareapoor[44] to train data mining methods, and the result performed the benefit of Bagging strategy. Since simple averaging or simple voting weighting method of Bagging strategy doesn't effectively reduce

dimensionality, the PMA algorithm is utilized to address these limitations.

In the empirical test below, it is observed that PMA not only solves the problem of Bagging, but also high accuracy is achieved when applied to small samples. Given that financial fraud datasets are typically characterized by limited sample sizes, PMA emerges as a more effective tool for fraud detection.

The objective of this research is to provide a novel approach that combines the Bagging strategy and PCA method, presenting a new model applicable not only to financial fraud detection but also to other real-world scenarios.

The structure of the remainder of the paper is as follows: The subsequent section briefly introduces the methodology of PLS, PCA and Bagging based PLS, and elaborates on the theory of PMA, along with discussing the factors that may affect its accuracy. Section 3 compares the PMA algorithm with other traditional dimension reduction methods, and presents the results of implementing experiment above. Section 4 demonstrates the superior accuracy of PMA in the application in financial fraud detection, while Section 5 presents the results and offers insights into potential future research directions.

2 Methods

2.1 Notation

Boldface uppercase and lowercase letters are used to denote matrices and vectors, respectively. Lowercase italic letters denote the scalars. The detailed notations are as follows:

X	$n * k$ matrix of samples
y	$n * 1$ vector of sample label
C	$k * k$ covariance matrix of PCA
w	$k * 1$ vector of the PCA loading
B	$k * p$ matrix of PLS regression coefficients
β	$k * 1$ vector of PLS1 regression coefficient
λ	the eigenvalue of PCA
n	number of samples
k	number of sample features
m	number of components

2.2 PLS and Bagging-based PLS

PLS is intended to project the high-dimensional predictor variables into a smaller set of latent variables, which have a maximal covariance to the responses. Given a training set (X, y), the decomposition of PLS algorithm is as follows[45–48]:

$$X = \sum_{i=1}^{a} t_i p_i^{\mathrm{T}} + E, \quad y = \sum_{i=1}^{a} u_i q_i^{\mathrm{T}} + F,$$

where t_i and u_i are score vectors, and p_i and q_i are loading vectors of X and y, respectively. E and F are residuals matrices. The a is the number of feature vectors. The PLS inner relation

between the projected score vectors is:

$$u_i = \beta_i t_i.$$

The detailed algorithm procedures of PLS are as follows:

1. Initialization: $E_0 = X, F_0 = y, i = 1$.
2. Computing weight vector: $w_i = E_{(i-1)}^{\mathrm{T}} F_{(i-1)}$, and making w_i to be normalization.
3. Computing the input's score vector: $t_i = E_{(i-1)} w_i$ and its loading vector: $p_i = \frac{E_{(i-1)}^{\mathrm{T}} t_i}{t_i^{\mathrm{T}} t_i}$.
4. Computing the output's loading vector: $q_i = \frac{F_{(i-1)}^{\mathrm{T}} t_i}{t_i^{\mathrm{T}} t_i}$ and making q_i to be normalization.
5. Computing the outputs score vector: $u_i = F_{(i-1)} q_i$.
6. Computing internal regression coefficient: $\beta_i = \frac{u_i^{\mathrm{T}} t_i}{t_i^{\mathrm{T}} t_i}$.
7. Computing residuals matrices: $E_i = E_{(i-1)} - t_i p_i^{\mathrm{T}}$ and $F_i = F_{(i-1)} - \beta_i t_i q_i^{\mathrm{T}}$.
8. Updating i to $i+1$, then go back Step 2 until the expected number of latent variables is achieved.

The accuracy and robustness of traditional algorithms are sought to be enhanced by the ensemble learning method, which combines the results of multiple sub-models. Bagging, a simple ensemble learning strategy, is extensively employed in addressing classification and regression problems, exemplified by Bagging SVM and Bagging PLS.

The general PLS method is usually found to produce bad or unstable results on the data with a very large number of collinear x-variables or the data with very limited training samples. By using the Bagging strategy, the Bagging PLS model could reduce the variance of the original unstable model without increasing the bias. Therefore, much more accurate and stable results can usually be achieved by the Bagging PLS compared to the traditional PLS method.

First, several sub-training sets are generated from the original data set by the Bagging-based PLS, based on the random sampling with replacement method. Then, a PLS model is trained on each sub-training set separately, and finally, the regression coefficients of all sub-PLS models are averaged, and the averaged regression coefficient is used for the model prediction. In detail, it is assumed that N sub-training sets are generated by random sampling with replacement, and the PLS regression coefficient vector corresponding to each sub-training set is β_i $(i = 1, \cdots, N)$. The final regression coefficient of Bagging-based PLS can be formulated as:

$$\beta = \frac{1}{N} \sum_{i=1}^{N} \beta_i.$$

2.3 PCA

An orthogonal transformation is employed by Principal Component Analysis (PCA) to convert a number of possibly correlated variables into a smaller number of uncorrelated variables known as principal components, with an attempt made to preserve the data variance. Given a data matrix X, computing the covariance matrix C, then the projection directions of PCA can be solved by:

$$\widehat{w} = \arg \max_{\|w=1\|} w^{\mathrm{T}} C w.$$

The above problem can be easily solved by the eigendecomposition methods, such as the singular-value decomposition (SVD) algorithm. The detailed algorithmic processes of PCA are as follows[49–52]:

1. Data standardization:

$$x_{ij}^{*} = \frac{x_{ij} - \bar{x}_j}{s_{ij}}, \quad s_{ij} = \sqrt{\frac{\sum_{j=1}^{m} (x_{ij} - \bar{x}_j)^2}{s_{ij}}},$$

where $X = (x_{ij})_{n*p}$ is the data matrix with n samples and p variables.

2. Computing the covariance matrix: $C = X^{\mathrm{T}}X$.

3. Eigendecomposition: $Cu = \lambda u$.

4. Denote the first m eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, their corresponding eigenvectors, $u_1, u_2, \cdots, u_m$, are principal components. The number of principal components m can be decided by the cumulative contribution rate of the principal components, i.e., choosing m such that

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{p} \lambda_j} \geq 0.95.$$

2.4 Principal Model Analysis

2.4.1 Theory and Algorithm

Combing the bagging strategy and PLS, we propose a principal model analysis (PMA) method in this section. The proposed PMA contains two steps. The first step is also to generate N sub-training sets from the original training set with replacement method and the corresponding PLS regression coefficient vector of each sub-training set is denoted by β_i ($i = 1, \cdots, N$). Different from the bagging PLS method which just simply averages the PLS sub-models, the second step adopted here is to use the PLS sub-models as the input of PCA algorithm to generate the final PMA model. This approach is chosen primarily because PCA can effectively find the "major" elements, remove the noise, and reveal the essential structure hidden behind the complex data. The flowchart is shown in Figure 1.

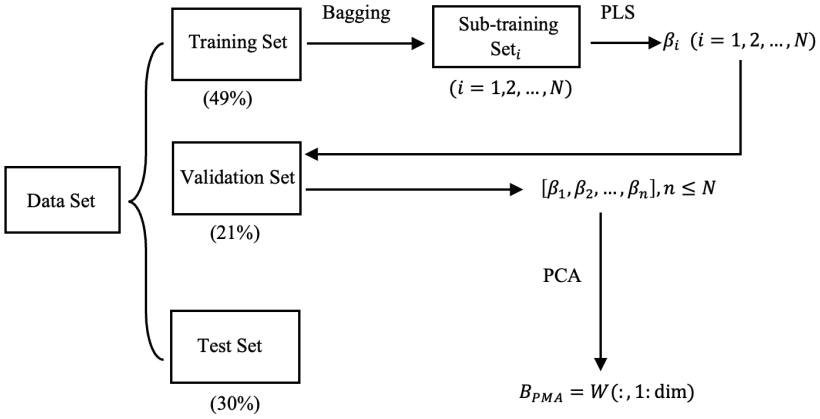

Figure 1 Flowchart of the proposed PMA algorithm

The original PCA algorithm is performed by decomposing the covariance matrix C, which is a symmetric and positive semi-definite matrix. However, the whole regression coefficient

matrix $B = [\beta_1, \cdots, \beta_n]$ in the PMA algorithm is not a symmetric and positive semi-definite matrix. In order to address this discrepancy, the regression coefficient matrix B is transformed into a symmetric positive semi-definite matrix by replacing it with BB^{T} for the eigenvalue decomposition. Through this process, the most representative models, known as the principal model, are obtained as the final PMA model. The optimization of PMA algorithm can be expressed as follows:

$$\begin{cases} B = [\beta_1, \cdots, \beta_n] \, [\beta_1, \cdots, \beta_n]^{\mathrm{T}}, \\ \widehat{w} = \arg \max\limits_{\|w\|=1} w^{\mathrm{T}} B w. \end{cases} \tag{1}$$

The above problem can be easily solved by the singular-value decomposition (SVD) algorithm:

1. Eigendecomposition: $Bu = \lambda u$.

2. Denote the first m eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, their corresponding eigenvectors, $u_1, u_2, \cdots, u_m$, are principal components. The number of principal components m can be decided by the cumulative contribution rate of the principal components, i.e., choosing m such that

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{p} \lambda_j} \geq 0.95.$$

The detailed processes of PMA method are shown as follows.

Algorithm PMA

Input: Training set and corresponding label vector, the number of PLS latent variables, the number of sub-models, the number of principal models (dim).

Output: The projection direction of PMA.

1. Preprocessing the training set T.

2. Dividing T using random sampling with replacement and generating PLS sub-models β_i.

3. Denote $B = [\beta_1, \cdots, \beta_n][\beta_1, \cdots, \beta_n]^{\mathrm{T}}$, and doing the eigenvalue decomposition in (3), sorting the eigenvalues in descending order and rearranging their corresponding eigenvectors.

4. Denote the rearranged eigenvector matrix as W, outputting the final PMA model $B_{PMA} = W(:, 1 : \mathrm{dim})$.

2.4.2 Determination of the Number of Latent Variables

The number of latent variables is an important parameter in the PLS model. The determination of the number of latent variables can be achieved through various approaches, such as genetic algorithm, F-test and cross-validation methods. Cross-validation methods include K-fold cross-validation (k-CV), leave-one-out cross-validation (LOOCV), Monte Carlo cross-validation (MCCV), and so on. In this paper, 10-fold cross-validation method is used to determine the number of latent variables.

2.4.3 The Sub-Models Selective Rule

For ensemble strategy, it's common practice to include sub-models that exhibit superior performance or provide diversity[33]. Therefore, it has been suggested by Zhou, et al. that using

part of sub-models instead of all of the them may be a better choice[53]. Herein, original data set is arbitrarily divided into three parts: Training set, validation set and test set[54]. Within the training sets, 100 PLS sub-models are established by employing sub-sampling and re-weighting of the existing training samples, respectively. The proposed method directly constructs diverse models using virtual samples derived from the training sets. This approach is proven beneficial in increasing ensemble diversity when the training samples are not enough[54]. By applying these 100 sub-models to the validation set, we get 100 different classification accuracies. Subsequently, we select the sub-models with the highest classification accuracy, following a descending order, to participate in the final ensemble.

2.4.4 Determination of Dimensions

Performing an EIG or SVD on a matrix is involved in PCA, resulting in the generation of an eigenvalue matrix. To select the principal components, only the first few eigenvalues have to be considered. How do we decide on the number of eigenvalues to retain from the eigenvalue matrix? Typically, the accumulative contribution rate is utilized to automatically retain useful eigenvalues.

When utilizing PMA for dimension reduction, scores are obtained by projecting the new samples onto the direction of the principal models. Consequently, the number of the final dimensions is equal to the number of selected principal models. In the experiment, if the parameter of fixed dimensions, which is one of the inputs, is greater than or equal to 1, the fixed dimensions are employed to select the principal models. Conversely, if the fixed dimensions are greater than 0 and less than 1, the cumulative contribution rate is utilized to determine the principal models.

From the selection of sub-models can be inferred, it can be inferred that a large number of final principal models may not be necessary. This conclusion arises from the fact that the selected sub-models exhibit nearly identical classification capabilities. Consequently, the classification abilities of all the sub-models can be effectively retained by just one principal model. In practical applications, it suffices to consider only the eigenvector with the largest eigenvalue as the principal model.

3 Experiment

3.1 Data Sets

In order to evaluate the performance of the proposed PMA method, we compare it with the PLS, LDA, PLS-LDA and Bagging PLS methods on three types of datasets:

1. Four UCI datasets, i.e., Breast data, Spambase data, Gas data, Musk data (Version 1) (obtained from http://archive.ics.uci.edu/ml/datasets.html);

2. Small data and Imbalanced data;

3. Raman spectral data (Raman).

The details of these datasets are shown in Table 1. Before using the datasets, the non-numerical and missing input data is removed, and the class label is converted to a numeric type.

Table 1 Data sets

Data Set	Number of Examples	Number of Attributes	Class label	Year
Breast	569	30	1 and 2	1995
Spambase	4601	57	0 and 1	1999
Gas	4782	128	5 and 6	2012
Musk (Version 1)	7074	166	0 and 1	1994
Small	70	166	0 and 1	1994
Imbalanced	350	166	0 and 1	1994
Raman	925	101	0 and 4	N/A

The datasets "Small" and "Imbalanced" are randomly sampled from the data set "Musk (Version 1)". The data set "Small" is a typical data set with high dimensionality and small samples, where the number of positive and negative samples are the same. The dataset "Imbalanced" is an imbalanced data set, where the ratio of positive and negative samples is $6:1$.

Spectral data set "Raman" is obtained by a standard Raman spectroscope (HR LabRam invers, Jobin-Yvon-Horiba). The excitation wavelength of the used laser (Nd: YAG, frequency doubled) is centered at 532 nm. There are 2545 spectra for 20 different strains available[55]. Herein two classes (B. subtilis DSM 10 and M. luteus DSM 348) are selected, and the spectra in the region 1100∼1200 are used in calculations[56].

3.2 Calculation

Five-dimension reduction methods, i.e., PLS, LDA, PLS-LDA, Bagging PLS and PMA, are compared in our experiments. For Bagging PLS algorithm, fifteen models are generated by the random sampling with replacement method and the final model is obtained by averaging these fifteen sub-models. For the PMA method, 100 sub-models are generated from the training set, and the best fifteen sub-models with higher accuracies are chosen to perform model fusion. Except for the LDA, the number of latent variables in the PLS, PLS-LDA, Bagging PLS and PMA are determined by the 10-fold cross-validation. In the experiment, the dimensionality of the original data is reduced to 1. For fair comparison, the linear Naive Bayes classifier is used to evaluate the results of the above different dimension reduction methods.

For each data set, we randomly choose 49%, 30% and 21% samples from the total samples to form the training set, test set, and validation set. The experiments are randomly run 20 times, and the averaged results are recorded.

All codes are written in Matlab Version R2021a (Math Works, Inc.) and run on a personal computer, APPLE Macbook Air 2020 with an APPLE M1 processor, 8GB RAM, and macOS Monterey 12.4 operating system.

3.3 Results and Discussion

3.3.1 Classification Performance of Different Algorithms

The classification performance of various algorithms is primarily investigated in this section. The results are reported for both the training and test datasets. The classification accuracies are reported in Table 2 and Table 3, respectively.

Table 2 Train classification accuracy of different comparative algorithms

Data Set	PLS	LDA	PLS-LDA	Bagging PLS	PMA
Breast	0.9201	0.6781	0.8908	0.9306	**0.9636**
Spambase	0.7976	0.7791	0.6823	0.853	**0.9076**
Gas	0.9705	0.9716	0.7917	0.9704	**0.9741**
Musk (Version 1)	0.9107	0.7174	0.9024	0.9115	**0.9166**
Small	0.9080	0.9009	0.9126	0.9323	**0.9854**
Imbalanced	0.9641	**1.0000**	0.9483	0.9715	0.9892
Raman	0.9516	0.7722	0.853	0.955	**0.9583**

Note: The bold value means the maximum accuracy among different methods.

Table 3 Test classification accuracy of different comparative algorithms

Data Set	PLS	LDA	PLS-LDA	Bagging PLS	PMA
Breast	0.9159	0.6501	0.8889	0.925	**0.9549**
Spambase	0.7942	0.7657	0.6817	0.8499	**0.9036**
Gas	0.97	0.9614	0.7924	0.9701	**0.9734**
Musk (Version 1)	0.9047	0.6994	0.8983	0.9054	**0.9106**
Small	0.6114	0.5681	0.6157	0.6229	**0.6233**
Imbalanced	0.8809	0.6085	0.8596	0.8921	**0.9069**
Raman	0.9335	0.6729	0.8399	0.9368	**0.9407**

Note: The bold value means the maximum accuracy among different methods.

The small-sample-size problem is often encountered in the field of pattern recognition, which may lead to the singularity of the within-class scatter matrix in the LDA. Consequently, for the datasets characterized as "Small" and "Imbalanced", poor results are shown by the LDA algorithm. Good overall classification performance is demonstrated by PLS. PLS-LDA algorithm, which firstly removes redundancy and noise in the dataset by the PLS method, and then performs the LDA algorithm on the PLS dimension reduction features, is seen to show better results than LDA, except for the datasets "Spambase" and "Gas". However, PLS-LDA still seems to show over-fitting phenomenon in the datasets "Small" and "Imbalanced". Because the PLS dimension reduction process may lose some information, the results of PLS-LDA are worse than PLS. Better results than PLS are achieved by Bagging PLS in the datasets "Breast", "Spambase", "Raman" and "Musk (Version 1)". Although many sub-models in Bagging PLS provide better performance than PLS, the improvement of Bagging PLS over PLS is not significant because of the average strategy. As observed in Table 2 and Table 3, an overfitting phenomenon is observed in all algorithms on the "Small" dataset. However, the best results in both training and test sets, except for the "Imbalanced" dataset, are achieved by the proposed PMA algorithm. Its superiority is particularly evident in datasets "Breast" and "Spambase."

Figure 2 is the box diagrams of the accuracy of different algorithms. For the datasets "Breast", "Spambase", "Raman" and "Musk (Version 1)", the results of LDA algorithm are obviously much worse than other methods. The results of PLS are also unstable on the data

set “Spambase”. The results of PLS-LDA are worse than others on the data sets “Spambase” and “Gas”. In the “Small” dataset, all algorithms show over-fitting phenomenon. Except for the “Small” dataset, PMA algorithm gets more stable results than other methods.

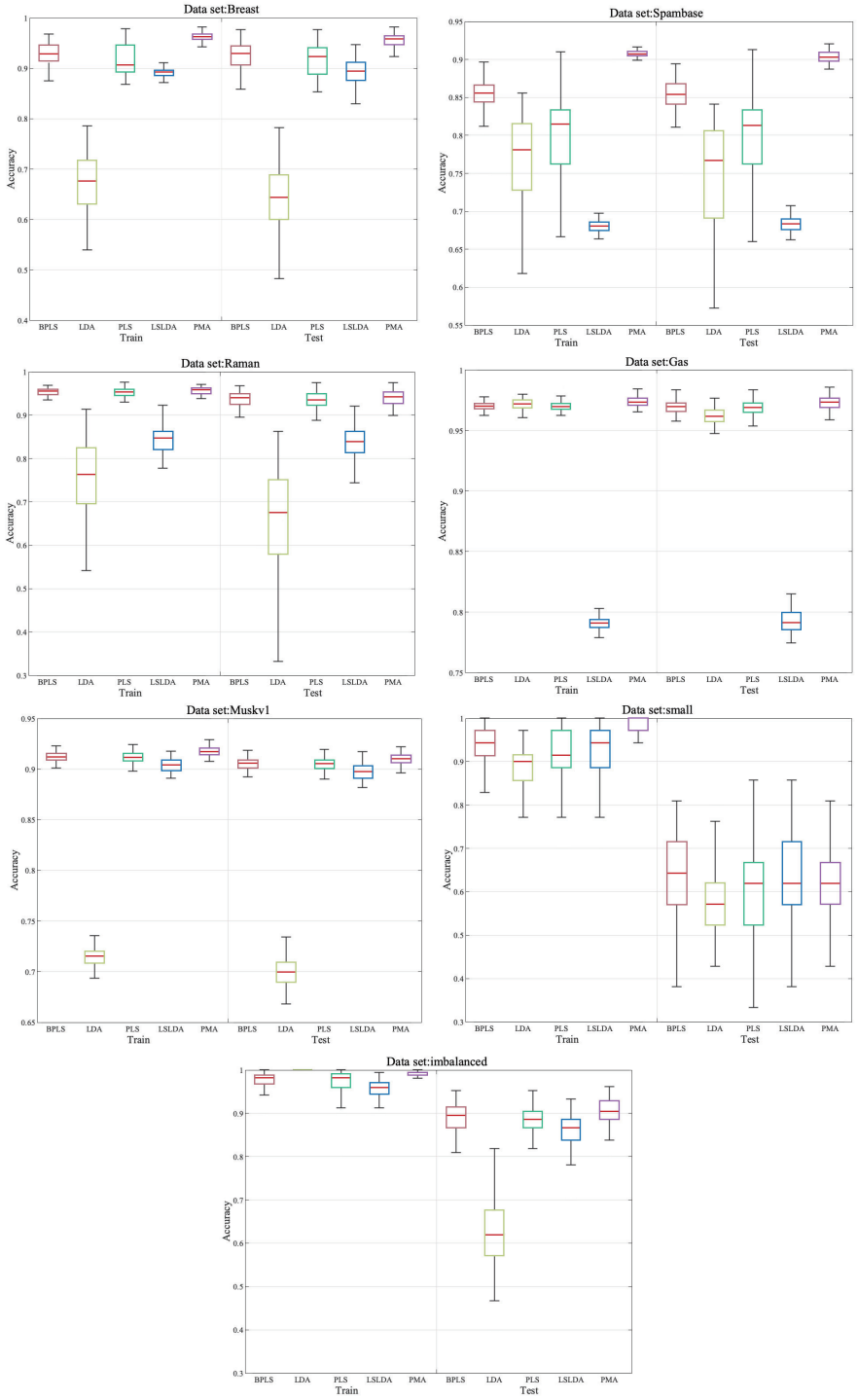

Figure 2 Classification accuracy box of data sets

3.3.2 Investigation on the Number of Sub-Models

From Figure 3, it can be observed that the number of sub-models is less sensitive to the PMA model. In general, the classification accuracies on each dataset basically tend to be stable with the change of the number of sub-models except for the datasets "Small" and "Imbalanced". It demonstrates that not all of the sub-models are valid. Meanwhile, it indicates the potential for enhancing classification performance by choosing some good sub-models. For the datasets "Breast" and "Raman", the classification results are significantly affected by the number of sub-models. The number of sub-models can be empirically determined by the cross-validation method.

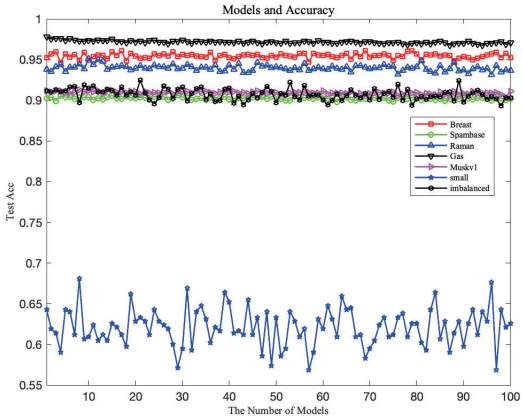

Figure 3 Classification accuracy vs. number of sub-model

3.3.3 Impact of PMA Dimensions

To investigate the effect of PMA dimensionalities, the classification results on different dimensionalities ranging from 1 to 30 are presented. As depicted in Figure 4, the classification accuracy across all datasets does not improve with the increase of dimensionality. A possible explanation could be that the first principal component already contains the majority information of the entire data. Notably, the results on the datasets "Gas" and "Muskv1" are relatively stable.

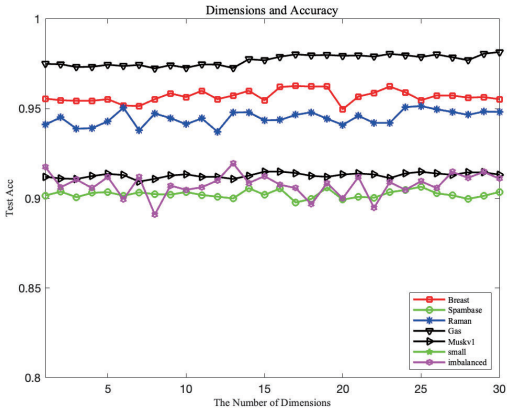

Figure 4 The impact of PMA dimensions

3.3.4 Discussions of the Proposed Method

The proposed PMA algorithm extends the original Bagging PLS for qualitative analysis. The results on the six datasets show that PMA algorithm can improve the classification accuracy to a certain extent. Various advantages are offered by model ensembles, such as enhancing the robustness. However, the number of sub-models and the number of dimensionalities must be carefully chosen.

4 Application in Financial Statement Fraud

4.1 Data

This research data used in this study was sourced from the CSMAR and the Wind databases in the period of 2012 to 2020. These company samples were committed by the China Securities Regulatory Commission (CSRC), the Shanghai Stock Exchange and the Shenzhen Stock Exchange for various irregularities, including "falsely listed assets", "fictitious profits", "false record (misleading statement)", "delayed disclosure", "material omission", "false disclosure". For the purposes of presentation, "financial statement fraud Company" and "non-financial fraud company" are also replaced by the terms "negative sample" and "positive sample" below. To obtain matched samples, considering different industries of different sizes are not comparable, non-fraud firms with the same year, size, and business characteristics were randomly selected into our dataset. The industry classifications are based on the standards issued by CSRC in 2012 and later. Finally, negative samples without matching positive samples are deleted. In total, 9178 samples are subsequently determined, of which 4589 are negative and positive samples respectively. These sample datasets are divided into training sets and test sets, with a 7:3 ratio between the training and test data.

After reviewing prior relevant literature[57–66], 21 financial and non-financial indicators were selected, as outlined in Table 4, among which indicators with more than 10% missing data were deleted.

Some of indicators contained missing values. To address this issue, the data was first processed using the Z-score normalization method. The processed data conforms to the normal distribution, and its transformation function form is:

$$y_i = \frac{x_i - \mu}{\sigma}, \quad i = 1, \cdots, n.$$

Then the KNN algorithm is used to impute the missing values. The KNN algorithm (K Nearest Neighbor) is a univariate method for handling missing data, employing a clustering method to estimate values. The missing data is replaced by the mean value of the column features, the nearest K records (the distance is represented by L) are calculated, and the mean value of K records is used to replace the missing value.

$$L = \sqrt{(x_1 - a_1)^2 + (x_2 - a_2)^2 + \cdots + (x_n - a_n)^2}.$$

It is necessary to establish evaluation criteria for the models before the comparison of financial fraud models. In binary classification, model samples are divided into positive and negative classes, with corresponding predictions falling into the same categories. However, not

Table 4 Financial items that used in the detection of financial statement fraud

No.	Financial items
1	total debt/total equity
2	total liabilities/total assets (TLTA)
3	the inventory/total assets
4	net profit/total assets (NP/TA)
5	Log (total assets)
6	the inventory/sales ratio (INV/SAL).
7	the sales/total assets ratio (SAL/TA)
8	net profit/sales ratio (NP/SAL).
9	Z-score
10	accounts receivable/sales ratio (REC/SAL)
11	gross profit/total assets (GP/TA)
12	current assets/total assets (CATA)
13	Working Capital/Total Assets (WCTA)
14	cash on sales ratio
15	current liabilities/total liabilities
16	total assets turnover
17	other receivables
18	whether the chairman and the general manager are combined
19	ownership concentration
20	balance of ownership
21	remuneration of company officers

all predictions are accurate. Accuracy measures the proportion of correct predictions to the total sample count, evaluating the model's correctness and suitability for financial statement fraud detection in academic research.

4.2 Results and Discussion

The final accuracy results are presented in Table 5. After a thorough comparison and analysis of each model, the primary conclusions are as follows: The PMA model is significantly better than other traditional financial fraud model based on PLS, LDA and other data mining algorithms. The accuracy of PMA in the training set and test set is 62.53% and 61.79%, which is higher than the simulation accuracy of other models, and has better stability and prediction accuracy.

Table 5 Classification accuracy of different comparative algorithm

	PLS	LDA	PLS-LDA	Bagging PLS	PMA
Training Set	0.6246	0.5036	0.6250	0.6250	**0.6253**
Test Set	0.6168	0.4979	0.6171	0.6180	**0.6179**

The results drawn from this section indicates that PMA algorithm holds significant value in predicting corporate financial fraud. This method will play an important role in preventing financial fraud, and will also challenge the traditional forecasting methods.

5 Conclusions

In this paper, a PMA method for classification has been proposed. By means of ensemble strategy, the proposed PMA method fuses the results of PLS sub-models and finds the principal model by performing PCA on the joint coefficient matrix of all sub-models. Experimental results indicate that the proposed PMA method outperforms the original PLS and Bagging methods in terms of classification performance. Furthermore, better results can be achieved by applying it in the detection of financial fraud. Our future work will focus on establishing more comprehensive evaluation criteria for the selection of sub-models. In addition, PMA will be applied to semi-supervised problems by adding a large number of unsupervised data.

References

[1] Pan J, Chai H F, Sun Q, et al. Joint feature screening method for ultrahigh dimensional censored data. Systems Engineering — Theory & Practice, 2023, 43(1): 169–190.
[2] Bellman R. Adaptive control processes: A guided tour. The University Press, 1961.
[3] Jolliffe I T, Cadima J. Principal component analysis: A review and recent developments. Philosophical Transactions, 2016, 374(2065): 201–202.
[4] Mendes-Moreira J, Soares C. Ensemble approaches for regression: A survey. ACM Computing Surveys, 2011, 45(1): 10.
[5] Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and Intelligent Laboratory Systems, 1987, 2(1): 37–52.
[6] Lin B, Song D, Liu Z Y. A model of aircraft support concept evaluation based on DEA and PCA. Journal of Systems Science and Information, 2018, 6(6): 563–576.
[7] Ginkel J R V, Kroonenberg P M. Using generalized procrustes analysis for multiple imputation in principal component analysis. Journal of Classification, 2014, 31(2): 242–269.
[8] Barker M, Rayens W. Partial least squares for discrimination. Journal of Chemometrics, 2012, 30(3): 446–452.
[9] Cheng L, Wang G Z. Soft sensing based on pls with iterated bagging method. Journal of Tsinghua University, 2008, 48: 86–90.
[10] Chen S, Peng J. Multivariate calibration with support vector regression based on random projection. International Journal of Wavelets Multiresolution and Information Processing, 2012, 10(6): 1250056.
[11] Liu Y, Rayens W. PLS and dimension reduction for classification. Computational Statistics, 2007, 22(2): 189–208.
[12] Zhou R X, Xiong Y H, Wang N, et al. Coupling degree evaluation of China's internet financial ecosystem based on entropy method and principal component analysis. Journal of Systems Science and Information, 2019, 7(5): 399–421.
[13] Ye J. Least squares linear discriminant analysis. Proceedings of the 24th International Conference on Machine Learning, 2007: 1087–1093.
[14] Chen L F, Liao H Y, Ko M T, et al. A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognition, 2000, 33(10): 1713–1726.
[15] Huang R, Liu Q, Lu H, et al. Solving the small sample size problem of LDA. Pattern Recognition, 2002, 3: 29–32.
[16] Zheng W, Zhao L, Zou C. An efficient algorithm to solve the small sample size problem for LDA. Pattern Recognition, 2004, 37(5): 1077–1079.
[17] Ye J P, Janardan R, Li Q. Two-dimensional linear discriminant analysis. Advances in Neural Information Processing Systems, 2005: 1431–1441.

[18] Martinez A M, Kak A C. Pca versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(2): 228–233.

[19] Boulesteix A L. PLS dimension reduction for classification with microarray data. Statistical Applications in Genetics and Molecular Biology, 2004, 3(1): 392.

[20] Wold S, Sjostrom M, Eriksson L. PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 2001, 58(2): 109–130.

[21] Goodhue D L, Lewis W, Thompson R. Does PLS have advantages for small sample size or non-normal data? MIS Quarterly, 2012, 36(3): 981–1001.

[22] Chen Y G. Application of partial least squares to performance multi-evaluation for public sectors. Systems Engineering — Theory & Practice, 2009, 29(1): 89–96.

[23] Jin X, Zhang X, Rao K, et al. Semi-supervised partial least squares. International Journal of Wavelets, Multiresolution and Information Processing, 2020, 18(3): 2050014.

[24] Marigheto N A, Kemsley E K, Defernez M, et al. A comparison of midinfrared and raman spectroscopies for the authentication of edible oils. Journal of the American Oil Chemists' Society, 1998, 75(8): 987–992.

[25] Shao X, Bian X, Cai W. An improved boosting partial least squares method for near-infrared spectroscopic quantitative analysis. Analytica Chimica Acta, 2010, 666(1–2): 32–37.

[26] Zhang M H, Xu Q S, Massart D L. Boosting partial least squares. Analytical Chemistry, 2005, 77(5): 1423–1431.

[27] Afara I, Singh S, Oloyede A. Application of near infrared (nir) spectroscopy for determining the thickness of articular cartilage. Medical Engineering and Physics, 2013, 35(1): 88–95.

[28] Bi Y, Xie Q, Peng S, et al. Dual stacked partial least squares for analysis of near-infrared spectra. Analytica Chimica Acta, 2013, 792(16): 19–27.

[29] Ni W, Brown S D, Man R. Stacked partial least squares regression analysis for spectral calibration and prediction. Journal of Chemometrics, 2010, 23(10): 505–517.

[30] Montanari A. Linear discriminant analysis and transvariation. Journal of Classification, 2004, 21(1): 71–88.

[31] Qin X, Gao F, Chen G. Wastewater quality monitoring system using sensor fusion and machine learning techniques. Water Research, 2012, 46(4): 1133–1144.

[32] Tan C, Wang J Y, Wu T, et al. Determination of nicotine in tobacco samples by near-infrared spectroscopy and boosting partial least squares. Vibrational Spectroscopy, 2010, 54(1): 35–41.

[33] Breiman L. Bagging predictors. Machine Learning, 1996, 24(2): 123–140.

[34] Bradley E. An introduction to the bootstrap. Chapman and Hall, 1995.

[35] Xu L, Jiang J H, Zhou Y P, et al. Mccvstacked regression for model combination and fast spectral interval selection in multivariate calibration. Chemometrics and Intelligent Laboratory Systems, 2007, 87(2): 226–230.

[36] Rengganis R, Sari M M R, Budiasih I. The fraud diamond: Element in detecting financial statement of fraud. International Research Journal of Management, IT and Social Sciences, 2019, 6(3): 1–10.

[37] Li J, Chang Y, Wang Y, et al. Tracking down financial statement fraud by analyzing the supplier-customer relationship network. Computers & Industrial Engineering, 2023, 178: 109118.

[38] Spathis C T. Detecting false financial statements using published data: Some evidence from Greece. Managerial Auditing Journal, 2002, 17(4): 179–191.

[39] Perols J. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. Auditing: A Journal of Practice & Theory, 2011, 30(2): 19–50.

[40] Wen S, Li J, Zhu X, et al. Analysis of financial fraud based on manager knowledge graph. Procedia Computer Science, 2022, 199: 773–779.

[41] Zhang J S, Lei Z Q, Cheng R K, et al. Short-term electricity price forecasting using random forest model with parameters tuned by grey wolf algorithm optimization. Journal of Systems Science and Information, 2022, 10(2): 167–180.

[42] Kong A, Zhu H L. Predicting trend of high frequency CSI 300 index using adaptive input selection and machine learning techniques. Journal of Systems Science and Information, 2018, 6(2): 120–133.

[43] Li Y. Corporate financial fraud identification and crisis forewarning based on the partial least squares method. International Journal of Engineering Intelligent Systems, 2022, 30(3).

[44] Zareapoor M, Shamsolmoali P. Application of credit card fraud detection: Based on bagging ensemble classifier. Procedia Computer Science, 2015, 48(2015): 679–685.

[45] Maclin R, Opitz D. Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research, 2011, 11: 169–198.
[46] Ren D, Qu F, Lü K, et al. A gradient descent boosting spectrum modeling method based on back interval partial least squares. Neurocomputing, 2012, 171(C): 1038–1046.
[47] Bian X, Li S, Shao X, et al. Variable space boosting partial least squares for multivariate calibration of near-infrared spectroscopy. Chemometrics and Intelligent Laboratory Systems, 2016, 158: 174–179.
[48] Folch-Fortuny A, Arteaga F, Ferrer A. PLS model building with missing data: New algorithms and a comparative study. Journal of Chemometrics, 2017, 31: 1–2.
[49] Shao-Hong G U, Wang Y S, Wang G X. Application of principal component analysis model in data processing. Journal of Surveying and Mapping, 2007, 24(5): 387–390.
[50] Kambhatla N, Leen T. Dimension reduction by local principal component analysis. Neural Computation, 1997, 9(7): 1493–1516.
[51] Trendafilov N T, Unkel S, Krzanowski W. Exploratory factor and principal component analyses: Some new aspects. Kluwer Academic Publishers, 2013.
[52] Chiang K Y, Hsieh C J, Dhillon I S. Robust principal component analysis with side information. International Conference on Machine Learning, 2016: 2291–2299.
[53] Zhou Z H, Wu J, Tang W. Ensembling neural networks: Many could be better than all. Artificial Intelligence, 2002, 137(1): 239–263.
[54] Hu Y, Peng S, Peng J, et al. An improved ensemble partial least squares for analysis of near-infrared spectra. Talanta, 2012, 94: 301–307.
[55] Peschke K D, Haasdonk B, Ronneberger O, et al. Using transformation knowledge for the classification of raman spectra of biological samples. Proceedings of the 4th Iasted International Conference on Biomedical Engineering, 2006: 288–293.
[56] Ferrari A C, Robertson J. Interpretation of raman spectra of disordered and amorphous carbon. Physical Review B, 2000, 61(20): 14095–14107.
[57] Kim Y J, Baik B, Cho S. Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. Expert Systems with Applications, 2016, 62: 32–43.
[58] Dechow P M, Ge W, Larson C R, et al. Predicting material accounting misstatements. Contemporary Accounting Research, 2011, 28(1): 17–82.
[59] Dalnial H, Kamaluddin A, Sanusi Z M, et al. Accountability in financial reporting: Detecting fraudulent firms. Procedia-Social and Behavioral Sciences, 2014, 145: 61–69.
[60] Pai P F, Hsu M F, Wang M C. A support vector machine-based model for detecting top management fraud. Knowledge-Based Systems, 2011, 24(2): 314–321.
[61] Song X P, Hu Z H, Du J G, et al. Application of machine learning methods to risk assessment of financial statement fraud: Evidence from China. Journal of Forecasting, 2014, 33(8): 611–626.
[62] Spathis C, Doumpos M, Zopounidis C. Detecting falsified financial statements: A comparative study using multicriteria analysis and multivariate statistical techniques. European Accounting Review, 2002, 11(3): 509–535.
[63] Persons O S. Using financial statement data to identify factors associated with fraudulent financial reporting. Journal of Applied Business Research (JABR), 1995, 11(3): 38–46.
[64] Kirkos E, Spathis C, Manolopoulos Y. Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications, 2007, 32(4): 995–1003.
[65] Ma X P, Wang Y R, Li J P, et al. Analyst coverage, forecasting bias, and corporate innovation: Evidence from China. Journal of Systems Science and Information, 2024, 12(1): 1–24.
[66] Gan X Y, Xu T Y, Li Z H, et al. The impact of industrial policy on photovoltaic enterprise risk using an LDA based-deep neural network model. Journal of Systems Science and Information, 2022, 10(2): 181–192.