Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6198317, 9 pages
https://doi.org/10.1155/2021/6198317

Research Article

A High-Dimensional Modeling System Based on Analytical Hierarchy Process and Information Criteria

Tuba Koç

Cankiri Karatekin University, Faculty of Science, Statistics Department, Cankiri, Turkey

Correspondence should be addressed to Tuba Koç; tubakoc@karatekin.edu.tr

Received 27 July 2021; Accepted 17 August 2021; Published 29 August 2021

Academic Editor: Ishfaq Ahmad

High-dimensional data sets frequently occur in several scientific areas, and special techniques are required to analyze these types of data sets. Especially, it becomes important to apply a suitable model in classification problems. In this study, a novel approach is proposed to estimate a statistical model for high-dimensional data sets. The proposed method uses analytical hierarchical process (AHP) and information criteria for determining the optimal PCs for the classification model. The high-dimensional "colon" and "gravier" datasets were used in evaluation part. Application results demonstrate that the proposed approach can be successfully used for modeling purposes.

1. Introduction

High-dimensional data refer to the state of $n < p$, where the number of unknown parameters is p and the sample size is n. Analyzing high-dimensional data is encountered in many areas. In statistical modeling, the classical solutions fail to produce useful results when there is high-dimensional data. High-dimensional modeling techniques are used to overcome this problem. The dimensional reduction is a convenient approach for high-dimensional modeling [1]. Some of the advantages of size reduction techniques are as follows: (i) it reduces the number of dimensions and data storage; (ii) it requires less time to calculate; (iii) irrelevant, noisy, and unnecessary data can be deleted; (iv) data quality can be well optimized; (v) it helps an algorithm work efficiently and improves accuracy; (vi) it allows visualizing data; and (vii) it simplifies classification and improves performance as well.

The researchers have studied on PCA for reducing the dimension of the explanatory variable sets. Liu et al. [2] used kernel PCA for gene expression classification. Nyamundanda et al. [3] gave a novel extension of PPCA called probabilistic principal component and covariates analysis (PPCCA) which provides a flexible approach to jointly model metabolomics data and additional covariate information. In [4], PCA was applied for EEG-based emotion recognition classification. Khedher et al. [5] established a classification system using PCA for early diagnosis of Alzheimer's disease. Gaikwad and Joshi [6] used PCA within probabilistic neural network to classify the brain tumors. Kondo et al. [7] applied PCA and logistic regression analysis for the diagnosis of lung cancer. Pamukçu et al. [1] used probabilistic PCA and several classification algorithms in gene expression data sets. Passemier et al. [8] developed new statistical theory for probabilistic principal component analysis models in high dimensions. Smallman et al. [9] enhanced a sparse method for unsupervised dimension reduction for data from an exponential-family distribution. In [10], deep neural networks, PCA, and linear support vector machine algorithms were used for Hoda dataset classification. Yao and Lopes [11] demonstrated the role of transformations in bootstrapping methods for high-dimensional PCA with their simulation and numeric experiments. Hung and Huang [12] proposed the generalized information criterion (GIC) for high-dimensional PCA sequence selection. Ayesha et al. [13] presented the state-of-the art size reduction techniques and their suitability for different types of data and application areas. Choubey et al. [14] proposed PCA, particle swarm optimization (PSO), and different machine learning algorithms as feature reduction or feature selection or attribute selection method for the

detection of diabetes. The literature can be extended with different studies.

While constructing the model with PCA, it is very important issue to determine the optimal number of principal components (PCs). However, there is not an objective criterion of identifying the optimal PCs. Although probabilistic PCA considers information criteria to select, this approach has two drawbacks: (1) probabilistic PCA only considers the explanatory variable set and does not take into account the classification model and (2) each criterion can determine the different number of PCs.

In this article, a novel hybrid approach is proposed to optimal model for high-dimensional data sets. The proposed method uses analytical hierarchical process (AHP) and information criteria for determining the optimal PCs for the classification model. The multivariate adaptive regression splines (MARS) algorithm is used as the classification algorithm within selected PCA scores. This article is an important attempt to construct an objective, high-dimensional model in the context of principal component regression modeling.

The article is organized as follows. 2 introduces the proposed hybrid system for determining the optimal PCs in high-dimensional settings and presents the employment of the MARS algorithm. Section 3 provides the application results on data sets. Section 4 presents the conclusion part.

2. Hybrid System Process for Selecting the Optimal PCs

A hybrid system is proposed to perform classification in cases where high-dimensional data are available. The multivariate adaptive regression splines (MARS) model is used as the classification model. The main reason of choosing MARS is inherently having variable selection capability. MARS can exclude redundant dimensions intelligently. The proposed system has three steps:

(1) Once, the principal component analysis is implemented on the explanatory variables. However, for multivariate techniques based on an accurate estimation of true covariance, where the $n < p$ problem exists, the classical sample covariance matrix has a systematically distorted Eigen structure. In this case, the structure of the covariance matrix undergoes a distortion such that the largest eigenvalues are upbiased and the smallest eigenvalues are down-biased. To overcome the limitations of the sample covariance matrix, shrinkage approaches are often used to estimate the high-dimensional covariance matrix [15]. Therefore, the shrinkage covariance matrix is used instead of the classical one. The principal component analysis is performed via the shrinkage covariance matrix, and the PCA scores are used as the explanatory variables. In this way, fewer predictors are used to build the MARS model.

(2) The PCA process is performed by selecting the optimal number of PC according to the information criteria using AHP. While performing PCA, the number of PCs should be chosen carefully. Since the PC scores directly affect the classification model, the process of PCA is linked within the MARS model. For each number of PCs, the MARS model is constructed and information criteria values are computed. Obviously, each criterion has the different decision on selecting the number of PCs, and we require a unique solution. This solution is obtained via AHP approach. The TOPSIS method is used for the AHP process. TOPSIS evaluates the several information criteria for every number of PCs, and it gives common scores. Upon these scores, one may select the optimal number of PCs.

(3) After selecting the optimal number of PCs, the MARS model is constructed. Due to the nature of MARS, some redundant dimensions can be reduced to improve the inference capability.

2.1. Principal Component Analysis with Shrinkage Covariance Matrix. Principal components analysis (PCA) was first introduced by Pearson [16] and developed by Hoteling [17] and Rao [18]. PCA is a vector-based approach in which the aim is to convert high-dimensional and interrelated vectors into small-sized unrelated vectors. The basic component analysis provides the explanation of the variance-covariance structure through a few linear combinations of the original variables. The general objective is to reduce dimension and to make an interpretation, as well as to take measures against the rank problem and to remove the linear relationship in the variance-covariance matrix. Let X be an $(n \times p)$ matrix where n is the number of samples and p is the number of variables or properties. Let us show the covariance matrix of the X data set with C. For the purpose of PCA, when $D = (1/n)YY'$ is a diagonal matrix, in the $Y = PX$ transformation, there must be an orthonormal V matrix such that the rows of P are the principal components of X. If the D matrix is rewritten as

$$D = \frac{1}{n}(PX)(PX)' = \frac{1}{n}PXX'P' = P\left(\frac{1}{n}XX'\right)P' = PCP'. \tag{1}$$

C variance-covariance matrix (by definition of self-decomposition) can be written as $C = V\Lambda V'$. In this case, the D matrix is $D = PCP' = P(V\Lambda V')P'$, and if we take $P \equiv V'$, we obtain

$$D = V'\left((V\Lambda V')V = \left(V'V\right)\right)\Lambda\left(V'V\right) = \Lambda. \tag{2}$$

That is, selecting $P \equiv V'$ makes D diagonal, and this means that the covariance of the newly obtained variables is zero, which is the purpose of PCA. Thus, the principal components of X are C's eigenvectors [19]. The general process of the principal component analysis is summarized above. In this study, the eigenvalues of the covariance matrix C, which was used in the analysis of principal components, were negative. The shrinkage covariance matrix is

considered instead of C matrix to eliminate possible bias in the analysis results.

Shrinkage estimators of the C variance-covariance matrix shrink the eigenvalues of $\widehat{C}_{\mathrm{MLE}}$ to the center. The purpose of shrinkage estimators is to take a convex combination of $\widehat{D}$ which is a target diagonal matrix that has been selected appropriately with the sample variance $\widehat{C}_{\mathrm{MLE}}$ of C [20]. Then, the covariance matrix's shrinkage estimator is as follows:

$$\widehat{C}_S = (1-\widehat{\rho})\widehat{C}_{\mathrm{MLE}} + \widehat{\rho}\widehat{D}, \tag{3}$$

where $\widehat{\rho}$ is the optimal shrinkage coefficient (or density) and takes the values between 0 and 1. This value can also be a function of observations. Then, the $\widehat{D}$ matrix is called the shrinkage target. $\widehat{D}$ Naive form is given as follows:

$$\widehat{D} = \frac{\mathrm{tr}\left(\widehat{C}_{\mathrm{MLE}}\right)}{p} I_p = \left(\frac{1}{p}\sum_{j=1}^{p}\lambda_j\right) I_p = \overline{\lambda} I_p. \tag{4}$$

Further information on the different shrinkage target matrices of the shrinkage covariance matrix is given in [21].

2.2. Multivariate Adaptive Regression Splines with Several Information Criteria. Multivariate adaptive regression splines (MARS) are a nonparametric regression technique which models the nonlinear relationship between a response variable and the set of predictors via basis functions [22]. Friedman [23] contrived MARS to capture the nonlinearities in the model. In MARS, it is not necessary to know the functional relationship. MARS can detect the relationships among predictors and response using basis functions. A general model form of MARS is shown as follows:

$$f(x) = \sum_{r=1}^{p} \beta_r B\left(X_r\right) + \varepsilon, \tag{5}$$

where β represents the coefficient vector, $B(.)$ shows the basis functions, and ε indicates the random error term. The basis functions are described as follows:

$$\begin{aligned} (x-t)_+ &= \begin{cases} x-t, & x>t, \\ 0, & x\geq t, \end{cases} \\ (x-t)_- &= \begin{cases} x-t, & x<t, \\ 0, & x\geq t, \end{cases} \end{aligned} \tag{6}$$

where "–" shows the negative, "+" positive region, and "t" shows the knot points. Knots can be defined as the numbers that controls the starting and ending points in local relationships. MARS uses these knots for identifying the linearity of nonlinear relationships in the model.

Figure 1 shows a graphical representation for the knots. Two knots points x_1 and x_2 are selected according to behaviors of the relationship between predictors and response. MARS handles the optimal knot selection with a goodness of fit measure. The most common used measure is generalized cross validation (GCV) which is defined as follows:

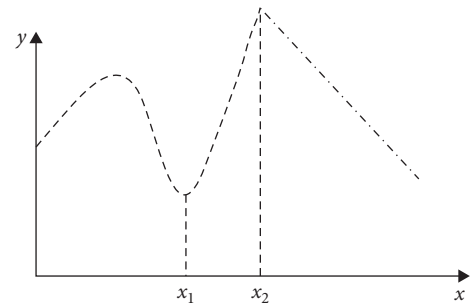

Figure 1: Example graph for knot.

$$\mathrm{GCV} = \sum_{i=1}^{n} \frac{\left(y_i - \widehat{y}_i\right)^2}{\left(1 - C(M)/n\right)^2}, \tag{7}$$

where $\widehat{y}$ indicates the predicted values and $C(M)$ shows a penalty measure which is related with the number of selected parameters.

In this study, we used the following information criteria after the most commonly used GCV measurement in MARS. It should be noted that the following criteria are used for the model evaluation part. The included information criteria are Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and Consistent Information Complexity Criteria (CICOMP) [24–26]. The formulations of these criteria are as follows:

$$\begin{aligned} \mathrm{AIC} &= -2\log L(\widehat{M}) - 2k, \\ \mathrm{BIC} &= -2\log L(\widehat{M}) + k\log(n), \\ \mathrm{CAIC} &= -2\log L(\widehat{M}) + k(\log(n)+1), \end{aligned} \tag{8}$$

where $L(\widehat{M})$ denotes the log-likelihood of the regression model, k shows the number of variables, and n indicates the sample size.

2.3. Analytical Hierarchy Process for Choosing the Number of PCs. The analytical hierarchy process (AHP) that has developed by Saaty [27] is a mathematical decision-making method that enables more efficient tools for researchers to organizing and analyzing complex decisions. It is easy to use and take into account both measurable and nonmeasurable criteria. The AHP method organizes the selected factors in a hierarchical structure according to the criteria, subcriteria, and alternatives under a general target and tries to reach the result. The following steps are followed in the AHP method:

By making pairwise comparisons, decision-makers create one matrix for each alternative and one matrix for criteria

These matrices are normalized, and their consistency is checked (whether the decision-makers two-point comparisons are consistent)

Then, with the help of matrix algebra, an average score is obtained for each alternative

The alternative with the highest score is the most appropriate alternative to the decision-makers comparisons [28].

In this study for each number of PCs, the MARS model is constructed and information criteria values are computed. Since each criterion has different decisions on selecting the number of PCs, we require a unique solution. We obtained this solution via AHP approach. The TOPSIS method is used for the AHP process. TOPSIS evaluates the several information criteria for every number of PCs, and it gives common scores.

3. Application

In this part, we applied the proposed method based on the hybridization of principal component analysis (PCA), analytical hierarchical processing (AHP), and information criteria to high-dimensional data sets.

Example 1. Colon data set as gene expression data is available in *R* software of packages "plsgenomics" [29]. The data set includes $p = 2000$ genes and $n = 62$ samples. The response variable has two groups as tumor tissues and normal tissues [30]. The task is to classify the tissues using 2000 predictors. Obviously, the data set is high dimensional for the reason of $n < p$.

If there exists $n < p$ situation for multivariate techniques that is based on an accurate estimation of true covariance, the classical sample covariance matrix has a systematically deteriorated self-structure. As mentioned earlier, the negative eigenvalues of the covariance matrix are a major problem for the analysis.

Table 1 presents the eigenvalues of both the classical and shrinkage covariance matrices. As it is seen, the classical covariance matrix produced negative eigenvalues because of the high dimensionality. This problem was handled by using the shrinkage covariance matrix which carries the negative eigenvalues into the positive range. After obtaining positive eigenvalues, we applied the principal component analysis with a positive definite shrinkage covariance matrix.

Table 2 shows the information criteria of the MARS model for each number of principal components. In Table 2, while AIC and BIC select 14 principal components, CAIC and MARS model's GCV measurement selects 15 and more principal component. The reported results clearly denote that each criterion determined different number of principal components. In this case, deciding the number of principal components for researchers is becoming a problem. We have addressed the AHP process to solve this problem and find an optimum number of principal components.

Table 3 shows the results for the AHP scores for each criterion. According to the results in Table 3, the number of principal components with a rank value of 1 is 14. So that this information shows us that the number of optimum principal components for this study is 14. After the principal component number was determined by the AHP process, we fitted the MARS regression model.

Table 1: Eigenvalues of covariance matrices.

Index	Classical-S	Shrinkage-S
1	0.41422	0.25941
2	0.36706	0.23006
3	0.28062	0.17702
4	0.26521	0.16741
5	0.20503	0.12669
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
96	$-8.66E-18$	0.01132
97	$-1.20E-17$	0.01110
98	$-1.70E-17$	0.01081
99	$-4.40E-17$	0.01075
100	$-6.37E-17$	0.01048

Table 2: Information criteria for each principal component of the MARS model.

PC	Information criteria			
	AIC	BIC	CAIC	GCV
1	−87.40951	−83.15524	−81.15524	0.23650
2	−87.40951	−83.15524	−81.15524	0.23650
3	−91.30137	−84.91997	−81.91997	0.22988
4	−102.30053	−93.79199	−89.79199	0.19971
5	−112.04647	−97.15653	−90.15653	0.19352
6	−114.09778	−97.08070	−89.08070	0.19638
7	−111.78619	−94.76912	−86.76912	0.20384
8	−111.78619	−94.76912	−86.76912	0.20384
9	−111.78619	−94.76912	−86.76912	0.20384
10	−111.78619	−94.76912	−86.76912	0.20384
11	−114.09778	−97.08070	−89.08070	0.19638
12	−110.15391	−97.39111	−91.39111	0.19080
13	−109.15987	−94.26993	−87.26993	0.20274
14	**−122.57154**	**−101.30020**	−91.30020	0.19041
15	−109.21318	−98.57751	**−93.57751**	**0.18578**
16	−109.21318	−98.57751	**−93.57751**	**0.18578**
17	−109.21318	−98.57751	**−93.57751**	**0.18578**
18	−109.21318	−98.57751	**−93.57751**	**0.18578**
19	−109.21318	−98.57751	**−93.57751**	**0.18578**
20	−109.21318	−98.57751	**−93.57751**	**0.18578**

Table 4 shows the results of the MARS regression model. We applied logit transformation in the MARS regression model, and the link function is as follows:

$$\text{Pred}_{\text{class}} = \frac{1}{1 - e^{-z}}. \tag{9}$$

The fitted model for the MARS regression is given as follows:

Table 3: Number of basic components by AHP process.

PC	Score	Rank
1	0	20
2	$6.91E-16$	19
3	0.11123	18
4	0.55448	17
5	0.75040	10
6	0.75683	3
7	0.64720	15
8	0.64720	13
9	0.64720	13
10	0.64720	13
11	0.75683	2
12	0.74076	11
13	0.61901	16
14	**0.92986**	**1**
15	0.75330	6.5
16	0.75330	6.5
17	0.75330	6.5
18	0.75330	6.5
19	0.75330	6.5
20	0.75330	6.5

$$z = 0.45763 + 4.31099 \times \mathrm{BF1} - 18.35969 \times \mathrm{BF1} + 3.728584 \times \mathrm{BF2} - 2.031415 \times \mathrm{BF3} - 7.754506 \times \mathrm{BF4} + 22.28331 \times \mathrm{BF5} - 7.823195 \times \mathrm{BF6} + 22.29569 \times \mathrm{BF7}. \tag{10}$$

Also, as shown in Table 4, the MARS regression model is used to select variables as well as model selection (the model did not select PC1, PC2, and PC7 principal components among 14 principal components).

Table 5 shows the performance measurements of the model. Performance measurement formulas we use in the study are as follows:

$$\begin{aligned}
\text{sensitivity} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\\
\text{specifity} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}},\\
\text{accuracy} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}},\\
\text{precision} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\\
\text{recall} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\\
F-\text{measure} &= \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}},\\
\text{MCCR} &= \frac{\mathrm{TP}*\mathrm{TN}-\mathrm{FP}*\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}.
\end{aligned} \tag{11}$$

In these equations, TP (True positive): means correct positive prediction, FP (False positive) means incorrect positive prediction, TN (True negative) means correct negative prediction, and FN (False negative) means incorrect negative prediction.

When the performance measurement values of the MARS model that we applied to colon cancer data are examined, it is seen that all measurement values are good. The point of particular interest in Table 5 is that the MCCR value (0.789) is quite good because this measurement generally does not give good results in model selection methods.

To evaluate the performance of the proposed hybrid approach, we used two high-dimensional modeling techniques: least absolute shrinkage and selection operator (Lasso) and adaptive elastic net (Aenet). Lasso and Aenet models are appropriate for modeling the high-dimensional data sets. Both of these models require a tuning parameter selection, and generally the information criteria are used to select the optimal one. We need to choose an information criterion which is adopted for high-dimensional data sets, and therefore we prefer to use Extended Bayesian Information Criterion (EBIC) during the selection of the tuning parameter in Lasso and Aenet [27]. EBIC is formulated as follows:

$$\mathrm{EBIC} = \mathrm{BIC} + 2\gamma \log\binom{p}{k}. \tag{12}$$

In the EBIC formula, $\gamma = 0.5$ and $\log\binom{p}{k}$ shows the combination function. The Lasso and Aenet applications are handled with glmnet and msanet packages [28, 29].

Table 6 shows the performance measures for Lasso-EBIC, Aenet-EBIC, and the proposed model. The

Table 4: MARS regression model.

Basis function	Formula
BF1	*P*max (0, 0.23409 – PC3)
BF2	*P*max (0, 0.26568 – PC4)
BF3	*P*max (0, PC5 + 0.38728)
BF4	*P*max (0, PC6 – 0.15427)
BF5	*P*max (0, PC8 – 0.10529)
BF6	*P*max (0, PC8 – 0.46778)
BF7	*P*max (0, PC11 – 0.53962)
BF8	*P*max (0, PC14 – 0.03291)
BF9	*P*max (0, PC8 – 0.46778)
BF10	*P*max (0, PC11 – 0.53962)
BF11	*P*max (0, PC14 – 0.03291)

Table 5: Performance measurements for the MARS model.

Performance measure	Value
Sensitivity	0.864
Specificity	0.925
Accuracy	0.903
Precision	0.925
Recall	0.925
F-measure	0.925
MCCR	0.789

Table 6: The performance measures for the classification models.

Performance measure	Model		
	Proposed model (colon data)	Lasso-EBIC	Aenet-EBIC
Sensitivity	0.864	0.867	0.824
Specificity	0.925	0.809	0.822
Accuracy	**0.903**	0.823	0.823
Precision	0.925	0.809	0.822
Recall	0.925	0.950	0.925
F-measure	**0.925**	0.874	0.871
MCCR	**0.789**	0.604	0.602

comparisons are mainly based on three important measures: accuracy, *F*-measure, and MCCR values. It is pretty clear that the proposed model has the highest accuracy, *F*-measure, and MCCR values. Especially, the MCCR represents the classification performance for the proposed modeling approach. According to the results, the proposed hybrid approach gives more accurate results when comparing with the other classification models in high-dimensional data sets.

Example 2. Gravier data set as breast cancer data is available in *R* software of packages "datamicroarray." The data set includes $p = 2{,}905$ genes and $n = 168$ samples. The response variable has two groups in which patients with no events after diagnosis were labeled good and early metastases were labeled as poor [31]. Analyzes were made by selecting the first $p = 1000$ features from the data set. The data set is again high dimensional for the reason of $n < p$.

Table 7 presents the eigenvalues of both the classical and shrinkage covariance matrices. The negative eigenvalue problem is solved using the shrinkage covariance matrix. After obtaining positive eigenvalues, we applied the principal component analysis with a positive definite shrinkage covariance matrix.

Table 8 shows the information criteria of the MARS model for each number of principal components. In Table 8, while AIC, BIC, and MARS model's GCV measurement selects 19 principal components, CAIC selects 20 principal components. We discussed the AHP process to find the optimum number of principal components.

Table 9 shows the results for the AHP scores for each criterion. According to the results in Table 9, the number of principal components with a rank value of 1 is 19. After the principal component number was determined by the AHP process, we fitted the MARS regression model.

Table 10 shows the results of the MARS regression model. We applied logit transformation in the MARS regression model. The fitted model for the MARS regression is given as follows:

Table 7: Eigenvalues of covariance matrices.

Index	Classical-S	Shrinkage-S
1	2.79003	1.88230
2	1.59543	1.17074
3	0.94476	0.68492
4	0.81998	0.61029
5	0.74720	0.54854
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
996	$-1.27E-16$	0.0020
997	$-1.62E-16$	0.0020
998	$-7.78E-16$	0.0020
999	$-1.67E-15$	0.0019
1000	$-2.69E-15$	0.0017

Table 8: Information criteria for each principal component of the MARS model.

PC	Information criteria			
	AIC	BIC	CAIC	GCV
1	−247.21820	−240.97030	−238.97030	0.22686
2	−247.21820	−240.97030	−238.97030	0.22686
3	−269.25770	−253.63780	−248.63780	0.20657
4	−289.79220	−271.04840	−265.04840	0.18521
5	−288.36040	−266.49260	−259.49260	0.18932
6	−295.30230	−270.31060	−262.31060	0.18417
7	−295.30230	−270.31060	−262.31060	0.18417
8	−305.26260	−274.02290	−264.02290	0.17858
9	−311.14600	−276.78240	−265.78240	0.17500
10	−311.14600	−276.78240	−265.78240	0.17500
11	−318.66110	−278.04960	−265.04960	0.17255
12	−326.31440	−279.45490	−264.45490	0.17025
13	−326.31440	−279.45490	−264.45490	0.17025
14	−326.31440	−279.45490	−264.45490	0.17025
15	−326.31440	−279.45490	−264.45490	0.17025
16	−311.89210	−277.52850	−266.52850	0.17422
17	−338.05580	−281.82440	−263.82440	0.16710
18	−331.97960	−281.99620	−265.99620	0.16737
19	**−339.77790**	**−283.54650**	−265.54650	**0.16540**
20	−311.67530	−277.31170	**−266.31170**	0.17445

Table 9: Number of basic components by AHP process.

PC	Score	Rank
1	$0.00E+00$	20
2	$1.34E-15$	19
3	0.29486	18
4	0.60325	16
5	0.55120	17
6	0.62968	14.5
7	0.62968	14.5
8	0.72226	13
9	0.77997	11.5
10	0.77997	11.5
11	0.83481	8
12	0.89028	6
13	0.89028	7
14	0.89028	4
15	0.89028	5
16	0.78997	9
17	0.96687	2
18	0.94268	3
19	**0.99237**	**1**
20	0.78710	10

Table 10: MARS regression model.

Basis function	Formula
BF1	Pmax (0, 0.265434 – PC3)
BF2	Pmax (0, PC3 – 0.265434)
BF3	Pmax (0, PC3 – 2.178976)
BF4	Pmax (0, PC4 + 0.06377673)
BF5	Pmax (0, PC4 – 0.5227402)
BF6	Pmax (0, PC4 – 0.8542159)
BF7	Pmax (0, PC6 – 0.5927528)
BF8	Pmax (0, PC8 – 0.08261739)
BF9	Pmax (0, PC8 – 0.6397357)
BF10	Pmax (0, PC8 – 1.085249)
BF11	Pmax (0, PC9 + 0.8624089)
BF12	Pmax (0, PC15 + 0.7457308)
BF13	Pmax (0, −0.3014017 – PC15)
BF14	Pmax (0, PC15 + 0.3014017)
BF15	Pmax (0, PC15 – 0.4687581)
BF16	Pmax (0, PC17 – 0.3960287)

Table 11: Performance measurements for the MARS model.

Performance measure	Model		
	Proposed model (gravier data)	Lasso-EBIC	Aenet-EBIC
Sensitivity	0.975	0.642	0.692
Specificity	0.947	0.302	0.526
Accuracy	**0.966**	0.642	0.654
Precision	0.947	0.752	0.526
Recall	0.947	0.523	0.333
F-measure	**0.947**	0.656	0.408
MCCR	**0.922**	0.233	0.191

$$\begin{aligned} z = {} & 5.057191 + 4.150164 \times \mathrm{BF1} + 2.523352 \times \mathrm{BF2} - 3.483849 \times \mathrm{BF3} - 12.72563 \\ & \times \mathrm{BF4} + 28.27023 \times \mathrm{BF5} - 17.27563 \times \mathrm{BF6} - 6.074134 \times \mathrm{BF7} - 5.962305 \times \mathrm{BF8} \\ & + 12.79007 \times \mathrm{BF9} - 9.163705 \times \mathrm{BF10} + 0.6654436 \times \mathrm{BF11} - 17.70061 \times \mathrm{BF12} \\ & - 6.683544 \times \mathrm{BF13} + 22.49807 \times \mathrm{BF14} - 6.951832 \times \mathrm{BF15} + 5.628957 \times \mathrm{BF16}. \end{aligned} \tag{13}$$

Table 11 shows the performance measures for Lasso-EBIC, Aenet-EBIC, and the proposed model. In Table 11, the proposed hybrid approach gives more accurate results than other classification models.

4. Conclusions

In high-dimensional data sets where the number of variables is greater than the number of samples, the classical covariance matrix structure has a systematic degeneration. The degeneration of the classical covariance matrix structure leads to a negative value of the eigenvalues. This causes the results of PCA analysis to be misleading. Besides that fact, the reduction phase requires choosing the optimal number of components. In this study, we have introduced a new approach in model estimation to overcome this problem. The proposed approach objectively identifies the number of PCs for a high-dimensional setting. The proposed system also enables to reduce the irrelevant component through MARS. One of the major advantages of this approach is to associate the selection process within the estimation model. Empirical findings prove that the developed hybrid system produces accurate results for the high-dimensional datasets.

Data Availability

Colon data set as gene expression data is available in *R* software of packages "plsgenomics" [29]. Gravier data set as breast cancer data is available in *R* software of packages "datamicroarray" [31].

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] E. Pamukçu, H. Bozdogan, and S. Çalık, "A novel hybrid dimension reduction technique for undersized high dimensional gene expression data sets using information complexity criterion for cancer classification," *Computational and*

mathematical methods in medicine, vol. 2015, Article ID 370640, 9 pages, 2015.

[2] Z. Liu, D. Chen, and H. Bensmail, "Gene expression data classification with kernel principal component analysis," *Journal of Biomedicine and Biotechnology*, vol. 2005, no. 2, p. 155, Article ID 905863, 2005.

[3] G. Nyamundanda, L. Brennan, and I. Gormley, "Probabilistic principal component analysis for metabolomic data," *BMC Bioinformatics*, vol. 11, no. 1, p. 571, 2010.

[4] S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena, "EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation," *Science World Journal*, vol. 2014, Article ID 627892, 11 pages, 2014.

[5] L. Khedher, J. Ramírez, J. M. Górriz, A. Brahim, and F. Segovia, "Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images," *Neurocomputing*, vol. 151, pp. 139–150, 2015.

[6] S. B. Gaikwad and M. S. Joshi, "Brain tumor classification using principal component analysis and probabilistic neural network," *International Journal of Computers and Applications*, vol. 120, no. 3, 2015.

[7] T. Kondo, J. Ueno, and S. Takao, "Logistic GMDH-type neural network using principal component-regression analysis and its application to medical image diagnosis of lung cancer," *Artificial Life and Robotics*, vol. 20, no. 2, pp. 137–144, 2015.

[8] D. Passemier, Z. Li, and J. Yao, "On estimation of the noise variance in high dimensional probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B*, vol. 79, no. 1, pp. 51–67, 2017.

[9] L. Smallman, A. Artemiou, and J. Morgan, "Sparse generalised principal component analysis," *Pattern Recognition*, vol. 83, pp. 443–455, 2018.

[10] A. Bossaghzadeh, "Improving persian digit recognition by combining deep neural networks and SVM and using PCA," in *Poceedings of the 2020 International Conference on Machine Vision and Image Processing (MVIP)*, pp. 1–5, IEEE, Tehran, Iran, February 2020.

[11] J. Yao and M. E. Lopes, "Rates of bootstrap approximation for eigenvalues in high-dimensional pca," 2020, https://arxiv.org/abs/2104.07328.

[12] H. Hung and S. Y. Huang, "A generalized information criterion for high-dimensional pca rank selection," 2020, https://arxiv.org/abs/2004.13914.

[13] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, 2020.

[14] D. K. Choubey, P. Kumar, S. Tripathi, and S. Kumar, "Performance evaluation of classification methods with PCA and PSO for diabetes," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, pp. 1–30, 2020.

[15] Y. Chen, "Robust shrinkage estimation of high dimensional covariance matrices," in *Proceedings of the IEEE Workshop on Sensor Array and Multichannel Signal Processing (SAM)*, Jerusalem, Israel, October 2010.

[16] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[17] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[18] C. R. Rao, "The use and interpretation of principal component analysis in applied research," *Sankhyā: The Indian Journal of Statistics*, vol. 18, pp. 329–358, 1964.

[19] J. Shlens, "A tutorial on principal component analysis," 2014, https://arxiv.org/abs/1404.1100.

[20] S. Mohebbi, E. Pamukcu, and H. Bozdogan, "A new data adaptive elastic net predictive model using hybridized smoothed covariance estimators with information complexity," *Journal of Statistical Computation and Simulation*, vol. 89, no. 6, pp. 1060–1089, 2019.

[21] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, 2005.

[22] E. K. Koc and H. Bozdogan, "Model selection in multivariate adaptive regression splines (MARS) using information complexity as the fitness function," *Machine Learning*, vol. 101, no. 1–3, pp. 35–58, 2015.

[23] J. H. Friedman, "Multivariate adaptive regression splines," *Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.

[24] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[25] G. Schwarz, "Cumulative index-volumes I–XVI," *Progress in Optics*, vol. 6, no. 2, pp. 461–464, 1978.

[26] H. Bozdogan, "Akaike's information criterion and recent developments in information complexity," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 62–91, 2000.

[27] T. L. Saaty, *The Analytic Process: Planning, Priority Setting, Resources Allocation*, McGraw, New York, NY, USA, 1980.

[28] M. Timor, *Analitik Hiyerarşi Prosesi*, Türkmen Kitabevi, Istanbul, Turkey, 2011.

[29] A. Boulesteix, G. Durif, S. Lambert-Lacroix, J. Peyre, and K. Strimmer, "Plsgenomics: PLS Analyses for genomics," *R package version 1.5–1*, vol. 33, 2017, https://CRAN.R-project.org/package=plsgenomics.

[30] U. Alon, N. Barkai, D. A. Notterman et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.

[31] E. Gravier, G. Pierron, and A. V. Salomon, "A prognostic dna signature for t1t2 node-negative breast cancer patients," 2010.