ANOVA bootstrapped principal components analysis for logistic regression

Borislava Toleva
Sofia University "St Kliment Ohridski", Bulgaria, vrigazova@uni-sofia.bg

Abstract

Principal components analysis (PCA) is often used as a dimensionality reduction technique. A small number of principal components is selected to be used in a classification or a regression model to boost accuracy. A central issue in the PCA is how to select the number of principal components. Existing algorithms often result in contradictions and the researcher needs to manually select the final number of principal components to be used. In this research the author proposes a novel algorithm that automatically selects the number of principal components. This is achieved based on a combination of ANOVA ranking of principal components, the bootstrap and classification models. Unlike the classical approach, the algorithm we propose improves the accuracy of the logistic regression and selects the best combination of principal components that may not necessarily be ordered. The ANOVA bootstrapped PCA classification we propose is novel as it automatically selects the number of principal components that would maximise the accuracy of the classification model.

Keywords: ANOVA, bootstrap, classification, logistic regression.

JEL classification: C38, C52, C63.
DOI: 10.2478/crebss-2022-0002

Received: January 12, 2022
Accepted: April 01, 2022

Acknowledgments: The author would like to thank the ENTRENOVA 2021 conference for the comments and valuable feedback.

Introduction

Dimensionality reduction techniques are widely used in big datasets to decrease the size of the dataset. Dimensionality reduction can be achieved using two approaches. The first one is by decreasing the number of variables. This is feature selection. The second one is by transforming the original dataset into another dimension and then choosing a smaller set of transformed variables (James et al., 2013). This is called feature extraction (Maleki et al., 2020). This paper focuses on principal components analysis as a feature extraction technique. There are two main issues with principal components analysis – first, the criterion for choosing principal components is the percentage of variance explained, which can be misleading in cases with similar percentage of variance explained (James et al., 2013). Second, the selection of principal components is not automated, which can lead to time-consuming manual selection of principal components (James et al., 2013). The aim of this paper is to propose an automatic algorithm for selection of principal components based on the accuracy of the model.

Principal components analysis (PCA) is a feature extraction technique that transforms independent variables into principal components. Each principal component is a linear combination of independent variables. The aim is to select smaller number of principal components than the original number of variables to perform dimensionality reduction (James et al., 2013, Maleki et al., 2020). Selection of principal components is done by using the eigenvalues and eigenvectors to calculate the percentage of variance explained. The combination of principal components (principal components) that explains the highest percentage of variance in data is then selected (James et al., 2013, Maleki et al., 2020).

A central issue with this approach is how to select the number of principal components when the percentage of variance explained is similar for two or more principal components combinations (James et al., 2013). In this case manual selection based on prior knowledge or empirical results is the criterion for selecting the number of principal components (James et al., 2013, Maleki et al., 2020). The disadvantage is that manual selection may not result in the best accuracy and can introduce bias in the model. Also, it may be computationally exhaustive to produce many experiments to empirically select the number of principal components. The standard approach answers the questions: “What number of principal components should be selected to explain the highest percentage of variance in data?”.

The aim of this research is to propose an automatic algorithm to select the number of principal components in the case of logistic regression. The paper answers the question: “What number of principal components to be selected to achieve the highest accuracy using the logistic regression?”. The proposed algorithm is called the ANOVA-Bootstrapped Principal Components Analysis. It combines widely used models like the bootstrap and ANOVA with the principal components analysis. The novelty in the proposed algorithm is that it automatically selects the number of principal components that results in the highest accuracy. The advantages of the ANOVA-Bootstrapped PCA include fast automatic selection of the numbers of principal components, reduction of bias as no manual selection of the number of principal components is performed and selection of the combination of principal components that would result in the highest accuracy.

Next section provides overview of current modifications of the PCA to select the number of principal components more efficiently. Section 3 details the proposed methodology. Section 4 concludes.

Literature review

Feature selection methods keep the most informative features. They can be divided into three groups – embedded feature selection, filter methods and wrapper methods (Maleki et al., 2020). Embedded feature selection methods include lasso (Tibshirani, 1996), adaptive lasso (Zou, 2006), nonnegative garrotte (Breiman, 1995), etc. They perform feature selection as part of the classification/regression model. Filter methods select variables based on a criterion, for example, feature importance, correlation, etc. Such are ANOVA and correlation analysis (Maleki et al., 2020). A third group are wrapper methods, where a combination of feature selection technique with classification/regression model is used to select the important features. For example, recursive feature elimination with decision trees (Maleki et al., 2020).

Feature extraction models, on the other hand, transform the original space into another dimension, where smaller number of features are selected (Maleki et al., 2020). The two most common feature extraction techniques are the principal components analysis (PCA) and the linear discriminant analysis (LDA) (Maleki et al., 2020). The principal components analysis (PCA) is an unsupervised learning model, while the LDA is supervised learning (James et al., 2013). The principal components analysis can be used as a feature extraction technique and as a data exploratory technique (James et al., 2013). As a feature extraction technique, the PCA finds variables that are correlated and transforms them into principal components. Each principal component is a linear combination of the original variables in the dataset. Often the criterion to perform dimensionality reduction is by keeping the number of principal components that explain biggest percentage of the variance in data. Usually the first, second or third principal component are enough to build a model (James et al., 2013).

A central topic in PCA is how to identify the number of principal components that needs to be used for classification/ prediction (James et al., 2013, Maleki et al., 2020). A standard approach has been widely used by researchers and academia (James et al., 2013, Maleki et al., 2020). It involves exploring the percentage of variance in data that each principal component explains alone and when combined with the previous ones. For example, explore what percentage of the variance in data the first two, three, four, etc. principal components explain (James et al., 2013, Maleki et al., 2020). The issue with the standard approach is that the researcher should select among two or three options for the number of principal components in cases when the percentage of variance explained by several principal components is similar. As James et al. (2013) outline, in some cases the researcher may need to select between the first three and the first four principal components and the selection is made based on researcher's experience and many other subjective factors. This is because the first three or four principal components may be enough to explain bigger percentage of the variance in Y.

The standard approach (James et al., 2013) has been applied in many research papers, including recent ones (Salata et al., 2021). Some researchers, however, propose updated PCA algorithms in order to solve this issue. For example, Pacheco et al. (2013) proposes exact methods for variable selection in principal component analysis. Kim and Rattakorn (2011) use weighted principal components to perform unsupervised feature selection. Prieto-Moreno et al. (2015) use discriminant information to select principal components. Sharifzadeh et al. (2017) proposes SSPCA - Sparse supervised principal component analysis. Gajjar et al. (2017) uses a novel algorithm to select non-zero loadings to select number of principal components. Rahoma et al. (2021) uses the bootstrap to perform sparce principal components analysis. Although all these examples are a new way to find the number of principal components, none of them offers an automatic algorithm for principal components selection.

Like Rahoma et al. (2021) this research examines the bootstrap procedure and its use in the principal component analysis. Unlike Rahoma et al. (2021) and existing academic literature, the author uses the bootstrap procedure to split data into training and test set. She proposes a novel PCA method called the ANOVA-bootstrapped PCA. Using ANOVA and PCA transformation, the number of principal components necessary for building a classification model can be identified immediately. Eigenvectors and eigenvalues are not necessary to extract the important principal components as principal components are combined based on accuracy. Therefore, this research extends (Vrigazova, 2021) by providing more experiments with the ANOVA-bootstrapped PCA for logistic regression.

Research methodology

This section introduces the classical PCA approach that is often used in academic literature (Classical PCA). It is compared with the ANOVA-bootstrapped PCA proposed in this paper. The logistic regression is used to compare two models in terms of accuracy. The author uses Python 3.6 to conduct her experiments. Figure 1 compares the two algorithms.

Classical principal component analysis

The general approach for choosing the number of principal components consists of several steps. It is described in (James et al., 2013), (Mitchel et al., 1997). To run the classical PCA the built-in functions in scikitlearn in Python are used. These include LogisticRegression(), sklearm.decomposition.PCA() and sklearn.model_selection.kFold(). Left part of figure 1 illustrates the classical approach for selecting the number of principal components (James et al., 2013).

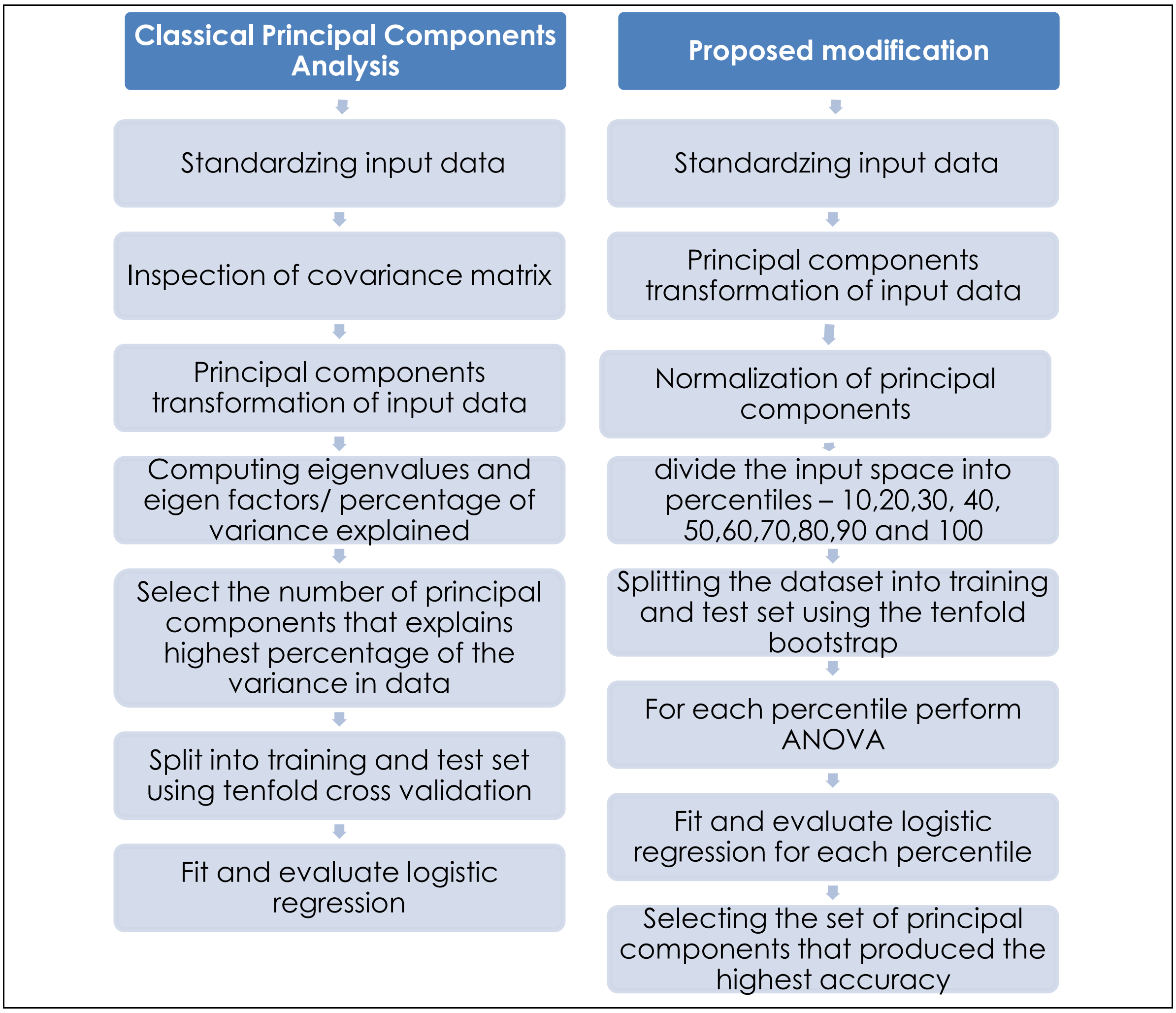
Classical Principal Components Analysis
Standardzing input data
Inspection of covariance matrix
Principal components transformation of input data
Computing eigenvalues and eigen factors/ percentage of variance explained
Select the number of principal components that explains highest percentage of the variance in data
Split into training and test set using tenfold cross validation
Fit and evaluate logistic regression
Proposed modification
Standardzing input data
Principal components transformation of input data
Normalization of principal components
divide the input space into percentiles – 10,20,30, 40, 50,60,70,80,90 and 100
Splitting the dataset into training and test set using the tenfold bootstrap
For each percentile perform ANOVA
Fit and evaluate logistic regression for each percentile
Selecting the set of principal components that produced the highest accuracy

Figure 1 Comparison between the classical approach for selecting the number of principal components and the proposed new algorithm

First, the input variables should be standardized to avoid bias resulting from the measurement unit of the independent variables. Then the covariance matrix should be inspected to identify and remove highly correlated variables that contain noise. The input variables can further be transformed into principal components. Each principal component is a linear combination of input variables (James et al., 2013). The transformed dataset contains as many principal components as the number of the input variables. Identifying the most informative principal components leads to dimensionality reduction as only the important principal components participate in the final model.

To select the important principal components, eigenvectors and eigenvalues are computed. They provide information about the percentage of variance explained by each principal component. The combination of principal components that explain the highest percentage of variance in data is then selected. These steps summarize the process of selecting the number of principal components according to textbooks. They are the same regardless of the type of model and resampling method used. The next steps involve splitting the dataset into training and test set and fitting the logistic regression with the selected principal components. Tenfold cross validation can be used to divide the input data into training and test set to evaluate logistic regression using the principal components selected.

However, an issue occurs in the classical approach: "In case two or more combinations of principal components result in similar accuracy, which one should be selected?". Textbooks provide several solutions to this issue. For example, select the number of principal components based on prior knowledge, the highest accuracy or the combination that provides the smallest number of principal components (James et al., 2013). Therefore, in case two or more combinations of principal components are appropriate, steps 6 and 7 are repeated with each combination of principal components to select the number of principal components that results in the highest accuracy and high percentage of variance explained.

The disadvantage of this approach is that it involves manual operations when several combinations of principal components are possible. Often, the researcher lacks a criterion to select among combinations. Therefore, bias can be introduced to the model. To overcome these disadvantages of the classical approach, a novel algorithm called the ANOVA-BOOTSTRAPPED PCA is proposed in this paper.

New approach - the ANOVA bootstrapped PCA

The Throughout the paper, the ANOVA-BOOTSTRAPPED PCA algorithm will be denoted as ANOVA-Boot-PCA-LR. For running ANOVA, PCA and the logistic regression, existing functions in Python (LogisticRegression(), sklearm.decomposition.PCA() and sklearn.Pipeline (ANOVA)) are used, while a script for running the tenfold bootstrap is created by the author. The tenfold bootstrap used in step 5 and its software realization in Python 3.6 can be found in author's previous study (Vrigazova, 2020). Right part of figure 1 illustrates the proposed algorithm.

Like in the classical approach, the input data are first standardized. Then, PCA transformation is applied to the standardized data. Then, the principal components are normalized between 0 and 1 to avoid negative values in the principal components. The input space is divided into percentiles – 10,20,30, 40, 50,60,70,80,90 and 100. This is necessary to run the proposed algorithm for each percentile of principal components and compare the output. At each percentile the dataset with principal components is split into training and test set in proportion 70/30 using the tenfold bootstrap described in (Vrigazova, Ivanova, 2020). ANOVA is performed to rank the importance of principal components. For each percentile of principal components and iterations of the bootstrap, the accuracy of the logistic regression is averaged. The percentile of principal components that results in the highest accuracy is selected. Then, the number of principal components selected and the accuracy from the ANOVA-Boot-PCA-LR and the classical approach are compared.

The ANOVA ranks the principal components by importance. The ANOVA tests principal components for equality of means with respect to the dependent variable by using the f-statistics. If a set of principal components have equality of means, they are not related to the dependent variable, so they are ranked lower at the table. Principal components that are ranked highest in the table are the ones that have greater inequality of means with respect to the dependent variable. Each percentile contains combination of the most importance principal components as ranked by the ANOVA. For instance, if the 10th percentile corresponds to 3 principal components, then the accuracy is based on the three most important principal components selected by the ANOVA. If the 20th percentile corresponds to 15 principal components, then logistic regression is fitted using the first fifteen principal components as ranked by the ANOVA. Note that the first n most important principal components ranked by the ANOVA are not the first n principal components ranked by index as it is with the classical approach.

In the standard approach the selected number of principal components follows the index of the components, e.g. first two, first three as the first components have bigger variance than the last ones. However, in the ANOVA-BOOTSTRAPPED PCA selecting 3 principal components means the three principal components with the highest ANOVA rank. Therefore, the three identified principal components can be the fifth, seventh and second, for example. This is an important difference between the proposed approach and the classic approach as the interpretation of selected principal components is not the same.

Results and discussion

In (Vrigazova, 2021) the output of three datasets using the ANOVA-bootstrapped PCA is presented. This study extends previous experiments by providing results on three additional datasets. The datasets are public.

The adult income dataset

The adult income dataset is also known as the Census income dataset (Kaggle, 2021a). It contains data about the income of adults exceeding or not $50K/yr (the dependent variable) and 13 independent variables. The classic PCA approach requires calculating the percentage of variance that each principal component (PC), and the components together contribute to explain the variance in the dataset. Table 1 shows the results from the classical approach on the adult income dataset.

The second column in table 1 shows the percentage of variance explained of each principal component. For example, the first principal component alone explains 15.8% of data variance. The second – 9.9%, etc. However, the goal is to find that combination of principal components that explains bigger part of the variance in data. For this purpose, the cumulative percentage of variance explained is necessary. This percentage is calculated in the fourth column. For example, if the first five components are taken, they explain 51% of data variance. If the first 11 principal components are used, then 92% of data variance is explained. The task is to decide what combination of principal components is necessary to fit logistic regression and get the best accuracy.

Table 1 Number of principal components selected in the adult income dataset (textbook approach)

Principal component	% of var explain.	Principal components combination	Cum. var. expla. (%)	Accuracy
Principal component 1	15.8%			
Principal component 2	9.9%	First 2 principal components	26%	-
Principal component 3	8.8%	First 3 principal components	35%	-
Principal component 4	8.0%	First 4 principal components	43%	-
Principal component 5	7.9%	First 5 principal components	51%	-
Principal component 6	7.8%	First 6 principal components	58%	-
Principal component 7	7.4%	First 7 principal components	66%	-
Principal component 8	7.2%	First 8 principal components	73%	-
Principal component 9	7.0%	First 9 principal components	80%	-
Principal component 10	6.5%	First 10 principal components	86%	-
Principal component 11	5.5%	First 11 principal components	92%	82.0%
Principal component 12	5.2%	First 12 principal components	97%	81.9%
Principal component 13	3.0%	First 13s principal components	100%	82.1%

The classic approach (James et al., 2013) advises to select the combination of principal components that explains the highest percentage of variance. At the same time, as the task is a dimensionality reduction task, the aim is to select a smaller number of principal components that are originally contained in the dataset. In the case of table 1 – to select smaller number of 13 principal components. Looking at table 1, several possible combinations of principal components exist. For example, the first 11 and 12 principal components explain respectively, 92% and 97% of the variance in data. On the other hand, the first 10 explain 86%, which can be considered low or high depending on the purpose of the research. In the presented case, the first 11 and 12 principal components explain more than 90% of the variance in data, so one of the two combinations should be chosen.

The last column of table 1 shows the accuracy of the logistic regression using the first 11, 12 and 13 principal components. Although the first 12 principal components account for 97% of the variance in data, their accuracy (81.9%) is very close to that of the first 11 (82%). In this case, the first 11 principal components can be selected. Choosing the first eleven principal components will not lead to loss of accuracy, however, it will lead to a smaller dataset to use for the logistic regression. In comparison, table 2 shows the results of the proposed ANOVA-bootstrapped PCA.

Table 2 Number of principal components: the ANOVA-Bootstrapped PCA logistic regression in the adult income dataset

Percentile	Number of principal components	Accuracy
10%	1	75.8%
20%	3	78.9%
30%	4	79.7%
40%	5	80.5%
50%	7	81.2%
60%	8	81.9%
70%	9	81.9%
80%	10	82.0%
90%	12	82.0%
100%	13	81.9%

Unlike table 1, where the researcher manually needs to select the number of principal components to use in the logistic regression, the ANOVA-Bootstrapped PCA Logistic regression provides automatic solution. Table 2 results directly from the proposed algorithm. Column one shows the percentile of principal components, column two – the corresponding number of principal components and column 3 – the corresponding accuracy of the logistic regression. The criterion for choosing the number of principal components is to select the smallest number of principal components without loss of accuracy.

For instance, choosing 5 principal components is not the optimal choice as the resulting accuracy of 80.5% is not the highest one in the table. On the other hand, choosing 10 principal components is not the best solution although the resulting accuracy is the highest – 82%. Based on table 2, 60% of the principal components (8 principal components) can be selected. Eight principal components are the smallest number of principal components that does not lead to loss of accuracy if the logistic regression is fitted. Eight principal components would result in 81.9% accuracy, which is almost the same as 82% accuracy when 10 principal components are used. Table 2 leads to an automatic decision what number of principal components to use only by looking at the table and applying the criterion: "Select the smallest number of principal components without loss of accuracy".

An important note should be made that choosing 8 principal components does not mean the first eight as it is in the classical approach. The ANOVA provides ranking of the importance of the principal components, so choosing 8 principal components means choosing the first eight principal components that are the most important according to the ANOVA ranking. Table 3 shows the ANOVA ranking of the principal components in the adult income dataset. The first eight principal components in table 3 are selected. Those are the first eight most important according to the ANOVA ranking.

Table 3 Ranking of the importance of principal components in the adult income dataset

Index of principal component	Importance
4	2333.3
8	1736.6
2	1515.8
9	723.9
3	714.1
7	673.6
5	325.2
13	273.7
6	124.6
11	56.3
12	25.6
10	4.9
1	0.8

Another note should be made that the ANOVA ranking of the principal components remains the same regardless of the classification model used. Running a different classification model, for example – decision tree, would not change the ranking of principal components. Rather it will change the accuracy of the model and the combination of principal components to take so that the highest possible accuracy using the decision tree might be achieved. So, the ANOVA-Bootstrapped

PCA selects the number of principal components that would achieve the best performance given the classification model applied.

The EPICA Dome C Ice Core 800KYr Temperature Estimates (ED) dataset

Similar experiments are conducted on the EPICA Dome C Ice Core 800KYr Temperature Estimates (ED) dataset (Vincentarelbundock, 2021). The dataset contains temperature record from the EPICA (European Project for Ice Coring in Antarctica) Dome C ice core covering 0 to 800 kyr BP. It has 5 independent variables/ principal components respectively. Table 4 shows the output from the classical approach.

As table 4 shows the first two principal components explain 97.4% of the variance in the data. The contribution of the other three components is very small. The author runs the logistic regression with the first 2 principal components and the first 3 principal components. The resulting accuracy is respectively 99% and 98%. In this case, she selects the first two components as they explain high percentage of the variance in data (97.4%) and results in the best accuracy - 99%. Choosing the number of principal components in table 4 can be confusing. Selecting the combination of principal components that explains the highest percentage in the variance, would be inefficient as this combination may not result in the best accuracy. Therefore, the researcher manually should decide which number of principal components should be selected.

Table 4 Number of principal components in the ed dataset: classical approach

Principal component	% of var., expl.	Principal component combination	Cum. var., expl. (%)	Accuracy
Principal component 1	57.5%			
Principal component 2	39.9%	Principal components 1+2	97.4%	99%
Principal component 3	2.5%	Principal components 1+2 + 3	99.9%	98%
Principal component 4	0.1%	Principal components 1+2+3+4	100.0%	
Principal component 5	0.0%	Principal component 1+2+3+4+5	100.0%	

That is not the case with the ANOVA-Bootstrapped PCA, which provides automatic selection of the number of principal components. Table 5 shows the output in the ED dataset.

Table 5 The ANOVA-Bootstrapped PCA logistic regression in the ED dataset

Percentile	Number of principal components	Accuracy
10%	1	96.6%
20%	1	96.6%
30%	2	97.5%
40%	2	97.5%
50%	3	97.5%
60%	3	98.0%
70%	4	98.0%
80%	4	98.0%
90%	5	98.0%
100%	5	97.4%

By following the rule to select the smallest number of principal components resulting to the best accuracy, table 5 identifies 3 principal components to be selected. They result in accuracy of 98%. Table 6 shows the index of the principal components selected based on the ANOVA ranking.

Table 6 Ranking of the principal components in the ED dataset

Principal component index	Importance
3	3045.4
4	41.1
1	30.3
5	8.1
2	2.2

To achieve the result in table 5, the first, third and fourth principal components are selected, unlike the classical approach, where the first two PCs are selected.

The Monica dataset

This dataset is also called the 'MONICA WHO' dataset (Kaggle, 2021b) and it was created in 1980 to record data in cardiovascular disease in 21 countries for 10 years. It contains 11 principal components. Table 7 shows the output from the classical approach.

Table 7 Classical approach for choosing the number of principal components in the Monica dataset

Principal component	% of var. expl.	Principal component combination	Cum. var. expl. (%)	Accur.
Principal component 1	40.4%			
Principal component 2	11.0%	First two principal components	51%	69.1%
Principal component 3	9.5%	First three principal components	61%	73.1%
Principal component 4	8.7%	First four principal components	70%	75.1%
Principal component 5	7.6%	First five principal components	77%	87.3%
Principal component 6	6.2%	First six principal components	83%	87.5%
Principal component 7	5.2%	First seven principal components	89%	87.5%
Principal component 8	3.7%	First eight principal components	92%	87.6%
Principal component 9	3.3%	First nine principal components	96%	87.6%
Principal component 10	2.7%	First ten principal components	98%	87.5%
Principal component 11	1.7%	First eleven principal components	100%	87.5%

Table 7 shows that using only the cumulative percentage of variance explained may be misleading. The first six principal components account for 83% of the variance in data. As the end of the table comes, the cumulative percentage of variance explained increases so that the first nine principal components account for 96% of variance explained. Should the first six, seven, eight, nine or ten principal components be selected? Based only on the cumulative variance explained, the researcher can select the first ten principal components as they account for 98% of the variation of the data. But this number would not reduce the size of the dataset.

On the other hand, when the accuracies achieved in table 7 are considered, it turns out that whether 6,7,8,9,10 or 11 principal components are selected, the accuracy remains almost unchanged (87.6%, 87.5%). Therefore, the researcher should decide whether to select bigger number of principal components or a smaller number. As the author's purpose is to perform dimensionality reduction, the first six principal components are selected, which account for 83% of the variance in data but the resulting accuracy is similar to that of a bigger number of principal components. However, another researcher may select, for example the first nine components as they explain 96% of the variance in data and the accuracy achieved is higher – 87.6%. As table 7 shows, the highest variance explained may not guarantee the best accuracy. Also, table 7 shows that the seventh, eighth, nineth, tenth, eleventh and twelfth principal components may be redundant as the accuracy remains unchanged, despite them explaining higher percentage of the variance in data. As in the previous case, the researcher should manually decide what number of principal components to take.

Table 8 Automatic selection of principal components in the Monica dataset

Percentile	Number of principal components	Accuracy
10%	1	65.9%
20%	2	77.5%
30%	3	82.9%
40%	4	84.9%
50%	6	88.3%
60%	7	86.7%
70%	8	87.8%
80%	9	88.2%
90%	10	86.8%
100%	11	87.0%

Table 9 Principal components automatically selected in the Monica dataset

Principal component index	Importance
3	1056
10	790
6	431
7	299
9	216
1	113
5	94
2	28
8	26
11	1
4	0

As table 8 shows the proposed algorithm gives a straightforward answer: the researcher needs to select six principal components to achieve accuracy of 88.3%. Using table 8, the manual application of criteria as it was in table 7 can be avoided. Also, the researcher should not decide whether to take smaller or bigger number of principal components when the accuracy remains similar. Instead, table 8 provides an automatic answer to the question:" How many principal components should be used for classification?". Table 8 shows the output of the automatic algorithm proposed. Table 9 shows the indices of the principal components used to achieve accuracy of 88.2%.

Discussion

Machine learning textbooks and many authors recommend detecting the number of principal components based on the percentage of variance explained (James et al., 2013) but this algorithm is not automatic. It is an issue when the percentage of variance explained identifies two possible numbers of principal components and their accuracy is close. The classic approach recommends manually deciding which number is most appropriate. Many authors try to solve this problem by modifying parts of the equation of the PCA (for instance, eigenvectors, eigenvalues, etc.) or by using variable selection with PCA (Salata, Grillenzoni, 2021; Pacheco et al., 2013; Kim, Rattakorn, 2011). However, none of the algorithms provides automatic selection of principal components. The algorithm proposed in this paper, however, fixes this issue by conducting automatic principal component selection. It performs automatic selection of the number of principal components to be used in the logistic regression.

This advantage of the author's algorithm is important in big datasets where the number of principal components can be big. The traditional manual selection of principal components can result in several alternatives and the choice between them can be hard and time consuming. For instance, if the traditional theory outlines the first five, six, seven or eight principal components as possible options, the researcher needs a criterion to choose among them. In many cases, the criteria are subjective. On the other hand, the ANOVA-Boot-PCA-LR algorithm proposed in this paper removes the subjectiveness coming from manually choosing the number of principal components. This makes the choice of principal components faster.

The author's approach has several limitations that need to be further researched, though. First, experiments with other classification methods are necessary to test the feasibility of the ANOVA-Boot-PCA-LR to other classification models. This would expand the practical applications of the model proposed and show its universal nature, outlining in what cases it can and cannot be applied to other classification methods. Second, the resulting selection of principal components depends on the accuracy of the model. The accuracy would change when either the classification model type changes or when its parameters change.

Despite this, the proposed algorithm provides the first step to an efficient and fast automatic algorithm for principal components selection in classification problems. Thus, the manual selection of principal components can be avoided, and better accuracy achieved, which is a novel approach in academic literature.

As a conclusion, the author develops a simple algorithm for automatic detection of the number of principal components to be used in the logistic regression. The advantages of this algorithm include simplicity as it is based on existing algorithms, easiness to interpret and providing more efficient results in terms of accuracy.

The ANOVA-BOOT-PCA algorithm can be viewed as an extension of the textbook approach for finding the number of principal components. However, instead of choosing them manually when several combinations of principal components are possible, the ANOVA-bootstrapped PCA decides automatically.

Conclusion

The classical PCA approach for selecting the number of principal components has a set of disadvantages. First, it involves manual selection of principal components in cases when the variance explained is similar or the accuracy from several combinations of principal components is similar. Therefore, the researcher introduces bias into the dataset. Second, in many cases, using the variance explained may not be enough to select the number of principal components. Calculation of accuracy may help deciding what combination to select. However, calculating the accuracy of all possible combinations of principal components may be computationally exhaustive. Also, the accuracy and the variance explained may not facilitate the researcher to decide the number of principal components. Therefore, an automatic way for selection of the number of principal components should be applied.

In this research the author proposes the ANOVA-Bootstrapped PCA to select the number of principal components and then fit a classification model (the logistic regression, in the presented example). The advantages of the ANOVA-Bootstrapped

PCA algorithm are automatic identification of the number of principal components, automatic overview of how the accuracy varies across different principal components combinations. This advantage can be used for data exploratory purposes.

The proposed algorithm also provides ranking of the importance of the principal components and gives the right combination to achieve the best accuracy. The right combination of principal components may not be the first several components but rather principal components with different indices. However, the number of principal components selected according to the ANOVA-Bootstrapped PCA depends on the classification model selected. So, different classification models would lead to different accuracy and number of principal components selected. However, the importance of the principal components as identified by the ANOVA model would not change.

References

1. Breiman L. (1995). Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, Vol. 37, No. 4, pp. 373-384.
2. Gajjar S., Kulahci M., Palazoglu A. (2017). Selection of non-zero loadings in sparse principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, Vol. 162, pp. 160-171.
3. James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
4. Kaggle (2021a). *Adult Income dataset*. Available at https://www.kaggle.com/wenruliu/adult-income-dataset [01 July 2021].
5. Kaggle (2021b). *Monika dataset*. Available at https://www.kaggle.com/ukveteran/who-monica-data [01 July 2021].
6. Kim, S., Rattakorn, P. (2011). Unsupervised feature selection using weighted principal components. *Expert Systems with Applications*, Vol. 38, No. 5, pp. 5704-5710.
7. Maleki, N., Zeinali, Y., Niaki, S.T.A. (2020). A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. *Expert Systems with Applications*, Vol. 164.
8. Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York.
9. Pacheco, J., Casado, S., Porras, S. (2013). Exact methods for variable selection in principal component analysis: Guide functions and pre-selection. *Computational Statistics & Data Analysis*, Vol. 57, No. 1, pp. 95-111.
10. Prieto-Moreno, A., Llanes-Santiago, O., García-Moreno, E. (2015). Principal components selection for dimensionality reduction using discriminant information applied to fault diagnosis. *Journal of Process Control*, Vol. 33, pp. 14-24.
11. Rahoma, A., Imtiaz, S., Ahmed, S. (2021). Sparse principal component analysis using bootstrap method. *Chemical Engineering Science*, Vol. 246.
12. Salata, S., Grillenzoni, C. (2021). A spatial evaluation of multifunctional Ecosystem Service networks using Principal Component Analysis: A case of study in Turin, Italy. *Ecological Indicators*, Vol. 127, pp. 1-13.
13. Sharifzadeh, S., Ghodsi, A.,Clemmensen, L., Ersbll B. (2017). Sparse supervised principal component analysis (SSPCA) for dimension reduction and variable selection. *Engineering Applications of Artificial Intelligence*, Vol. 65, pp. 168-177.
14. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, Vol. 73, No. 3, pp. 267-268,
15. Vincentarelbundock (2021). *EPICA Dome C Ice Core 800KYr Temperature Estimates dataset*. Available at https://vincentarelbundock.github.io/Rdatasets/datasets.html [01 July 2021].
16. Vrigazova, B. (2021). Novel Approach to Choosing Principal Components Number in Logistic Regression. *ENTRENOVA-ENTerprise REsearch InNOVAtion*, Vol. 7, No. 1, pp. 1-12.

17. Vrigazova, B., Ivanov, I. (2020). Tenfold bootstrap procedure for support vector machines. *Computer Science*, Vol. 21, No. 2, pp. 241-257.
18. Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, Vol. 101, No. 476, pp. 1418-1429.

About the author

Borislava Toleva obtained a PhD in Data Science at the Faculty of Economics and Business Administration, Sofia University, Bulgaria. She obtained a master's degree in Statistics, financial econometrics and actuarial studies in 2015 after a bachelor's degree in Economics at the same university. Her research areas include practical applications of machine learning algorithms for prediction and how their performance can be boosted. Also, applications of big data techniques to small datasets in the field of economics as alternative to traditional econometrics theory. She challenges traditional econometric modelling techniques used to find connections among variables from institutional economics by combining feature selection methods and big data prediction models. As a result, new applications of machine learning techniques to economic data appear. The author can be contacted at vrigazova@uni-sofia.bg.