DRIVING INSIGHTS: DIMENSIONALITY REDUCTION WITH PRINCIPAL COMPONENT ANALYSIS ON AUTOMOBILE MPG DATASET

Abstract

Dimensionality reduction is a fundamental technique in machine learning and data analysis to effectively handle high-dimensional datasets. In this study, we focus on applying Principal Component Analysis (PCA) to reduce the dimensionality of the Auto MPG dataset, a popular dataset that contains information about various car models and their fuel efficiency. The objective of this research is to explore the application of PCA for dimensionality reduction in the context of the Auto MPG dataset. PCA works by transforming the original features into a new set of uncorrelated variables called principal components. By selecting a subset of these components, we can retain the most informative aspects of the data while reducing its dimensionality. Through the application of PCA to the Auto MPG dataset, we aim to achieve two primary outcomes. Firstly, we aim to identify the most influential features that contribute significantly to the variation in fuel efficiency across different car models. Secondly, we seek to determine the optimal number of principal components to retain, striking a balance between dimensionality reduction and preserving the information necessary for accurate analysis. By implementing PCA and evaluating its performance on the Auto MPG dataset, we hope to provide insights into the potential benefits of dimensionality reduction techniques for automotive data analysis. The findings of this study could have implications for enhancing fuel efficiency, understanding the key factors affecting it, and improving decision-making processes related to automobile design and engineering.

Keywords: Primary components, PCA object, dataset, data analysis.

Author

Dr. Renuka Sagar
Associate Professor
Department of Artificial Intelligence and Machine Learning
Ballari Institute of Technology and Management
Jnana Gangothri Campus
Bellary, Karnataka, India.
renukasagar83@gmail.com

I. INTRODUCTION

Although automobiles have been around for a while, the first one was created in 1769. That occurred more than three centuries ago. The modern automobile is distinctive and special due to its many diverse features. A car's average age can be anywhere from a few to 17 years. There are numerous methods for calculating MPG, one of the key elements of the vehicle. Principal component analysis is one of the most often used approaches to calculating MPG. In this study, fuel consumption is calculated using principal component analysis. A linear algebraic method called principal component analysis (PCA) can be used to decrease the dimensionality of a set of features. The singular value decomposition and other linear combinations of features can be discovered using PCA, which is especially helpful for exploratory data analysis and visualization. PCA is a versatile tool that is most frequently used for multivariate data processing and visualisation. Each car's mileage over a certain time period is recorded in the automotive dataset MPG. For every car in the dataset, there are a total of 30 variables. Finding the variables that affect an automobile's MPG is the major goal of this investigation. A statistical method known as principal component analysis (PCA) identifies the linear combinations of the original variables that are most beneficial. It is employed in a variety of industries, including marketing, engineering, chemistry, and also biology In order to uncover intriguing patterns and insights, a mathematical methodology known as principal component analysis is used in this study on a dataset of automotive gasoline usage.

II. LITERATURE REVIEW

The literature review for a study on Principal Component Analysis (PCA) applied to an automobile MPG dataset using dimensionality reduction techniques could encompass the following key areas:

1. Dimensionality Reduction [1]and PCA is used to define dimensionality reduction and its significance in handling high-dimensional datasets. It also discusses the motivation behind using PCA as a popular linear dimensionality reduction technique. It also highlights the mathematical foundation of PCA and how it captures variance to create new orthogonal features.

2. Applications of PCA in Automotive Domain [2] , It Explores previous studies where PCA has been applied to automotive datasets. This study Discusses instances where PCA led to insights in areas such as vehicle performance analysis, emissions reduction, and fuel efficiency improvement.

3. Analysis of Automobile Datasets [3], this survey Presents notable studies that have used automobile-related datasets for analysis and modeling. It Discuss the types of data commonly found in automobile datasets, including features related to vehicle specifications, engine performance, and emissions.

4. PCA and Fuel Efficiency Analysis [4], This paper Showcase research that utilized PCA to analyze vehicle fuel efficiency. It also explains how PCA can identify key features that contribute most to fuel efficiency variations.

5. PCA in Emissions Analysis [5], This research discusses the literature that employs PCA to analyze emissions data from vehicles. It also Highlight how PCA can reveal underlying emission patterns and identify contributing factors.

6. Dimensionality Reduction for Automotive Insights [6], This paper discusses other dimensionality reduction techniques beyond PCA that have been applied in the automotive domain. It also compares PCA with other methods in terms of effectiveness, interpretability, and computational efficiency.

7. Challenges and Considerations [7], It Explores challenges specific to applying dimensionality reduction techniques to automotive datasets, such as missing data or categorical variables.

   It also discusses on how researchers have addressed these challenges and any trade-offs made.

8. Impact on Industry and Research [8], this research paper analyze the impact of dimensionality reduction techniques on the automotive industry, including areas like vehicle design, optimization, and diagnostics. It also highlights how insights gained from reduced-dimensional datasets have influenced decision-making and innovation.

9. Future Directions [9][10], This survey paper Propose potential future research directions in applying PCA and other dimensionality reduction techniques to automotive datasets. It Suggests areas for exploration, such as integrating PCA with machine learning models or investigating non-linear dimensionality reduction methods.

III. METHODOLOGY

A method for extracting sets of orthogonal (uncorrelated) variables is principal component analysis (PCA), which can subsequently be used to reduce the dimensionality of a data set. A common method in machine learning is PCA. A data set's dimensionality can be decreased using PCA, a linear treatment that extracts orthogonal variables.
I used PCA for the vehicle mpg dataset

1. **Data Preprocessing:** Start by importing the dataset and performing any necessary data cleaning, handling missing values, and scaling the data if required. Ensure that the dataset is in a suitable format for PCA analysis.

2. **Feature Selection:** Decide which features from the dataset you want to include in the PCA analysis. Depending on the specific objectives and characteristics of the dataset, you may want to exclude certain features that are irrelevant or redundant.

3. **Standardization:** Perform standardization on the selected features, which involves scaling them to have zero mean and unit variance. This step is crucial for PCA, as it ensures that the features are on a similar scale and prevents variables with larger ranges from dominating the analysis.

4. **PCA Application:** Apply PCA to the standardized feature matrix. This can be done using libraries such as scikit-learn in Python. PCA will transform the original feature space into a new set of orthogonal variables called principal components.

5. **Explained Variance:** Analyze the explained variance ratio associated with each principal component. The explained variance indicates the amount of information retained by each component. Plotting a scree plot or cumulative explained variance plot can help determine the number of principal components to retain.

6. **Selecting Components:** Based on the explained variance analysis, decide on the number of principal components to retain. You can choose a threshold (e.g., retaining components that explain 95% of the variance) or use domain knowledge to make an informed decision.

7. **Dimensionality Reduction:** Transform the original dataset into a reduced-dimensional space using the selected principal components. This reduced dataset will have fewer dimensions than the original dataset while still capturing a significant amount of information.

By applying PCA for dimensionality reduction on the AutoMPG dataset, you can reduce the number of features while preserving the most relevant information, facilitating further analysis and interpretation.

IV. EXPERIMENTAL SETUP

1. **Dataset Selection:** Choose a suitable publicly available automobile MPG dataset that contains relevant attributes such as vehicle specifications, engine details, and fuel efficiency measurements. During Data Preprocessing Handled missing values by imputation or removal, and the rationale for the chosen approach. Performed data normalization or standardization to ensure all features have comparable scales.

2. **PCA Application:** Implement the PCA algorithm on the preprocessed dataset using a programming language like Python the number of principal components to retain based on an analysis of explained variance. For Visualization Created screed plots or cumulative explained variance plots to visualize the proportion of total variance explained by each principal component. Generate scatter plots or biplots to visualize the distribution of data points in the reduced-dimensional space.

3. **Analysis and Interpretation:** Interpret the principal components in terms of the original features to understand the underlying patterns captured by each component. Analyze the loadings of variables on each principal component to identify which attributes contribute most to the variance.

4. **Performance Evaluation:** The study involves a predictive task, such as regression for predicting MPG, split the dataset into training and testing sets. Trained a regression model (e.g., linear regression) on both the original dataset and the reduced-dimensional dataset obtained from PCA. Evaluate and compare the model performance on the test set using appropriate metrics like Mean Squared Error (MSE) or R-squared. Interpret the

results of the regression models to understand how the reduced-dimensional dataset from PCA affects predictive accuracy and model interpretability. compare the results of PCA with a baseline model that does not involve dimensionality reduction (e.g., using all original features). The trade-offs between dimensionality reduction, model performance, and interpretability. The findings from the experiment, including insights gained from the reduced-dimensional dataset and implications for using PCA in analyzing automobile MPG data has been carried out.

V. RESULTS & DISCUSSION

When using PCA for dimensionality reduction on the AutoMPG dataset, several outcomes have been observed:

1. **Reduced Dimensionality:** PCA will transform the original high-dimensional feature space into a lower-dimensional space. The number of dimensions will be determined by the number of principal components we choose to retain.

2. **Explained Variance:** PCA provides information about the amount of variance explained by each retained principal component. This allows us to understand the contribution of each component to the overall dataset variance and identify the most influential features.

3. **Feature Selection:** By examining the weights or loadings of each feature in the principal components, we determine which original features are most relevant for explaining the variation in the AutoMPG dataset. Features with higher absolute loadings in the retained components have a stronger influence on the data.

4. **Visualization:** With the reduced dimensionality, we created scatter plots or other visualizations to observe the distribution of car models based on their fuel efficiency and the retained principal components. This can help identify patterns or clusters in the data.

Principal Component Analysis Based on Covariance Matrix

Descriptive Statistics

	mpg	cylinders	displacement	horsepower	weight	acceleration	model	year	origin
Mean	23.224	10.997	11.452	19.820	20.773	11.043	11.779	14.906	12.917
Variance	94,733.734	30,399.254	30,158.120	64,025.394	64,239.027	32,411.895	37,742.916	50,341.828	47,013.041
S.D	307.788	174.354	173.661	253.032	253.454	180.033	194.275	224.370	216.825

Covariance Matrix

	mpg	cylinders	displacement	horsepower	weight	acceleration	model	year	origin
mpg	94,733.734	1,295.045	3,923.369	-248.344	5,531.903	44.880	5,329.900	7,216.801	49,584.829
cylinders	1,295.045	30,399.254	1,540.383	6,408.371	-10.429	341.827	10.130	4,647.741	1,814.769
displacement	3,923.369	1,540.383	30,158.120	2,949.142	6,268.671	37.523	350.851	180.012	3,283.674
horsepower	-248.344	6,408.371	2,949.142	64,025.394	2,559.591	4,492.834	-67.163	3,112.493	42.364
weight	5,531.903	-10.429	6,268.671	2,559.591	64,239.027	1,687.476	4,815.720	424.710	414.688
acceleration	44.880	341.827	37.523	4,492.834	1,687.476	32,411.895	2,067.622	5,620.058	18.461
model	5,329.900	10.130	350.851	-67.163	4,815.720	2,067.622	37,742.916	2,666.502	5,321.313
year	7,216.801	4,647.741	180.012	3,112.493	424.710	5,620.058	2,666.502	50,341.828	2,734.448
origin	49,584.829	1,814.769	3,283.674	42.364	414.688	18.461	5,321.313	2,734.448	47,013.041

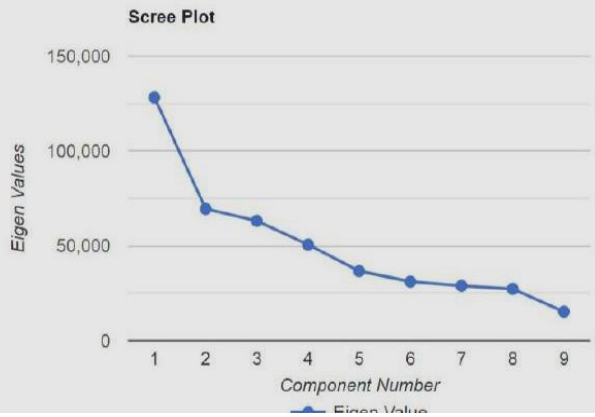
Correlation Matrix

	mpg	cylinders	displacement	horsepower	weight	acceleration	model	year	origin
mpg	1.000	0.024	0.073	-0.003	0.071	0.001	0.089	0.105	0.743
cylinders	0.024	1.000	0.051	0.145	-0.000	0.011	0.000	0.119	0.048
displacement	0.073	0.051	1.000	0.067	0.142	0.001	0.010	0.005	0.087
horsepower	-0.003	0.145	0.067	1.000	0.040	0.099	-0.001	0.055	0.001
weight	0.071	-0.000	0.142	0.040	1.000	0.037	0.098	0.007	0.008
acceleration	0.001	0.011	0.001	0.099	0.037	1.000	0.059	0.139	0.000
model	0.089	0.000	0.010	-0.001	0.098	0.059	1.000	0.061	0.126
year	0.105	0.119	0.005	0.055	0.007	0.139	0.061	1.000	0.056
origin	0.743	0.048	0.087	0.001	0.008	0.000	0.126	0.056	1.000

Gleason-Staelin Phi0.141985594324081

Eigenvalues of Covariance Matrix

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Eigenvalues	**128,143.251**	**69,561.621**	**63,243.634**	**50,705.964**	**36,702.329**	31,110.400	28,959.429	27,333.653	15,304.928
Proportion	0.284	0.154	0.140	0.112	0.081	0.069	0.064	0.061	0.034
Cumulative Proportion	0.284	0.438	0.579	0.691	0.772	0.841	0.905	0.966	1.000



Eigen value

Loadings (Eigenvectors) of Covariance Matrix

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
mpg	**0.834**	-0.074	-0.026	-0.072	-0.089	0.081	-0.072	-0.012	**-0.522**
cylinders	0.027	0.144	-0.138	0.093	-0.113	**-0.520**	0.303	**-0.756**	-0.068
displacement	0.058	0.141	0.082	-0.053	-0.113	-0.312	**0.746**	**0.545**	-0.069
horsepower	0.012	**0.739**	**-0.559**	-0.317	0.044	0.033	-0.162	0.110	0.003
weight	0.089	**0.581**	**0.777**	0.006	-0.133	0.043	-0.129	-0.095	0.070
acceleration	0.011	0.150	-0.069	0.217	0.185	**0.743**	**0.521**	-0.260	-0.014
model	0.088	0.088	0.106	0.157	**0.935**	-0.250	-0.032	0.049	-0.094
year	0.102	0.186	-0.193	**0.897**	-0.189	-0.086	-0.164	0.201	0.060
origin	**0.523**	-0.075	-0.059	-0.084	0.059	-0.031	0.069	-0.030	**0.837**

Correlation of Principal Components with Original Variables

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	Communality k=5
mpg	0.970	-0.064	-0.022	-0.053	-0.056	0.046	-0.040	-0.006	-0.210	0.952
cylinders	0.056	0.217	-0.199	0.120	-0.124	-0.526	0.296	-0.716	-0.048	0.120
displacement	0.119	0.215	0.119	-0.069	-0.124	-0.317	0.731	0.518	-0.049	0.095
horsepower	0.017	0.770	-0.556	-0.282	0.033	0.023	-0.109	0.072	0.002	0.983
weight	0.126	0.605	0.771	0.005	-0.101	0.030	-0.086	-0.062	0.034	0.987
acceleration	0.021	0.220	-0.096	0.272	0.197	0.728	0.493	-0.239	-0.010	0.171
model	0.163	0.120	0.137	0.182	0.922	-0.227	-0.028	0.042	-0.060	0.942
year	0.163	0.219	-0.216	0.900	-0.162	-0.067	-0.124	0.148	0.033	0.957
origin	0.863	-0.091	-0.068	-0.087	0.052	-0.025	0.054	-0.023	0.478	0.768

Principal Component Scores from Covariance Matrix

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
O1	330.108	2,149.605	2,679.631	-34.427	-434.361	49.332	-245.406	-155.748	221.398
O2	81.138	2,843.568	-2,159.176	-855.896	50.602	-85.373	-392.685	276.735	8.915
O3	40.022	2,590.885	-1,953.419	-1,048.320	101.725	-46.111	-322.915	203.935	-15.361
O4	591.348	482.000	282.381	-218.167	-412.170	-1,131.909	2,583.407	1,742.268	-146.399
O5	328.994	574.499	-611.788	650.532	-462.970	-1,815.532	975.645	-2,516.019	-116.922
O6	3,801.683	18.974	176.566	69.626	-531.296	309.249	-382.904	14.335	-2,024.457
O7	5,921.284	-579.384	-350.174	-424.770	219.806	70.780	-16.278	-87.655	1,305.130
O8	460.605	903.157	-874.155	4,093.103	-556.328	-131.086	-483.466	729.777	238.772
O9	441.449	561.928	685.814	633.334	3,571.691	-806.308	-73.533	110.712	-314.907
O10	36.017	900.983	-340.825	709.570	642.306	2,649.708	1,763.772	-892.213	-33.333
O11	330.535	2,238.266	2,748.451	-44.429	-452.021	41.703	-242.006	-142.916	216.148
O12	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704
O13	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704
O14	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704
O15	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704
O16	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704
O17	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704
O18	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704
O19	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704
O20	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704
O21	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704
O22	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704
O23	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704
O24	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704
O25	-31.864	-32.692	-0.730	-9.098	-4.477	2.308	-8.154	1.847	1.704

VI. CONCLUSION

A common statistical method for reducing dimensions while retaining as much variance as possible within a given data set is principal component analysis (PCA) (Pearson, 1901; Hotelling, 1933; Ringner, 2008). Principal components analysis (PCA) is simply a

straightforward dimensionality reduction technique that converts the columns of the dataset into a new feature group (PCs). We are interested in the features of the data, and PCA, which is closely related to factor analysis, frequently leads to comparable results. The first principal components produced by PCA, as the name implies, are those that capture the most information or variance in the data set. The typical setting for PCA as a tool for data analysis and exploration involves a data. An observational data set containing P distinct variables for each of N entities or individuals and P different variables for each of N entities or individuals constitutes the typical context for PCA as an analysis tool for data exploration. Here, we employ some PCA optimality criteria to select the best subsets of the p output variables—the main variables. Imagine running PCA on a dataset with hundreds of variables and seeing that the first few components represent the majority of the explained variance, the NxP data matrix column choices, X, that capture the most variance. The primary components of a data set are these two linear combinations. First, I initialise a PCA class and call fit transform just on X. to transform X to obtain the new set of X's principle components, then compute the weights of the primary components all at once. the sklearn.decomposition module offers the PCA object that can be simply mapped and converts data to the principle components, i.e., it initialises a PCA object from sklearn and changes the data in accordance with the computed components. When the principal component input features are entered, the customised PCA object's Inverse Transform method returns the original data. A lesser quantity of information, but more nuance, is added to the data set with each subsequent component.

REFERENCES

[1] Indirect PCA Dimensionality Reduction Based Machine Learning Algorithms for Power System Transient Stability Assessment, Publisher: IEEE
[2] Dimensionality Reducti on Techniques to Aid Parallelization of Machine Learning Algorithms, Publisher: IEEE
[3] Efficient Machine Learning Model for DDoS Detection System Based on Dimensionality Reduction Saad Ahmed Dheyab 1, Shaymaa Mohammed Abdulameer 2, Salama Mostafa.
[4] On the Use of Machine Learning Models for Prediction of Compressive Strength of Concrete: Influence of Dimensionality Reduction on the Model Performance Zhi Wan 1, Yading Xu 1, Branko Šavija 1 PMID: 33546376 PMCID: PMC7913490 DOI 10.3390/ma14040713 Free PMC article
[5] Hybridized Dimensionality Reduction Method for Machine Learning based Web Pages Classificat ion, Thabit Sulaiman Sabbah,IJCCCE.
[6] Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer.This article provides an in-depth understanding of PCA, including its concepts, challenges, and considerations.
[7] Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459. A comprehensive review of PCA, covering its applications, challenges, and strategies to address various issues.
[8] Kumar, N., & Kaur, P. (2019). Principal Component Analysis: A Comprehensive Review. Archives of Computational Methods in Engineering, 26(4), 891-918. An overview of PCA's challenges, along with discussions on its applications and strategies to optimize its usage.
[9] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems, 2(1-3), 37-52. Insightful exploration of PCA's challenges and considerations, particularly in the context of chemometrics and data analysis.
[10] Chandramouli, M., & Jeganathan, P. (2015). Principal component analysis and its applications in image compression – A survey. Optik, 126(14), 1135-1145. While focused on image compression, this study discusses challenges and considerations relevant to PCA in the context of data reduction.

