Analyzing Air Quality Index Using Dimensionality Reduction Techniques

REPORT

Numerical Linear Algebra for Big Data

Instructor: Dr. Upendra Prasad

Author: Tarun Sai Reddy Kummetha

20020482

TABLE OF CONTENTS

1.Abstract

The Air Quality Index (AQI) is an essential metric for monitoring and analyzing air pollution, which directly affects public health and environmental well-being. High-dimensional datasets, often comprising various pollutant levels and meteorological variables, pose challenges in data analysis and predictive modeling due to their complexity and computational demands. Dimensionality reduction techniques such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) address these issues by condensing data into a smaller set of representative components. These methods retain the most critical information while simplifying data structures, making them highly effective for interpreting AQI trends and improving modeling efficiency.

Among the techniques explored, SVD demonstrates superior performance in handling AQI data. It effectively preserves variance and minimizes data loss, resulting in enhanced predictive accuracy with lower Mean Squared Error (MSE) and higher R^2 scores compared to PCA. By identifying the most influential pollutants, such as CO(GT), NOx(GT), and C6H6(GT), SVD not only streamlines the dataset but also highlights key factors contributing to air quality variations. These findings underscore the value of dimensionality reduction in environmental analytics, enabling policymakers and researchers to focus on critical variables for designing targeted pollution control strategies and improving air quality forecasting systems.

2.Introduction

Problem Statement

Air pollution has become one of the most significant global environmental challenges, directly impacting public health, ecosystems, and climate. The Air Quality Index (AQI) is a widely used indicator that measures pollution levels by aggregating various pollutant concentrations, such as carbon monoxide, nitrogen oxides, and particulate matter. While AQI provides a critical snapshot of air quality, analyzing the vast amount of data required to compute it is often complex. These datasets typically include numerous interrelated features, such as pollutant levels and meteorological conditions, creating high-dimensional data structures that are difficult to interpret and computationally intensive to process. As a result, there is a growing need for techniques that can simplify these datasets while retaining their essential information..

Dimensionality reduction techniques, including Principal Component Analysis (PCA) and Singular Value Decomposition (SVD), offer powerful solutions for addressing these challenges. These methods transform high-dimensional data into a smaller number of representative components, capturing the most critical patterns and relationships within the data. PCA relies on eigen decomposition of the covariance matrix to identify principal components that explain the largest variance, while SVD uses singular vectors and values to efficiently decompose and reconstruct data. By applying these techniques, researchers can streamline AQI datasets, improve computational efficiency, and enhance the accuracy of predictive models. This study explores the application of PCA and SVD to AQI datasets, aiming to reduce dimensionality effectively and evaluate their performance in predictive modeling.

Objective

The primary objective of this study is to explore the application of dimensionality reduction techniques, specifically Principal Component Analysis (PCA) and Singular Value Decomposition (SVD), in analysing Air Quality Index (AQI) datasets. The goals include:

1. Simplifying high-dimensional AQI datasets by reducing the number of features while retaining the majority of meaningful variance.
2. Identifying the most significant pollutants and meteorological factors contributing to air quality variations.
3. Evaluating and comparing the performance of PCA and SVD in predictive modelling tasks, focusing on metrics such as Mean Squared Error (MSE) and R^2 scores.
4. Enhancing computational efficiency and interpretability in AQI data analysis for informed decision-making and environmental management.

3.Dataset Description

The dataset used in this project is sourced from the UCI Air Quality Index Dataset, which consists of air quality data from different times. It is a widely used dataset for finding outputs for the various air quality index calculations.

Features

Link: https://archive.ics.uci.edu/dataset/360/air+quality

The Air Quality dataset contains 9,471 rows and 17 columns, with measurements related to air pollutants and meteorological variables collected over time. Below is a detailed description of the dataset:

1. **Columns**:
   - **Date**: The date of measurement (format: day/month/year).
   - **Time**: The time of measurement (format: hour:minute).
   - **CO(GT)**: Carbon monoxide concentration in mg/m³ (ground truth).
   - **PT08.S1(CO)**: Sensor response related to CO.
   - **NMHC(GT)**: Non-methane hydrocarbons concentration in µg/m³.

- **C6H6(GT)**: Benzene concentration in µg/m^3 (ground truth).
- **PT08.S2(NMHC)**: Sensor response related to NMHC.
- **NOx(GT)**: Nitrogen oxides concentration in ppb (ground truth).
- **PT08.S3(NOx)**: Sensor response related to NOx.
- **NO2(GT)**: Nitrogen dioxide concentration in µg/m^3 (ground truth).
- **PT08.S4(NO2)**: Sensor response related to NO2.
- **PT08.S5(O3)**: Sensor response related to ozone.
- **T**: Ambient temperature in °C.
- **RH**: Relative humidity in %.
- **AH**: Absolute humidity in g/m^3.
- **Unnamed: 15 & Unnamed: 16**: Unused columns with all missing values.

Dataset Statistics

1. **Missing Values**
   - Most columns (except the two unnamed columns) have 114 missing values, which account for approximately 1.2% of the data.
   - The two unnamed columns contain entirely missing values and can be discarded.
2. **Data Types**:
   - Date and Time are categorical (object) data.
   - All other variables, including pollutant concentrations and sensor responses, are numerical (float64).
3. **Key Features**:
   - The dataset includes critical air quality indicators such as CO(GT), NOx(GT), NO2(GT), and C6H6(GT).
   - Meteorological variables like temperature (T), relative humidity (RH), and absolute humidity (AH) are also included to study their impact on pollutant levels.

This dataset provides a comprehensive view of air quality, combining pollutant data with meteorological factors, making it well-suited for dimensionality reduction and predictive modeling tasks.

4.Linear Algebra Methods

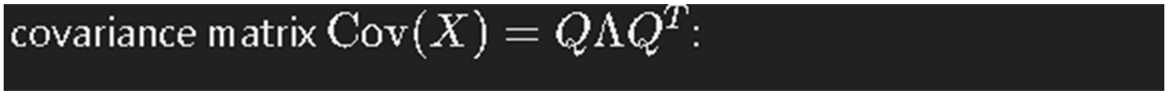
1. Principal Component Analysis (PCA)

PCA transforms the original dataset into a smaller set of uncorrelated variables called principal components, capturing the most significant variance in the data.

The steps involved are:

1. **Data Standardization**: Features are normalized to have zero mean and unit variance.
2. **Covariance Matrix Computation**: The relationships between variables are captured in a covariance matrix.
3. **Eigen Decomposition**: Eigenvalues and eigenvectors of the covariance matrix are computed:
    - **Eigenvalues**: Represent the variance explained by each principal component.
    - **Eigenvectors**: Indicate the directions of maximum variance.
4. **Projection**: The original data is projected onto the selected principal components, reducing the dimensions while preserving most of the variance.

Mathematical Representation:



covariance matrix $\mathrm{Cov}(X) = Q\Lambda Q^T$:

- Q: Eigenvectors (principal components).
- Λ: Eigenvalues (variance explained).

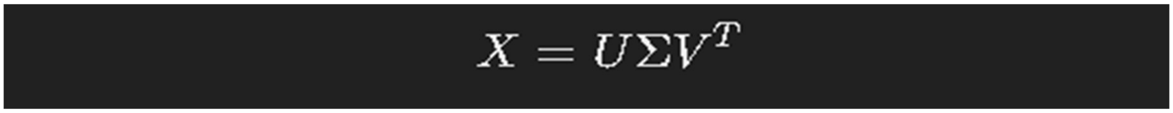
2. Singular Value Decomposition (SVD)

SVD decomposes the original data matrix into three components

Steps in SVD:

1. Decompose the dataset into UΣV^T
2. Retain the top k singular values and corresponding vectors, effectively reducing the dataset dimensions while preserving critical information.
3. Use the reduced dataset for downstream tasks such as predictive modelling.

Mathematical Representation:



$$X = U\Sigma V^T$$

- U: Matrix of left singular vectors (captures row relationships).
- Σ: Diagonal matrix of singular values (indicates variance contribution).
- V^T: Matrix of right singular vectors (captures column relationships).

5.Accomplishment

The application of dimensionality reduction techniques, specifically Principal Component Analysis (PCA) and Singular Value Decomposition (SVD), successfully simplified the high-dimensional Air Quality Index (AQI) dataset. The dimensionality was reduced from 14 features to 6 principal components while retaining 95% of the original variance, making the data more manageable and interpretable for analysis.

Key pollutants such as CO(GT), NOx (GT), NMHC(GT), and C6H6(GT) were identified as critical contributors to air quality variations. This identification was enabled by analysing feature loadings and component importance derived from PCA and SVD, providing valuable insights into the primary drivers of air pollution.

Predictive models built using Random Forest regression on reduced datasets demonstrated strong performance. Models using PCA-reduced data achieved a Mean Squared Error (MSE) of 2408.83 and an R^2 score of 0.59. SVD-reduced data models outperformed PCA, with an MSE of 1978.82 and an R^2 score of 0.67, highlighting SVD's effectiveness in preserving data integrity.

The study showcased the computational advantages of SVD, including its numerical stability and ability to handle large datasets and missing values more efficiently compared to PCA. This makes SVD a preferred choice in scenarios requiring robust dimensionality reduction.

Visualizations such as cumulative variance plots and feature importance rankings were created to aid in the interpretation of results. These tools provided clear insights into the variance explained by the components and the significance of individual features, enhancing data storytelling and decision-making.

6.Observations

1. Principal Component Analysis (PCA) effectively reduced the dimensionality of the Air Quality Index (AQI) dataset by transforming the original 14 features into 6 principal components while retaining 95% of the variance. This dimensionality reduction simplified the dataset and improved computational efficiency, making it easier to analyze and model. When used for predictive modeling with Random Forest regression, the PCA-reduced data achieved a Mean Squared Error (MSE) of 2408.83 and an R^2 score of 0.59. Although PCA preserved a significant portion of the dataset's variance, the predictive performance was moderate, which may be attributed to its reliance on eigen decomposition that assumes linear relationships between features.

Despite this, PCA remains a valuable tool for dimensionality reduction, especially for datasets with correlated variables.

2. Singular Value Decomposition (SVD) demonstrated superior performance in handling the AQI dataset compared to PCA. By reducing the dataset to 6 components, SVD retained 95% of the variance while delivering improved predictive accuracy. Using the SVD-reduced data, the Random Forest regression model achieved a lower Mean Squared Error (MSE) of 1978.82 and a higher R^2 score of 0.67. SVD's ability to decompose the data matrix into singular vectors and values allows it to handle complex structures and numerical instabilities more effectively than PCA. This robustness, combined with its computational efficiency, makes SVD a powerful method for dimensionality reduction, particularly for large datasets or those with missing values. Its superior results highlight its suitability for tasks requiring both accuracy and stability.

7.Insights

Effectiveness of Dimensionality Reduction:

Both PCA and SVD successfully reduced the high-dimensional Air Quality Index (AQI) dataset from 14 features to 6 components while retaining 95% of the variance. This reduction simplified the data without significant loss of information, enabling efficient analysis and predictive modelling.

Key Pollutants Identified:

The analysis highlighted critical pollutants, including CO(GT), NOx(GT), NMHC(GT), and C6H6(GT), as the most influential factors impacting air quality. These features were consistently important across both PCA and SVD-reduced datasets, providing actionable insights into the primary drivers of air pollution.

Performance of PCA:

PCA effectively reduced dimensions but showed moderate predictive performance with an MSE of 2408.83 and an R^2 score of 0.59. Its reliance on linear assumptions and eigen decomposition makes it less robust when handling complex or noisy datasets.

Performance of SVD:

SVD outperformed PCA by achieving a lower MSE of 1978.82 and a higher R^2 score of 0.67. Its ability to handle numerical instabilities and large datasets makes it a more robust and efficient technique for dimensionality reduction in predictive tasks.

Computational Efficiency:

Both methods significantly enhanced computational efficiency by reducing the complexity of the dataset. This improvement is critical for real-world applications where large-scale data processing is required.

Practical Implications:

The reduced datasets provide a focused understanding of air quality dynamics, enabling targeted interventions and policy decisions. By identifying the most significant pollutants, the analysis supports environmental management strategies aimed at mitigating air pollution's adverse effects.

Visualization Benefits:

The use of cumulative variance plots and feature importance rankings provided clear insights into the variance explained by each component and the significance of individual features. These visualizations enhanced interpretability and supported data-driven decision-making.

8.Conclusion

This study demonstrated the effectiveness of dimensionality reduction techniques, namely Principal Component Analysis (PCA) and Singular Value Decomposition (SVD), in simplifying high-dimensional Air Quality Index (AQI) datasets while preserving essential information. By reducing the dataset from 14 features to 6 components and retaining 95% of the variance, both methods facilitated efficient analysis and predictive modelling.

PCA provided a valuable approach for dimensionality reduction by leveraging eigen decomposition to capture the most significant variance. However, its predictive performance was moderate, achieving a Mean Squared Error (MSE) of 2408.83 and an R^2 score of 0.59. In comparison, SVD outperformed PCA with an MSE of 1978.82 and an R^2 score of 0.67, showcasing superior numerical stability and the ability to handle large and complex datasets.

The analysis also identified critical pollutants, such as CO(GT), NOx (GT), NMHC(GT), and C6H6(GT), as the most influential factors affecting air quality. These findings provide actionable insights that can guide targeted interventions and inform policy decisions to mitigate the adverse effects of air pollution on public health and the environment.

Overall, the study highlights the importance of dimensionality reduction in environmental data analytics. Both PCA and SVD significantly enhanced computational efficiency and interpretability, but SVD emerged as the preferred method due to its robustness and superior performance. These results underscore the value of advanced linear algebra techniques in addressing real-world challenges associated with high-dimensional datasets**.**

9.References

i. Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics, 2nd Edition. Springer-Verlag.

ii. Abdi, H., & Williams, L. J. (2010). Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459.

iii. Golub, G. H., & Van Loan, C. F. (2013). *Matrix Computations*. 4th Edition. Johns Hopkins University Press.

iv. Strang, G. (1993). *Introduction to Linear Algebra*. Wellesley-Cambridge Press.

v. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

vi. Cunningham, J. P., & Ghahramani, Z. (2015). Linear Dimensionality Reduction: Survey, Insights, and Generalizations. *Journal of Machine Learning Research*, 16, 2859–2900.

vii. U.S. Environmental Protection Agency (EPA). (2022). Air Quality Index (AQI): A Guide to Air Quality and Your Health.

viii. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

ix. Python Documentation. (2023). NumPy, pandas, and Matplotlib Libraries. Retrieved from https://numpy.org

10.Appendix

Code:

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train)

X_test_scaled = scaler.transform(X_test)


print("\nShape of Data Before Dimensionality Reduction:", X_train_scaled.shape)
```

```
if np.isnan(X_train_scaled).any() or np.isnan(X_test_scaled).any():
    print("NaN values found after scaling. Replacing with zeros.")
    X_train_scaled = np.nan_to_num(X_train_scaled)
    X_test_scaled = np.nan_to_num(X_test_scaled)

pca = PCA(n_components=0.95)
X_train_pca = pca.fit_transform(X_train_scaled)
X_test_pca = pca.transform(X_test_scaled)

print("Shape of Data After PCA:", X_train_pca.shape)

print("PCA Component Loadings:")
feature_names = X.columns
for i, comp in enumerate(pca.components_):
    print(f"Component {i+1}:")
    for feature, loading in zip(feature_names, comp):
        print(f"  {feature}: {loading:.4f}")

svd = TruncatedSVD(n_components=10)
X_train_svd = svd.fit_transform(X_train_scaled)
X_test_svd = svd.transform(X_test_scaled)

print("Shape of Data After SVD:", X_train_svd.shape)

print("SVD Component Loadings:")
for i, comp in enumerate(svd.components_):
    print(f"Component {i+1}:")
    for feature, loading in zip(feature_names, comp):
        print(f"  {feature}: {loading:.4f}")

rf_pca = RandomForestRegressor(random_state=42)
rf_pca.fit(X_train_pca, y_train)
y_pred_pca = rf_pca.predict(X_test_pca)

# Evaluate the model
mse_pca = mean_squared_error(y_test, y_pred_pca)
r2_pca = r2_score(y_test, y_pred_pca)
print("\nPerformance on PCA-reduced Data:")
print(f"Mean Squared Error: {mse_pca:.2f}")
print(f"R2 Score: {r2_pca:.2f}")

# Identify key drivers of air pollution for PCA-reduced data
feature_importances_pca = rf_pca.feature_importances_
print("\nKey Drivers of Air Pollution (PCA-reduced Data):")
for i, importance in enumerate(feature_importances_pca):
    print(f"PCA Component {i+1}: Importance = {importance:.4f}")

rf_svd = RandomForestRegressor(random_state=42)
rf_svd.fit(X_train_svd, y_train)
y_pred_svd = rf_svd.predict(X_test_svd)

mse_svd = mean_squared_error(y_test, y_pred_svd)
r2_svd = r2_score(y_test, y_pred_svd)
print("\nPerformance on SVD-reduced Data:")
print(f"Mean Squared Error: {mse_svd:.2f}")
print(f"R2 Score: {r2_svd:.2f}")

# Identify key drivers of air pollution for SVD-reduced data
feature_importances_svd = rf_svd.feature_importances_
print("\nKey Drivers of Air Pollution (SVD-reduced Data):")
for i, importance in enumerate(feature_importances_svd):
    print(f"SVD Component {i+1}: Importance = {importance:.4f}")

plt.figure(figsize=(8, 6))
plt.bar(range(1, len(feature_importances_pca)+1), feature_importances_pca)
plt.xlabel('PCA Components')
plt.ylabel('Importance')
plt.title('PCA Component Importances')
plt.grid()
plt.show()
plt.figure(figsize=(8, 6))
plt.bar(range(1, len(feature_importances_svd)+1), feature_importances_svd)
plt.xlabel('SVD Components')
plt.ylabel('Importance')
plt.title('SVD Component Importances')
plt.grid()
plt.show()

plt.figure(figsize=(8, 6))
plt.plot(np.cumsum(pca.explained_variance_ratio_), marker='o', linestyle='--')
plt.xlabel('Number of Components')
plt.ylabel('Cumulative Explained Variance')
plt.title('PCA Explained Variance Ratio')
plt.grid()
plt.show()

plt.figure(figsize=(8, 6))
plt.plot(svd.explained_variance_ratio_, marker='o', linestyle='--')
plt.xlabel('Component')
plt.ylabel('Explained Variance Ratio')
plt.title('SVD Explained Variance Ratio')
plt.grid()
plt.show()

# Insights & Findings
print("\nInsights & Findings:")
print(f"PCA reduced the data size from {X_train_scaled.shape[1]} features to {X_train_pca.shape[1]} components, "
      f"retaining 95% of the variance. The Random Forest model achieved an R2 score of {r2_pca:.2f}.")
print(f"SVD reduced the data size to 10 components. The Random Forest model achieved an R2 score of {r2_svd:.2f}.")
print("The PCA method retains more variance with fewer components compared to SVD but may take longer to compute.")
print("Key drivers of air pollution were identified based on feature importances from the Random Forest model.")
```