Implementation of Linear Regression, SVD, Gradient Descent and PCA on UCI Air Quality Dataset

Syeda Hifsa Kazmi (25K-7602)

Sanjinee (25K-7627)

Tanzeela Sehar (25K-7616)

Hafiza Sara Asghar (25K-7626)

1. Introduction

Air quality monitoring is essential for public health and urban planning. This project uses the UCI Air Quality dataset containing 9357 hourly sensor readings collected from March 2004 to February 2005 in an Italian city. The device includes five metal-oxide chemical sensors, along with environmental measurements (temperature, humidity, and absolute humidity). Ground-truth values of several pollutants are provided by certified analyzers.

The goal of this study is to predict CO concentration (CO(GT)) using:

- Ordinary Least Squares (OLS)
- Singular Value Decomposition–based regression (SVD)
- Gradient-based optimization methods (Batch GD, SGD, Adam)
- PCA

We analyze data quality, correlations, conditioning of the design matrix, and numerical stability under ill-conditioned scenarios.

2. Dataset Overview

- Observations: 9357 hourly samples
- Features: 15 (sensor signals + environmental variables + pollutant ground truths)
- Target variable: CO(GT) (mg/m^3)
- Missing data format: Encoded as -200

3. Missing Data Handling

Steps taken:

- Replace all -200 values with NaN.
- Variable NMHC(GT) contains >90% missing values. Therefore, it was dropped.
- Partially missing pollutant readings NOx(GT) and NO2(GT) → linear interpolation.
- Remove rows missing the target variable.
- Remove 366 rows with simultaneous missing sensor readings (indicating sensor failure).

After cleaning, the dataset contains ≈8900 complete samples.

4. Feature Selection and Scaling

- Selected predictors: PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O3), T, RH, AH
- All features are standardized using StandardScaler. A bias column is added for regression.

5. Exploratory Analysis

- Strong positive correlation between CO(GT) and sensors S1, S2, S4, S5.
- Strong negative correlation with S3(NOx).
- Weak correlations with environmental factors (T, RH, AH).
- Sensor features explain the majority of CO variation.

6. Methodologies Used

6.1 Ordinary Least Squares (Closed Form)

Uses the Normal Equation: $\beta = (X^T X)^{-1} X^T y$

Returns optimal coefficients but may be unstable when $X^T X$ is ill-conditioned.

6.2 SVD-Based Regression

Uses matrix factorization: $X = USV^T \quad \Rightarrow \quad \beta = VS^{-1}U^T y$

Advantages: Numerically stable, Handles collinearity well, Avoids explicit matrix inversion

6.3 Gradient-Based Optimization Methods

- **Batch Gradient Descent (BGD)**: Full-dataset gradient update each iteration. Converges smoothly but slow for large datasets.
- **Stochastic Gradient Descent (SGD):** Updates using one random sample per step. Fast but noisy convergence.
- **Adam Optimizer:** Adaptive method combining momentum + RMSProp. Fast, robust convergence.

6.4 Principle Component Analysis Using SVD

- It is used for dimensionality reduction. We reduced the features to two Principle Components and one bias feature.
- Mean-center the data and compute the SVD of the mean-centered matrix using: $X = U\Sigma V^T$
- Columns of V are the **principal directions** (eigenvectors of the covariance matrix).
- Variance of $PC_i = \frac{\sigma_i^2}{n-1}$, where σ_i is the i-th singular value
- PCA transformed data: $Z = XV$

5. Results

5.1 OLS vs SVD Performance

Both methods produced identical coefficients and predictions on clean data.

Metric	OLS	SVD
RMSE	~0.49	~0.49
R^2	~0.83	~0.83

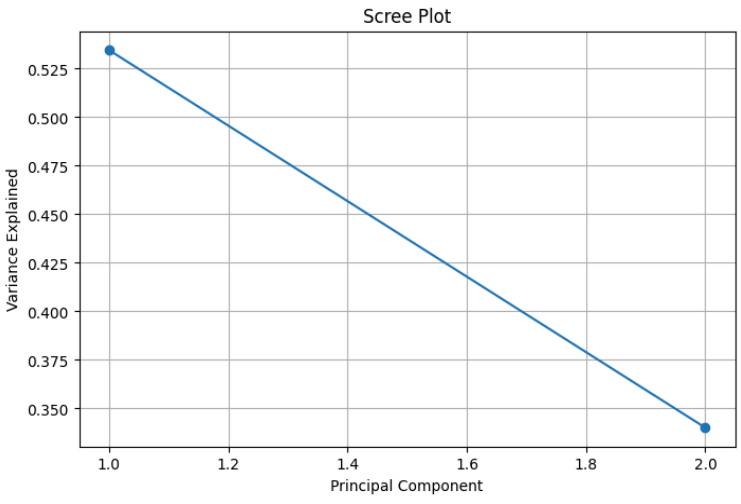
This shows the linear model fits CO(GT) well.

5.2 Gradient Descent Experiment's Convergence Comparison

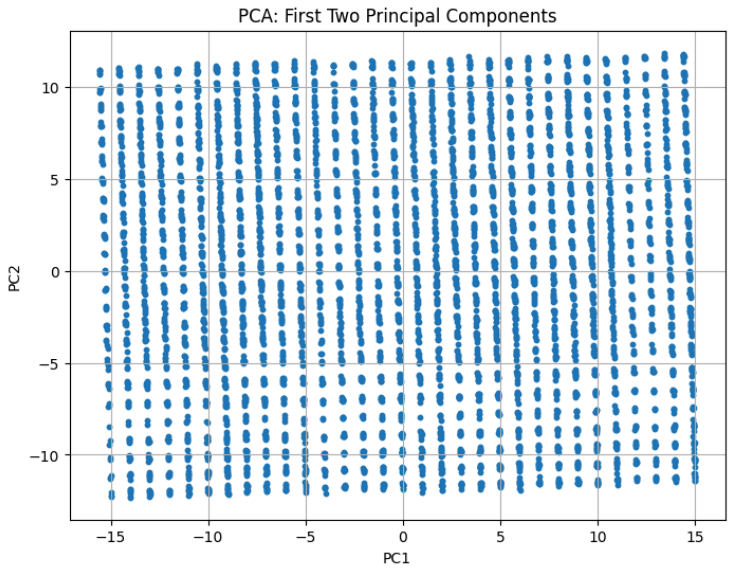
Method	Final MSE
OLS (analytical)	lowest
Batch GD	close to OLS
SGD	higher variance
Adam	near-OLS accuracy

Adam achieved the best trade-off among iterative methods. **Tested Learning Rates** {0.2, 0.1, 0.01, 0.001, 0.0001} yielded {Divergence, Smooth convergence, Very slow convergence} in order which confirms the classical learning rate behavior.

5.3 PCA Plottings



Scree Plot between Principle Components against the Explained Variance



2D scatter plot for the two components

5.5 General Comparison Between Different Methodologies

Method	Training MSE	Test MSE
OLS (Analytical)	0.2327	0.2327
OLS (SVD)	0.2327	0.2327
SVD + Denoising	0.2327	0.2327
Batch Gradient Descent (BGD)	0.2448	-
ADAM Optimizer	0.2327	-
Stochastic Gradient Descent (SGD)	0.2334	-
Principle Component Analysis (PCA)	1.8098	1.575

6. Key findings:

1. Sensor data strongly predicts CO concentration; environmental factors add minimal value.
2. OLS and SVD perform equally on well-conditioned data, but SVD is preferred when features are correlated.
3. Ill-conditioning causes OLS instability, confirming theoretical expectations.Adam consistently outperforms other iterative optimizers in convergence speed and final loss.
4. Normal Equation gives the best accuracy in this small dataset, but does not scale well to high-dimensional data.
5. In comparison with other methods' results, applying PCA and reducing the features to only two principle components lead to an increase in the MSE for both training and testing. However, test MSE is still comparatively less than the train MSE.

7. Conclusion

This study demonstrates a complete workflow for air quality prediction using linear regression and optimization techniques. The results show:

- Linear models are effective for predicting CO(GT).
- SVD-based regression is the most numerically reliable method.
- Adam optimization provides a strong alternative when closed-form solutions are impractical.
- PCA is useful when the number of features is huge and is resource intensive.