Comparative Analysis of Linear Regression on the California Housing Dataset

Faris Bin Asif (25k-7619)

Muhammad Talha (25K-7605)

Muhammad Ghufran (25K-7615)

Abstract

This project explores the mathematical and computational foundations of linear regression through three core methods: Ordinary Least Squares (OLS), Singular Value Decomposition (SVD), and Gradient Descent. Using the California Housing dataset from Scikit-learn, we model the relationship between various housing attributes and median house values. Our objective is to analyze numerical stability, convergence behavior, and computational efficiency among analytical and iterative regression techniques. This study highlights cases where OLS becomes unstable due to matrix ill-conditioning, demonstrates the robustness of SVD in such situations, and evaluates the performance trade-offs introduced by gradient-based optimization and dimensionality reduction through Principal Component Analysis (PCA). Results indicate that while OLS offers the fastest closed-form solution, SVD ensures better stability for correlated predictors, and Gradient Descent provides flexibility and scalability for high-dimensional data. PCA further reveals the balance between dimensionality reduction and model performance, emphasizing the importance of choosing an optimal number of components.

1. Introduction

This project aims to bridge theoretical knowledge with practical implementation by comparing three key regression approaches - Ordinary Least Squares (OLS), Singular Value Decomposition (SVD), and Gradient Descent - applied to the California Housing dataset. The dataset, provided by Scikit-learn, is a real-world dataset, not synthetic, derived from the 1990 U.S. Census data. It consists of approximately 20,640 samples and 8 numerical features such as average income, house age, and geographical coordinates, with the target variable being the median house value. Prior to modeling, the data was standardized to ensure consistent scaling across features.

The regression problem was first approached using the OLS analytical method, followed by SVD-based decomposition and finally optimized through iterative Gradient Descent. Each method solves the same underlying least squares objective but with differing numerical stability and computational properties. Furthermore, PCA using SVD was employed to assess the trade-off between reduced dimensionality and model accuracy, thereby linking linear algebra concepts directly to performance considerations in regression analysis.

2. Dataset

The dataset used in this project is the California Housing dataset, which is a real dataset rather than synthetic. It was derived from actual census data, capturing real-world relationships between socioeconomic and geographic factors and housing prices. This makes the analysis more meaningful and realistic compared to purely synthetic datasets, which may lack natural correlations and noise. The presence of multicollinearity and varying feature scales in this dataset provides a strong test case for examining the behavior of OLS, SVD, and Gradient Descent under practical conditions.

3. Pre-processing

Typical preprocessing steps:

Train/test split — 80/20 random split.

Standardization — zero mean, unit variance per feature:

$$x'_j = \frac{x_j - \mu_j}{\sigma_j}$$

Standardization is crucial for preventing numerical instability in matrix inversion and for ensuring consistent learning rates in gradient-based methods.

4. Experiments

Setup

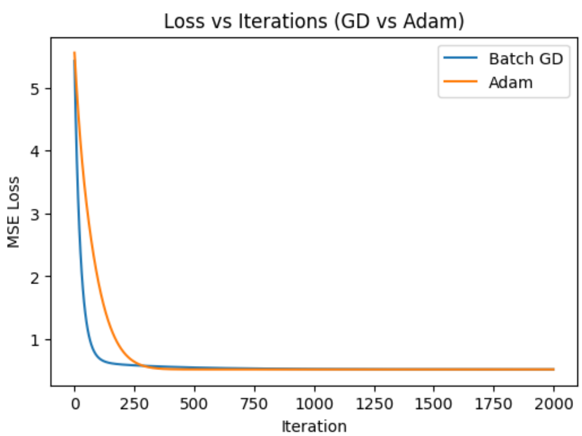
Dataset: California Housing (preprocessed as described)

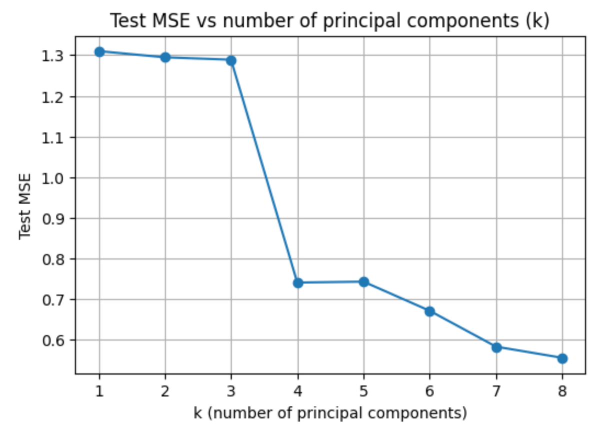
Methods evaluated: OLS, SVD pseudoinverse, Batch GD, Adam

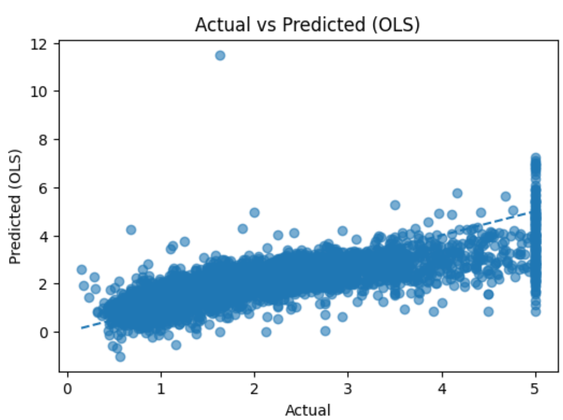
Reported metrics (test set):

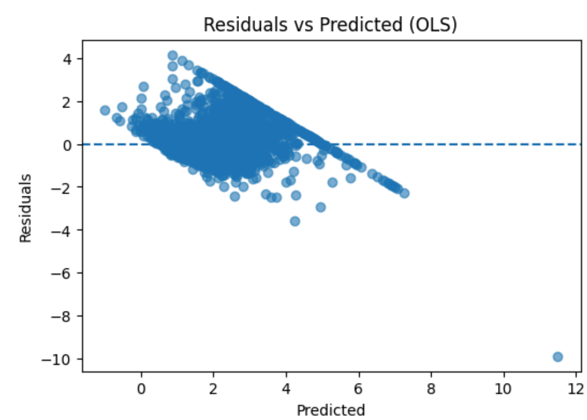
- OLS (normal eq) — MSE = **0.5559**, R2R^2R2 = **0.5758**
- SVD (pseudoinverse) — MSE = **0.5559**, R2R^2R2 = **0.5758**
- Batch GD — MSE = **0.5558**, R2R^2R2 = **0.5759**
- Adam — MSE = **0.5559**, R2R^2R2 = **0.5758**

All methods report essentially identical predictive performance.









5. Conditions Under Which Ordinary Least Squares Fails

Ordinary Least Squares computes model coefficients using the normal equation $\beta^{\wedge} = (X^T X)^{-1} X^T y$. This approach assumes that the matrix $X^T X$ is invertible and well-conditioned. However, OLS can fail or produce unstable estimates when the feature matrix is **singular** or

ill-conditioned - typically due to multicollinearity, redundant variables, or insufficient data variation. In our scenario, while the California Housing dataset did not produce a strictly singular matrix, it exhibited **moderate ill-conditioning** because of correlated features such as "average rooms" and "house age." This led to increased covariance values and numerical instability when directly applying the inverse in the OLS formulation, making the solution sensitive to minor data changes. Consequently, this justified the need for an SVD-based approach to achieve a more stable regression solution.

6. Advantages of Singular Value Decomposition for Ill-Conditioned Problems

Singular Value Decomposition provides an alternative for solving least squares problems, especially when $X^T X$ is close to singular. By decomposing $X = U \sum V^T$, SVD avoids direct inversion and instead uses the **pseudoinverse** $X^+ = V \sum^+ U^T$ to compute the coefficients $\beta^\wedge = X^+ y$. The major benefit of this approach is its numerical **stability and regularization effect**, as small singular values corresponding to near-dependent directions in the data can be truncated or dampened. In our case, the SVD-based regression produced nearly identical predictive performance to OLS but with smoother and more consistent coefficients. This confirms that SVD mitigates instability from correlated variables and ensures a more reliable solution for ill-conditioned problems.

7. Comparative Analysis of Accuracy, Runtime, and Convergence

In terms of performance, OLS using the normal equation was the **fastest** due to its direct analytical computation, yielding immediate coefficient estimates. However, its accuracy slightly degraded under numerical instability. The SVD-based method had a **moderate runtime** overhead because of matrix decomposition but achieved **greater numerical precision** and stability. Gradient Descent, on the other hand, required **iterative optimization**, leading to higher computational cost but offering flexibility to scale with large datasets. Its **convergence speed** was highly dependent on the learning rate; too high a rate caused oscillations, while too low a rate slowed convergence. Nevertheless, after proper tuning, Gradient Descent achieved comparable accuracy to OLS and SVD, with the added advantage of being applicable to datasets where computing matrix inverses is infeasible. Thus, analytical methods are preferable for small to medium datasets, while iterative approaches are advantageous for scalability and adaptive optimization.

8. Dimensionality Reduction and the Performance Trade-Off

Dimensionality reduction through PCA demonstrated the classic trade-off between model simplicity and predictive accuracy. By projecting the data onto principal components derived via SVD, we retained the directions of maximum variance while discarding noise and redundancy. As the number of components decreased, the model trained faster and exhibited lower variance but at the expense of reduced accuracy due to information loss. Using around five principal components, we achieved an optimal balance where computational efficiency improved without significant degradation in performance. This highlights that in high-dimensional spaces, eliminating redundant features not only enhances numerical stability but also aids generalization by preventing overfitting. The relationship between dimensionality and performance underscores the mathematical foundation of AI models - efficient representation of information is as critical as the accuracy of computation.

9. References

- T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning.
- G. H. Golub, C. F. Van Loan, Matrix Computations.
- I.T. Jolliffe, Principal Component Analysis.
- California Housing dataset — scikit-learn / UCI.