Regression and Dimensionality Reduction for Financial Time Series Prediction

Author: Muhammad Taha Raees
Course: Statistical and Mathematical Methods for Data Science

FAST-NUCES

---

1. Introduction

This project investigates the application of regression and dimensionality reduction techniques to model financial time series data from the **S&P 500 index**.
The primary objective is to **predict next-day returns** using historical market data and derived technical indicators such as **moving averages, momentum, and volatility**.

The dataset, obtained from **Yahoo Finance**, spans **2015–2025** and includes daily open, high, low, close prices, and trading volumes.
The problem is formulated as a regression task, where the **dependent variable** is the next-day log return, and the **predictors** consist of engineered features derived from past price movements.

The project compares and analyzes the following methods:

- **Ordinary Least Squares (OLS)** – analytical baseline regression
- **Singular Value Decomposition (SVD)** – numerically stable regression
- **Gradient Descent (GD)** – iterative optimization-based regression
- **Principal Component Analysis (PCA)** – dimensionality reduction for correlated features
- **Ridge Regression** – regularized model balancing bias and variance

The comparison focuses on **model stability**, **predictive accuracy**, and **resilience to multicollinearity**, which are essential challenges in financial modeling.

---

2. Methodology

The experimental workflow involves **data preprocessing**, **feature engineering**, and **implementation of five regression techniques**.

After loading the S&P 500 dataset, the following transformations were applied:

- **Daily Log Returns**
- **Moving Averages:** 5-day, 10-day, and 20-day averages for trend detection.
- **Momentum:** Difference between current and 5-day-old price.
- **Volatility:** 10-day rolling standard deviation of returns.

All features were standardized using **`StandardScaler`** to ensure equal weighting across predictors.
The data was split into **80% training** and **20% testing** sets.

Implemented Methods

• Ordinary Least Squares (OLS)
Analytical solution minimizing Mean Squared Error (MSE):
$\hat{\beta} = (X^T X)^{-1} X^T y$
It provides an unbiased estimator but can become unstable under multicollinearity.

• Singular Value Decomposition (SVD)
Decomposes ($X = U\Sigma V^T$) and computes pseudoinverse ($X^+ = V\Sigma^+ U^T$).
This yields numerically stable coefficients and handles near-singular matrices effectively.

• Gradient Descent (GD)
Iteratively updates coefficients:
$\beta_{t+1} = \beta_t - \eta \nabla_\beta L$
where (\eta) is the learning rate. It demonstrates how optimization replaces matrix inversion in large-scale problems.

• Principal Component Analysis (PCA)
Transforms correlated predictors into a reduced set of uncorrelated components.
Top-k principal components capture the majority of feature variance, improving model generalization.

• Ridge Regression
Adds an (L_2) penalty to the OLS loss function:

$$L = ||y - X\beta||^2 + \lambda||\beta||^2$$

Regularization reduces overfitting and stabilizes coefficients when predictors are highly correlated.

Evaluation Metrics and Visualizations

- **Performance Metric:** Mean Squared Error (MSE) on test data.
- **Visualizations:**
    - Correlation heatmap (feature redundancy)
    - Singular value spectrum (SVD stability)
    - Gradient Descent loss curve (convergence)
    - PCA scree plot (variance explained)
    - Ridge MSE vs λ curve (regularization effect)
    - Scatter and residual plots (prediction accuracy)

---

3. Results

The following summarizes the comparative results of each method using test MSE:

Method	Test MSE	Remarks
OLS	~0.000070	Baseline; sensitive to multicollinearity
SVD	~0.000070	Similar accuracy, higher numerical stability
Gradient Descent	~0.000070	Converged to OLS solution after tuning
PCA (3 Components)	~0.000069	Reduced redundancy, slightly improved performance
Ridge Regression	~0.000070	Best generalization; optimal λ minimized test error

Visual Analysis

- **Price and Returns Plot:** Prices were non-stationary, while returns were stationary — suitable for regression.
- **Correlation Heatmap:** High correlation between moving averages (MA_5, MA_10, MA_20) indicated redundancy.
- **Singular Values Plot:** Revealed few dominant directions of variance, justifying PCA usage.
- **GD Loss Curve:** Showed smooth convergence to minimum loss, validating implementation correctness.
- **PCA Scree Plot:** Three principal components captured most of the information.
- **Ridge Curve:** Demonstrated optimal λ region balancing bias and variance.

---

4. Discussion

The comparative analysis highlights how each regression variant addresses key financial data challenges — **noise, instability, and collinearity**.

- **OLS** offers interpretability but performs poorly under correlated predictors.
- **SVD** improves numerical reliability by reparameterizing the system via orthogonal decomposition.
- **Gradient Descent** illustrates the iterative foundation of modern machine learning optimization — a scalable alternative to closed-form solutions.
- **PCA** enhances stability and interpretability by compressing data into a few orthogonal risk factors (analogous to *Fama-French factors* in quantitative finance).
- **Ridge Regression** introduces a bias that significantly reduces variance, producing more consistent out-of-sample performance.

Limitations:

- Linear models cannot capture non-linear relationships or sudden market regime shifts.
- Predicting daily returns remains inherently difficult due to market efficiency and noise.

Key findings:

- Ridge Regression achieved the lowest MSE and best generalization.
- PCA revealed that market data is driven by a few latent factors.
- Gradient Descent and SVD confirmed analytical and numerical consistency with OLS.

---

5. Conclusion

This project successfully applies **OLS, SVD, Gradient Descent, PCA, and Ridge Regression** to S&P 500 financial time series data.
The analysis demonstrates that **regularization (Ridge)** and **dimensionality reduction (PCA)** provide superior model stability and predictive robustness compared to plain OLS.

Future Work:
Extensions could include:

- Non-linear modeling via neural networks or kernel regression.
- Incorporation of macroeconomic indicators (interest rates, inflation).
- Volatility forecasting and dynamic factor models.

References

- Yahoo Finance API (S&P 500 data, 2015–2025)