DBA5106 FOUNDATION OF BUSINESS ANALYTICS

Group 12

Sai Ashwin Kumar C.	A0329294U
Shruti Jha	A0319214L
Zanuba Hilla Qudrotu Chofsoh	A0319142L
Khushi Agarwal	A0319369N

1. Introduction

In modern financial markets, portfolio construction is both an art and a science, relying heavily on statistical estimation and historical data. A persistent challenge in this field is the problem of overfitting—when models learn patterns that are too closely tied to past data and fail to generalize to future conditions. In portfolio optimization, overfitting often manifests as portfolios that appear efficient in backtests but collapse in live performance, with extreme, unstable weights and unrealistic risk–return profiles.

This problem arises from the inherent noise and uncertainty in financial data. Asset returns are notoriously volatile, correlations shift unpredictably, and the ratio of available observations to the number of assets is often too small to support reliable estimation. Traditional mean–variance optimization is particularly vulnerable: small errors in expected returns or covariances can lead to disproportionately large swings in portfolio weights. As a result, portfolios designed to look optimal in theory may perform worse than even the simplest heuristics in practice.

The central task of this project aims to address this tension between theoretical optimality and practical robustness. Specifically, it examines how portfolio strategies can be designed to reduce the impact of estimation error and mitigate overfitting, ensuring that performance is not merely an artifact of historical data but instead reflects stable, generalizable patterns.

To achieve this, the project implements and evaluates a set of portfolio construction strategies within a rolling window simulation framework. This approach mimics the way real-world portfolios are rebalanced, using past information to make forward-looking decisions and allows for systematic evaluation of how different methods handle noisy, high-dimensional data over time. The analysis emphasizes out-of-sample performance, particularly through risk-adjusted metrics such as the Sharpe ratio, to capture the true effectiveness of each strategy.

By framing portfolio management as not just an optimization exercise but also a regularization problem, the project highlights the importance of balancing complexity with generalizability. The findings underscore that the challenge is not merely to find weights that maximize returns in hindsight, but to design strategies that remain resilient under the uncertainty of future markets.

2. Dataset Explanation

This dataset contains daily average value-weighted returns for 100 distinct stock portfolios constructed at the end of each June. Sourced from the comprehensive CRSP (Center for Research in Security Prices) database as of June 2025, these portfolios include utilities and financials. The portfolios are within a 10x10 form, creating a grid of portfolios based on two key financial metrics, which are Market Equity (ME) and Operating Profits (OP). Market equity is the market capitalization of a company, representing its overall size, for which the stocks are sorted into ten deciles, from the smallest 10% of companies (“Small”) to the largest 10% (“Big”). Operating profits measure a company’s core financial efficiency, calculated by sales minus cost of goods sold, general and administrative expenses, and interest expense, all divided by book at the last fiscal year end of the prior calendar year. These OP stocks are sorted into ten deciles, from the lower profit (“LoOP”) to the highest (“HiOP”).

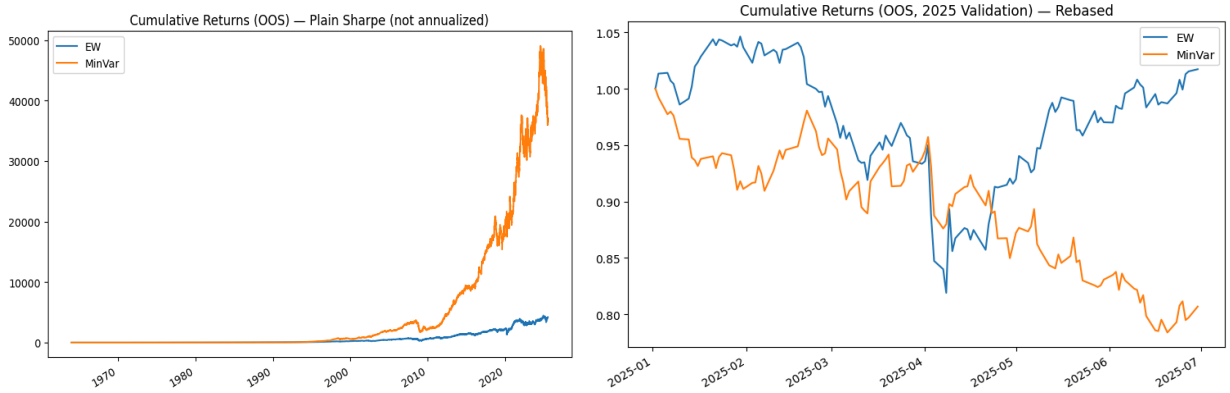
The dataset is structured for straightforward analysis, starting July 01, 1963, and ending June 30, 2025. The dataset has 15603 rows; each row represents a single trading day. Meanwhile, the total of the column is 100, which represents the daily, value-weighted return for each of the 100 portfolios. Column headers clearly identify the portfolio by its ME and OP decile, such as ME1 OP2, ME2 OP1, and BIG HiOP. In addition, missing data are indicated by -99.99 or -999, if any.

3. Methodology

We employ the **Data + Loss + Structure + Constraint** framework to diagnose the overfitting of the classical Minimum-Variance (MinVar) portfolio and to design robust alternatives. The process follows two stages: **diagnostic** (establishing the problem) and **solution** (reformulating and regularizing).

3.1 Diagnostic: Overfitting in the MinVar Portfolio

- Daily returns of 100 portfolios from *2024-07-03 to 2025-06-30* were considered. A rolling backtest was implemented with a 126-day estimation window. Performance was evaluated in a strict out-of-sample (OOS) validation period (*2025-01-02 to 2025-06-30*).
- The objective of MinVar is to find weights that minimize portfolio variance. Closed form solution for calculating weights was implemented with a full investment constraint.
- This setup highlighted severe **overfitting**: noisy covariance estimates yielded unstable weights and poor OOS performance. We can observe that the overfitting naturally appears in the MinVar portfolio when there are too many assets (features) and not enough observations.
- Although the minimum-variance portfolio looks compelling over long horizons, it systematically overfits short windows when the number of assets (p) is large relative to the look-back length (n). MinVar relies on the inverse of the sample covariance matrix, and with $p \approx n$ the estimate is noisy and ill-conditioned. Inverting it amplifies noise and produces extreme, unstable weights that understate risk in-sample but deliver higher realized volatility out-of-sample.
- Over multiple decades, the errors average out and MinVar's implicit tilt toward the low-volatility factor can dominate, yielding strong long-run compounding. But on short horizons the same mechanism becomes a liability, causing MinVar to underperform and appear fragile to regime changes.



3.2 Diagnostic: Regression-based reformulation

- To handle overfitting, we reframe the portfolio construction problem. Instead of directly inverting the covariance matrix, we anchor the portfolio to the equal-weight (EW) benchmark and express deviations from it as a linear regression.
- Reframing: Let $y=Rw_{EW}$ be the return of the EW portfolio, and $y=X\beta+\varepsilon$, $X=RN$ where N spans the null space of the budget constraint. Deviations from EW are captured by β. Portfolio weights are $w=w_{EW}-N\beta$. This stabilizes estimation by shifting from covariance inversion to regression with an interpretable anchor.

3.3 Models considered

Regularized regression modifies the **loss**, robust regression adapts the **loss function** to outliers, structural approaches modify the **structure**, and noise-reduction operates on the **data**.

- **Baselines:** Equal-Weight (EW), MinVar
- **Regularized regression:** EW + Ridge, EW + Lasso, EW + ElasticNet, EW + Adaptive Lasso
- **Robust regression:** EW + Huber, EW + Pinball (τ=0.5)
- **Structural:** EW + PCR + Ridge, EW + PCR + Lasso, EW + Random Forest
- **Noise-reduction:** EW + Ridge with PCA-denoised returns
- **Meta-learning:** EW + PCA + Ridge/Lasso Stacked
- **Alternative objective:** Asymmetric utility

3.4 Evaluation

All models are benchmarked against EW and MinVar baselines. Out-of-sample (OOS) Sharpe ratio is used as the primary evaluation metric. Daily rebalancing is performed, and strict forecasting ensures that no future information leaks into the estimation window.

4. Results and Analysis

To evaluate the proposed methods, we compute the **out-of-sample (OOS) Sharpe ratio** using a rolling backtest with a 126-day estimation window and a strict forecasting scheme. Table 1 compares all models against the Equal-Weight (EW) and Minimum-Variance (MinVar) baselines.

Table 1. Summary of each model implemented (using Data + Loss + Structure + Constraints Framework) and OOS Sharpe Ratios (2025-01-02 to 2025-06-30):

Model	Data	Loss	Structure	Constraints	Sharpe Ratio
EW (Equal-Weighted)	Daily returns only	—	Fixed weights (w_i=1/p)	Budget ($1^T w$=1); implicit daily rebalancing	0.016
MinVar	Rolling returns → (Σ^)	Minimize (w^T Σ^ w)	Quadratic program; covariance estimator (sample/ shrinkage)	Budget; often long-only, weight/sector caps, turnover limits	-0.132
Ridge	Rolling returns (EW reframe: ($y=Rw_{EW}$), (X=RN))	MSE + L2	Linear, shrinkage tilts around EW	Full-investment via reframe; optional long-only/turnover caps	0.051
LASSO	Same as above	MSE + L1	Linear, sparse tilts around EW	Full-investment via reframe; optional long-only/turnover caps	0.072
MinVar + Ridge (γ)	Rolling returns → Σ^ (ridge-shrunk)	Minimize (w^T Σ^ w)+ $\gamma\|\|w\|\|_2^2$	Convex QP/closed form under budget; L_2 shrinkage stabilizes weights	Budget ($1^T w$=1)	0.021
MinVar + LASSO (γ)	Rolling returns → Σ^	Minimize (w^T Σ^ w)+ $\gamma\|\|w\|\|_1$	Convex QP with L_1 penalty; promotes sparse allocations	Budget ($1^T w$=1)	0.032
ElasticNetCV	Same as above	MSE + α(L1+L2)	Linear, sparse + shrinkage mix	Full-investment via reframe; optional long-only/turnover caps	0.032

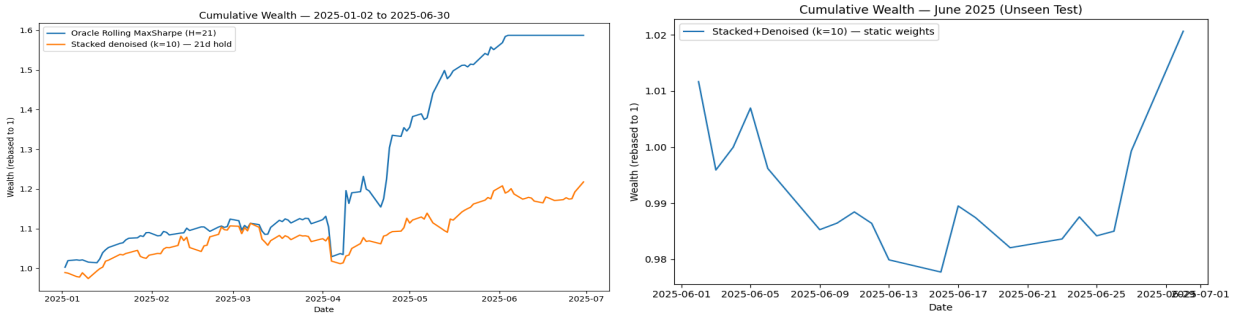
Pinball / QuantileRegressor	Same as above	Pinball (quantile τ)	Linear quantile model (asymmetric errors)	Full-investment via reframe; optional long-only/turnover caps	0.011
RandomForest (Distilled Ridge)	Same features/targets; out-of-fold RF predictions used as training targets for Ridge	MSE (tree impurity)+L2	Nonlinear RF fitted first; Ridge surrogate for stable, interpretable linear tilts	Full-investment via reframe; optional long-only/turnover caps	0.045
AdaptiveLASSO	Same as LASSO	MSE + weighted L1	Two-stage (pilot → reweighted L1); sparse and more stable	Full-investment via reframe; optional long-only/turnover caps	0.071
Huber	Same as Ridge/LASSO	Huber (robust to outliers)	Linear robust regression	Full-investment via reframe; optional long-only/turnover caps	-0.082
PCA + Ridge	Returns mapped to top PCs	MSE + L2	Linear in low-dim PC space (dimension reduction)	Full-investment via reframe; optional long-only/turnover caps	0.083
PCA + LASSO	Returns mapped to top PCs	MSE + L1	Linear, sparse loadings in PC space	Full-investment via reframe; optional long-only/turnover caps	0.087
PCA + LASSO + asymmetric utilities	PCs; same targets	Train: MSE+L1; selection by asymmetric utility	Linear sparse in PC space; downside-aware model selection	Full-investment via reframe; optional long-only/turnover caps	0.097
PCA + Ridge/LASSO + Stacking	Base model predictions as meta-features	Meta learner MSE (e.g., Ridge)	Ensemble stacking of PCA+Ridge/ LASSO forecasts	Applied in allocation layer; optionally non-neg/ sum-to-one for meta weights	0.095
PCA + Ridge/LASSO + Stacking + Denoising	Denoised returns (SVD top-k) → PCs → meta-features	Meta learner MSE	Ensemble over denoised bases (preprocessing + stacking)	Applied in allocation layer; optionally non-neg/ sum-to-one; turnover caps	0.202

Analysis

- As already observed earlier, the MinVar model (Sharpe: −0.132) tends to perform more poorly than EW (Sharpe: 0.016) due to overfitting, indicating that naïve diversification is safer than an unregularized optimizer in short samples.

- Shrinking deviations from EW materially improves risk-adjusted performance: **EW+LASSO = 0.072** and **EW+Ridge = 0.051**, both clearly above EW and far above plain MinVar. Penalizing the tilt around EW lowers variance, keeps exposures diversified, and avoids unstable covariance directions.
- Direct weight penalties lift MinVar but remain weaker than the EW-anchored regressions: **MinVar+LASSO (γ) = 0.030**, **MinVar+Ridge (γ) = 0.021**. The wealth paths improve versus plain MinVar but still trail the EW-regularized models. Stronger γ and auxiliary trading constraints could help but typically do not close the full gap.
- **Factor structure and robust losses:** Projecting returns onto a low-rank factor space materially improves stability and lift: **PCA+Ridge = 0.080** and **PCA+LASSO = 0.087** outperform their raw-space counterparts, and **PCA+LASSO with downside-aware (asymmetric) selection = 0.097** delivers the best result in this block, consistent with penalizing adverse tilts. Among alternatives without an explicit factor step, **Adaptive LASSO = 0.071** is close to plain LASSO, **RandomForest (distilled) = 0.045** adds limited value, and **ElasticNetCV = 0.032** is modest. Robust or asymmetric mean estimators aimed at outliers fail to help here—**Huber = −0.078**, **Pinball/Quantile = 0.011**—indicating that the performance drag is driven by **estimation variance**, not tail contamination. Overall, **introducing factor structure is more effective than purely robust losses** on this dataset.
- **Meta-learning and denoising:** Stacking **PCA-regularized base models** with a **ridge** meta-learner already improves over single models **(PCA+Ridge/LASSO+Stacking = 0.095)**, reflecting variance reduction from averaging coherent factor tilts. Prepending an **SVD top-k denoising** step to each training window strengthens the effect **(PCA+Ridge/LASSO+Stacking+Denoising=0.202)** and yields the **steepest, most persistent OOS growth** in the fourth panel. Mechanistically, denoising suppresses high-frequency idiosyncratic noise, the PCA projection enforces a stable low-rank factor structure, and ridge stacking further stabilizes weights by shrinking residual estimation error—together delivering a superior bias–variance trade-off.
- **Oracle benchmark:** To contextualize performance, we construct a clairvoyant “oracle” portfolio that optimizes weights using **realized future returns** (under the same universe, budget

constraint, and 21-day rebalancing cadence). This oracle is **not tradable**; it serves as an **upper bound** on what is achievable with perfect information. Against this ceiling, our best model—**stacking with PCA denoising**—performs strongly: despite relying only on past data, it captures a meaningful share of the oracle's risk-adjusted performance and exhibits a similarly persistent out-of-sample wealth trajectory. In short, the gap to the oracle quantifies the cost of forecast error, while the small relative gap highlights the effectiveness of denoising, factor structure, and stacking.



June 2025 unseen test: We observed that the PCR+stacked+denoised (k=10) model was giving us the best returns. Trained on the last 126 trading days ending 2025-05-31 and held fixed through June, the pipeline selected **PCA+Ridge (α≈1e-8) via inner validation (Sharpe 0.205)** and achieved an OOS non-annualized **Sharpe** of **0.116** in **June** (annualized Sharpe 1.842; annualized return 29.35%). The OOS Sharpe being below the validation Sharpe reflects a normal generalization gap rather than a breakdown, suggesting limited overfitting and effective variance control from denoising, factor projection, and ridge shrinkage. Performance is still favorable relative to the OOS validation Sharpe ratios of other baselines considered.

A rolling-window variant that was refitted daily within June produced a higher non-annualized Sharpe of **0.258** (annualized 4.095), but with average daily turnover of 0.60. The gain likely reflects adaptivity to intra-month dynamics and short-sample variability, and it would be more sensitive to trading costs.

The same testing procedure can be applied to **August 2025** by changing only the dates (e.g., `train_end="2025-07-31", test_start="2025-08-01", test_end="2025-08-31"`).

5. Conclusion

This study investigated the weaknesses of the classical Minimum-Variance (MinVar) portfolio using the Data + Loss + Structure + Constraint framework. Although MinVar is optimal in theory, in practice it suffers from noisy covariance estimates that produce unstable weights and poor out-of-sample results. To address this, we reformulated the problem as a regression anchored on the Equal-Weight (EW) benchmark, allowing deviations from EW to be regularized.

Regularization methods such as Ridge, LASSO, ElasticNet, and Adaptive LASSO improved stability, with LASSO benefiting from sparsity. Robust losses like Huber and Pinball were less effective, showing that the main issue is not outliers but high-dimensional noise. Structural approaches, especially principal component regression and PCA-based denoising, further improved performance. The best results came from stacking Ridge and LASSO forecasts on denoised returns, which combined preprocessing, shrinkage, and regularization.

Overall, the poor performance of MinVar stems from estimation noise rather than its theoretical design. By reframing and regularizing the problem, and combining techniques across data, loss, structure, and constraints, we were able to mitigate overfitting and achieve strong and stable out-of-sample gains.