# Lab Report #4: Portfolio Construction
## Course: RSM338H1S, Winter 2026

**Instructions:**

- This assignment may be completed in **groups of up to 3 students**. Group members may be from either section. If you work in a group, submit one report with all names listed.
- Submit your lab report as a **PDF** to Crowdmark via Quercus. Export your Jupyter Notebook to PDF before uploading. Only one group member should upload the submission, being sure to select their group mates at the time of submission.
- There is no page limit, but be concise. A good report is thorough but not padded.

**Marking:**

- **75%** — Coding and results (correct implementation, complete answers to all parts, appropriate choice of methods, accurate numerical output, properly formatted tables and figures)
- **25%** — Overall quality (clear and professional writing, thoughtful interpretation of results, demonstrated understanding of the underlying concepts, logical flow and narrative structure)

**Writing Expectations:** Your report should read as a **coherent narrative**, not just code with scattered comments. Use section headers to indicate which problem you're working on. Before each code block, briefly explain what you are about to do and why. After results appear, interpret what you see. A reader should be able to understand your analysis even if they skipped the code cells.

You may use AI coding assistants (ChatGPT, Copilot, Claude, etc.) to help write code, but you must be able to explain what every line does. The text you write around the code is what demonstrates your understanding. You are ultimately responsible for your own work. **If you use an AI tool, you must disclose this in a note at the end of your report. Mention which tool you used, which tasks you asked it to complete, and discuss your (dis)satisfaction with its assistance.**

**Assignment:** You are a quantitative analyst at a pension fund that currently tracks a market-cap-weighted equity index. The investment committee wants to know whether they can do better by constructing an optimized portfolio using mean-variance theory. Your job is to take S&P 500 stock data, estimate the inputs to mean-variance optimization, build efficient portfolios, and stress-test whether those portfolios are actually usable—or whether estimation error makes them fall apart. Along the way, you will test whether the CAPM holds in practice and implement MAXSER (Ao, Li, and Zheng, 2019), a regularized portfolio construction method designed to produce portfolios that survive contact with real data.

**Data:** You are provided with three CSV files:

- `prices.csv` — Daily adjusted close prices for S&P 500 constituents, 1970–2025.
- `market.csv` — Daily closing values for the S&P 500 index (`^GSPC`), same date range.
- `DGS10.csv` — Daily 10-year U.S. Treasury constant maturity rate from FRED, 1962–2025. This is an annualized yield expressed as a percentage. To use it as a risk-free rate proxy, convert to a daily continuously compounded rate: $r_f^{\text{daily}} = \ln(1 + y/100) \,/\, 252$.

## Problem 1: From Prices to Returns

The starting point of any empirical finance project is converting raw price data into returns and building the inputs needed for portfolio optimization.

**Data Preparation:** Use all S&P 500 stocks in `prices.csv` over the most recent 10 years of data. You will encounter missing values, stocks that enter or leave the index, and other data quality issues—these are decisions you need to make and document. Describe what you did and why.

Align your stock prices, market index, and risk-free rate to a common set of dates.

**Tasks:**

(a) Compute daily **log returns** for each stock and for the market index:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$$

Briefly explain why log returns are preferred over simple returns for statistical analysis.

(b) Build the components needed for mean-variance optimization, **annualized from daily data**. With $T$ trading days of daily log returns:

**Annualized expected excess returns:**

$$\hat{\mu}_i = 252 \cdot \bar{r}_i - \bar{r}_f, \qquad \text{where } \bar{r}_i = \frac{1}{T}\sum_{t=1}^{T} r_{i,t} \text{ and } \bar{r}_f = \frac{1}{T}\sum_{t=1}^{T} r_{f,t}^{\text{ann}}$$

**Annualized covariance matrix:**

$$\hat{\mathbf{\Sigma}} = 252 \cdot \hat{\mathbf{\Sigma}}_{\text{daily}}$$

where $\hat{\mathbf{\Sigma}}_{\text{daily}}$ is the sample covariance matrix of daily log returns.

The factor of 252 (trading days per year) converts daily moments to annual units under the assumption that daily returns are i.i.d. Report the dimensions of $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{\Sigma}}$.

(c) Pick 2–5 stocks that stand out statistically in some way—unusually high or low average returns, extreme volatility, anomalous correlations with the market, etc. For each, find **news stories** or firm-specific events that help explain the pattern you observe. Briefly discuss whether the statistical anomaly was foreseeable or only obvious in hindsight.

## Problem 2: The Security Market Line

The CAPM predicts that expected excess returns are proportional to systematic risk ($\beta$). In this section, you will estimate betas, construct the Security Market Line, and investigate stocks that deviate from it.

**Data Preparation:** Aggregate your daily log returns to monthly frequency before running regressions. Since log returns are additive, the monthly log return is simply the sum of the daily log returns within each month. Do the same for the market index and the risk-free rate.

**Tasks:**

(a) For each stock $i$, run the CAPM regression using **monthly** excess returns:

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i \left(r_{m,t} - r_{f,t}\right) + \epsilon_{i,t}$$

Report $\hat{\beta}_i$, $\hat{\alpha}_i$, $R^2$, and the standard error of $\hat{\beta}_i$ for each stock.

(b) For 2–3 representative stocks (one high-$\beta$, one low-$\beta$, one mid-range), show the scatter plot of excess stock returns vs. excess market returns with the fitted regression line. Interpret the slopes.

(c) Plot the **Security Market Line**: $\hat{\beta}_i$ on the horizontal axis and average realized excess return on the vertical axis. Include the theoretical SML line for comparison. Identify stocks that lie significantly above the line (positive $\alpha$) and below the line (negative $\alpha$).

(d) For each outlier (at least 2 above and 2 below the SML), research the company and the relevant time period. Find specific **news stories**, events, or firm-specific developments that could explain the abnormal performance (e.g., a major product launch, an accounting scandal, a merger, a regulatory change).

(e) Connect your findings to the CAPM framework. Are these deviations evidence of mispricing, or do they reflect firm-specific shocks that diversified investors would not have been exposed to? What does this tell you about the distinction between $\alpha$ and luck?

# Problem 3: The Efficient Frontier

## 3.1 The Closed-Form Solution

For a target return $\mu_{\text{target}}$, the minimum-variance portfolio solves:

$$\min_{\vec{w}} \frac{1}{2}\vec{w}^\top \boldsymbol{\Sigma}\vec{w} \quad \text{subject to} \quad \begin{cases} \vec{w}^\top \boldsymbol{\mu} = \mu_{\text{target}} \\ \vec{w}^\top \vec{1} = 1 \end{cases}$$

The optimal weights depend on two multipliers $\lambda_1$ and $\lambda_2$ (one per constraint):

$$\vec{w}^* = \boldsymbol{\Sigma}^{-1}(\lambda_1 \boldsymbol{\mu} + \lambda_2 \vec{1})$$

Define the scalars:

$$A = \vec{1}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \qquad B = \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \qquad C = \vec{1}^\top \boldsymbol{\Sigma}^{-1}\vec{1}, \qquad D = BC - A^2$$

The multipliers are:

$$\lambda_1 = \frac{C \cdot \mu_{\text{target}} - A}{D}, \qquad \lambda_2 = \frac{B - A \cdot \mu_{\text{target}}}{D}$$

For any target return $\mu_{\text{target}}$, plug in $\lambda_1$ and $\lambda_2$ to get the portfolio weights $\vec{w}^*$. Sweeping over a range of target returns traces the efficient frontier.

**Tasks:**

(a) Using your estimated $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ from Problem 1, implement this closed-form solution and trace the efficient frontier across a range of target returns. Plot the frontier in $(\sigma, \mu)$ space.

(b) Identify the **Global Minimum Variance (GMV)** portfolio and the **tangency portfolio** (maximum Sharpe ratio). Report their weights, expected return, and standard deviation. Comment on the weights—are they reasonable? Do you see extreme long or short positions?

## 3.2 Using a Solver

This is a standard optimization problem—minimize a quadratic objective subject to linear constraints—so off-the-shelf **quadratic programming** solvers can handle it directly. In Python, `cvxpy` is a good choice. The advantage of a solver is that you can easily add constraints that the closed-form solution cannot accommodate.

**Tasks:**

(a) Re-solve the efficient frontier using `cvxpy` (or another QP solver). Verify that it matches your closed-form frontier from part (a).

(b) Now add a **long-only constraint** ($w_i \geq 0$ for all $i$) and re-solve. Plot the constrained frontier alongside the unconstrained frontier. How do the two compare? What is the "cost" of prohibiting short sales in terms of achievable risk-return combinations?

# Problem 4: MAXSER

The plug-in tangency portfolio from Problem 3 is an "error maximizer"—it exploits noise in $\hat{\boldsymbol{\mu}}$. The MAXSER approach (Ao, Li, and Zheng, 2019) recasts the problem as a Lasso regression, sidestepping the unstable matrix inverse. This problem walks you through the full procedure step by step.

**Frequency convention.** Everything in this problem uses **monthly excess returns at their native frequency**—the same data you already computed in Problem 2. Do not annualize anything. All quantities ($\hat{\theta}$, $r_c$, $\sigma$) must be at the same frequency.

## 4.1 Choosing a Subpool of Stocks

You have $N = 472$ stocks but only $T \approx 121$ monthly observations. Because $N > T$, the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ is singular—you literally cannot invert it. So before doing anything else, we need to select a smaller group of stocks where $N < T$.

The question is: *which* 50 stocks? A natural criterion is to pick the subset that offers the best risk–return tradeoff. The quantity that measures this is the **squared Sharpe ratio of the tangency portfolio**:

$$\theta = \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$

A higher $\theta$ means a better frontier is available from those stocks. Of course, we don't know the true $\theta$—we only have the sample estimate $\hat{\theta}_s = \hat{\boldsymbol{\mu}}^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}$, which is biased upward. To partially correct for this, use the Kan and Zhou (2007) adjustment:

$$\hat{\theta}_{\mathrm{KZ}} = \frac{(T - N - 2)\,\hat{\theta}_s \; - \; N}{T}$$

**Setup:**

1. Randomly draw **1,000 subpools** of $N^- = 50$ stocks from your 472-stock universe.
2. For each subpool, compute $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ from the monthly excess returns of those 50 stocks, then compute $\hat{\theta}_{\mathrm{KZ}}$ using the formula above.
3. Select the subpool at the **95th percentile** of $\hat{\theta}_{\mathrm{KZ}}$ across all 1,000 draws. Why the 95th and not the maximum? Because the maximum is likely an outlier—a subpool that got lucky with estimation noise. The 95th percentile gives a strong subpool without chasing a fluke.

All remaining work in this section uses this selected subpool of 50 stocks.

**Tasks:**

(a) Report the distribution of $\hat{\theta}_{\mathrm{KZ}}$ across the 1,000 subpools (median, 95th percentile, max). How many subpools have negative $\hat{\theta}_{\mathrm{KZ}}$?

(b) Report the selected subpool's $\hat{\theta}_{\mathrm{KZ}}$ and list the tickers.

## 4.2 Estimating the Squared Sharpe Ratio $\hat{\theta}$

The plug-in estimate of the squared Sharpe ratio on your selected subpool is:

$$\hat{\theta}_s = \hat{\boldsymbol{\mu}}^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}$$

This is badly biased upward.

ALZ (2019, equation 1.32) provide a corrected estimator that (i) removes most of the upward bias and (ii) is guaranteed to be non-negative:

$$\hat{\theta} = \underbrace{\frac{(T-N-2)\,\hat{\theta}_s - N}{T}}_{\text{Kan–Zhou term}} + \underbrace{\frac{2\,\hat{\theta}_s^{N/2}\,(1+\hat{\theta}_s)^{-(T-2)/2}}{T\;\cdot\;I_x(a,b)\;\cdot\;B(a,b)}}_{\text{ALZ correction (always}\geq 0)}$$

where:

- $x = \hat{\theta}_s \,/\, (1+\hat{\theta}_s)$
- $a = N/2, \quad b = (T-N)/2$
- $I_x(a,b)$ is the **regularized incomplete beta function**: in Python, `scipy.special.betainc(a, b, x)`
- $B(a,b)$ is the **beta function**: in Python, `scipy.special.beta(a, b)`

The first term (Kan–Zhou) can be negative when the sample evidence is weak. The second term (ALZ correction) is always $\geq 0$ and acts as a floor, ensuring you always get a usable $\hat{\theta} \geq 0$.

Recall from lecture that the tangency portfolio weights solve a regression against a constant target $r_c$:

$$\vec{w}^* = \arg\min_{\vec{w}}\ \mathbb{E}\big[(r_c - \vec{w}^\top \vec{r}_t)^2\big]$$

The target depends on your risk budget $\sigma$ and the bias-corrected $\hat{\theta}$:

$$\hat{r}_c = \sigma_{\text{monthly}} \cdot \frac{1+\hat{\theta}}{\sqrt{\hat{\theta}}}$$

**Tasks:**

(a) Report $\hat{\theta}_s$, the Kan–Zhou term, and the ALZ correction separately. How large is the bias correction relative to the plug-in estimate?

(b) Comment on what this tells you about the reliability of in-sample Sharpe ratios when $N/T$ is not small.

(c) Choose a risk budget (e.g., $\sigma_{\text{ann}} = 15\%$, so $\sigma_{\text{monthly}} = 0.15/\sqrt{12}$) and compute $\hat{r}_c$.

## 4.3 The Lasso Regression

Now we solve the MAXSER Lasso:

$$\hat{\vec{w}}_{\text{MAXSER}} = \arg\min_{\vec{w}} \frac{1}{T}\sum_{t=1}^{T}(\hat{r}_c - \vec{w}^{\top}\vec{r}_t)^2 + \lambda\|\vec{w}\|_1$$

The penalty $\lambda$ controls sparsity: larger $\lambda$ means fewer stocks get nonzero weight.

Rather than fitting the Lasso repeatedly at a grid of $\lambda$ values (which is slow), use the **LARS algorithm**: in Python, `sklearn.linear_model.lars_path` with `method='lasso'`. A single call traces the full solution path—it gives you the optimal weights at *every* value of $\lambda$, from the most regularized ($\vec{w} = \vec{0}$, all cash) to the least regularized (the OLS solution, no penalty at all).

It helps to parameterize the path by:

$$\zeta = \frac{\|\vec{w}\|_1}{\|\vec{w}_{\text{OLS}}\|_1} \in [0,\, 1]$$

At $\zeta = 0$ the portfolio is empty; at $\zeta = 1$ it is the unregularized OLS solution. You need to find the $\zeta$ that delivers a portfolio whose volatility matches your risk budget $\sigma_{\text{monthly}}$.

Select $\zeta$ via **10-fold cross-validation**:

1. Split your $T$ monthly observations into 10 folds. You *can* shuffle the data—this is not a forecasting problem. We are estimating the distributional properties of returns (means and covariances), so temporal ordering does not matter. Shuffling gives each fold a more representative mix of market regimes.
2. For each fold: compute the full LARS path on the **training set**, then for each step along the path, compute the **portfolio volatility on the held-out validation set**. Select the $\zeta$ whose validation-set volatility is closest to $\sigma_{\text{monthly}}$.
3. Average the selected $\zeta$ across all 10 folds.

Finally, compute the full-sample LARS path and extract the weights at the averaged $\hat{\zeta}$.

**Tasks:**

(a) Report the MAXSER portfolio weights. How many stocks receive nonzero weight?

## 4.4 Plug-In vs. MAXSER: Comparison

To compare MAXSER against the plug-in approach, both portfolios need to target the **same risk level**. The plug-in tangency portfolio from Problem 3 has whatever volatility it has (likely much higher than 15%). To put it on equal footing, scale it to the risk budget:

$$\vec{w}_{\text{plug-in}} = \frac{\sigma_{\text{monthly}}}{\sqrt{\hat{\theta}_s}}\, \hat{\mathbf{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}$$

You can verify that this gives $\sigma_p = \sigma_{\text{monthly}}$ by construction. The weights will not sum to one—the remainder $(1 - \vec{1}^\top\vec{w})$ is held in cash. Compute this on the same 50-stock subpool.

**Tasks:**

(a) Create a **side-by-side visualization** of the plug-in and MAXSER portfolio weights (e.g., horizontal bar charts). Highlight the differences: sparsity, position size, implementability. Which portfolio is more reasonable to implement?

(b) Build a comparison table with the following metrics: number of nonzero weights, portfolio concentration ($\text{HHI} = \sum_i w_i^2$), in-sample Sharpe ratio, maximum absolute weight, and risk-free weight.

(c) Why does MAXSER produce a sparser, more stable portfolio? Connect your answer to the concepts of estimation risk and regularization from lecture. What are the practical advantages of a sparse portfolio for implementation? Is the plug-in's higher in-sample Sharpe ratio meaningful, or is it inflated by overfitting?

# References

- Ao, M., Li, Y., & Zheng, X. (2019). Approaching mean-variance efficiency for large portfolios. *Review of Financial Studies*, 32(7), 2890–2919.
- Kan, R., & Zhou, G. (2007). Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 42(3), 621–656.