IAE Nantes
Économie & Management
Pôle Sociétés

Nantes
Université

Master 1 – Econometrics and Statistics
Applied Econometrics Track

Dissertation

Index Tracking and Asset Selection Using Penalized Regression Techniques

Florian CROCHET
Supervised by Mr. Olivier DARNÉ

Academic Year 2024–2025

Source code: github.com/Flo-1-618/Dissertation

Acknowledgements

I would like to express my sincere gratitude to my dissertation supervisor, Professor Darné, for his guidance and support. His suggestion of the topic, based on my interests, as well as his feedback, answers to my questions, and the reference materials he provided, were very helpful throughout this work.

I also wish to thank all the professors who contributed to my academic development. In particular, I am grateful to Professor Sévi for his course on univariate time series, which was especially relevant to this research, and to Professor Yayi for his course on Financial Asset Valuation, which helped strengthen my understanding of the subject.

Finally, I would like to thank my mentor, a second-year master's student, for sharing her knowledge and providing me with lessons and projects that supported my learning.

Abstract – Keywords

This dissertation addresses the problem of index tracking through sparse asset selection using penalized regression methods. Focusing on the S&P 500 index from 2017 to 2024, it applies modern regularization techniques (Ridge, Lasso, Elastic Net, and Adaptive Lasso) combined with non-negativity constraints to replicate index returns while selecting only a limited number of constituent stocks. The analysis incorporates rigorous data preprocessing, outlier adjustment, and variable screening via Distance Correlation Sure Independence Screening (DC-SIS), allowing for effective dimensionality reduction while preserving relevant dependencies. A rolling-origin cross-validation framework ensures robustness to look-ahead bias and serial dependence. Empirical results demonstrate that Adaptive Lasso and Elastic Net (with data-driven mixing parameter) consistently identify stable and parsimonious asset subsets that achieve low tracking error and strong replication performance. The findings highlight the practical relevance of penalized regression techniques for constructing cost-efficient, interpretable, and stable index-tracking portfolios, offering valuable insights for ETF design and passive investment strategies.

Keywords: S&P 500, index tracking, asset selection, penalized regression, R

Summary of Contents

Glossary / Acronyms

Acronym	Definition
ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller test
AIC	Akaike Information Criterion
Adaptive Lasso	Lasso with data-driven weights
AR	Active Return
Beta	Systematic risk coefficient (CAPM slope)
CAPM	Capital Asset Pricing Model
CV	Cross-validation
DC	Distance Correlation
DC-SIS	Distance Correlation Sure Independence Screening
Elastic Net	Penalized regression combining L1 and L2 penalties
ETF	Exchange-Traded Fund
High-dimensionality	Setting where variables greatly exceed observations
Hyperparameter	Tuning parameter controlling penalty strength in regularization
IR	Information Ratio
Jensen's Alpha	Risk-adjusted return beyond CAPM expectations
KPSS	Kwiatkowski-Phillips-Schmidt-Shin test
Lasso	Least Absolute Shrinkage and Selection Operator
Log-returns	Natural logarithm of price relatives, used for return calculation
Look-ahead bias	Bias from using future information in model estimation
Ljung-Box	Portmanteau test for serial autocorrelation
Mahalanobis Distance	Multivariate distance metric used for outlier detection
MCD	Minimum Covariance Determinant: robust estimator of multivariate location and scatter
MCP	Minimax Concave Penalty
Non-negativity constraint	Long-only constraint: portfolio weights ≥ 0
Oracle property	Ability of penalized regressions to select true model asymptotically
Out-of-sample	Model validation on data not used for estimation
PACF	Partial Autocorrelation Function
PP	Phillips-Perron test
RMSE	Root Mean Squared Error
Ridge	Ridge Regression (L2 penalty)

Acronym	**Definition**
Rolling-origin CV	Rolling-origin Cross-validation (time-series cross-validation)
S&P 500	Standard & Poor's 500 Index
SCAD	Smoothly Clipped Absolute Deviation
Shrinkage	Regularization technique shrinking coefficients toward zero
Shapiro-Wilk	Normality test for univariate distributions
SIS	Sure Independence Screening
Spearman	Spearman Rank Correlation
Stationarity	Property of constant mean and variance over time
TE	Tracking Error
Winsorization	Outlier adjustment technique limiting extreme values

1 Introduction

1.1 Presentation of the Topic

Trackers, or Exchange-Traded Funds (ETFs), are financial instruments designed to replicate the performance of an underlying index, such as the S&P 500, often by holding only a subset of its constituents. This partial replication reduces transaction costs and simplifies portfolio management but raises a central challenge in asset selection: which stocks should be included to best track the index?

This dissertation addresses this problem by applying penalized regression techniques (Ridge, Lasso, Elastic Net, and Adaptive Lasso) to select sparse subsets of assets while maintaining strong replication accuracy. To reflect practical investment constraints, non-negativity restrictions are imposed, ensuring that portfolio weights remain consistent with long-only strategies. Using historical data from the S&P 500 between 2017 and 2024, combined with variable screening via Distance Correlation Sure Independence Screening (DC-SIS) and rolling cross-validation, the study constructs and evaluates efficient index-tracking portfolios.

1.2 Relevance of the Topic

The relevance of this research lies in bridging financial econometrics and portfolio optimization through modern variable selection techniques. A key challenge in index tracking is to approximate benchmark performance while reducing the number of assets held, thereby lowering transaction costs and management complexity (Qian Chen et al. 2022). Penalized regression methods such as Lasso, Elastic Net, and Adaptive Lasso excel in this setting, as they simultaneously estimate model parameters and identify the most influential assets (Wu, Y. Yang, and H. Liu 2014a). When used under non-negativity constraints reflecting real-world no-short-sale rules, these methods facilitate the construction of parsimonious, efficient portfolios (Wu, Y. Yang, and H. Liu 2014a). This approach is particularly well-suited for the development of ETFs and structured investment products that must adhere to both regulatory standards and operational constraints.

1.3 Limitations of the Topic (space, time, themes, etc.)

This study has several limitations. It focuses solely on the S&P 500 index, with data spanning from January 2, 2017 to March 14, 2024; therefore, the findings may not generalize to other indices or time periods. In addition, the explanatory variables are limited to historical return series of the constituent companies, excluding macroeconomic indicators or firm-level fundamentals that could provide deeper insight (Silva-Filho et al. 2023). Finally, the analysis does not incorporate dynamic rebalancing or real-world transaction costs, which may affect the portfolio's actual performance in a live trading environment (Y. Shu, Zhang, and Yan 2020).

1.4 History of the Topic

The idea of index tracking can be traced back to the development of Modern Portfolio Theory and the Efficient Market Hypothesis, where the belief in the impossibility of consistently beating the market led to the rise of passive investing (Silva and Almeida Filho 2023). Since the 1970s, with the creation of the first index fund,[1] tracking strategies have evolved alongside financial econometrics (Silva and Almeida Filho 2023). In recent years, with the growth of high-dimensional data and machine learning techniques, methods such as penalized regressions have become central in addressing complex asset selection problems (Xia, Y. Yang, and H. Yang 2023).

1.5 Definition of Terms

This dissertation focuses on tracking the S&P 500 index, a capitalization-weighted benchmark composed of 500 large-cap publicly traded companies in the United States. The dataset includes the daily closing returns of these companies from January 2, 2017, to March 14, 2024. Asset selection refers to the process of choosing an optimal subset of these companies to construct a tracking portfolio(Qingyu Chen and al. 2022). Penalized regression models are used to identify the most relevant predictors by applying penalties that shrink less important coefficients toward zero, allowing the selection of the assets that most effectively explain the index's movements(Wu, Y. Yang, and H. Liu 2014b).

[1] The first index fund was created by John Bogle in 1976 through the Vanguard Group.(The Vanguard Group 2024)

1.6 Relation to Literature or Existing Work

This research builds on a growing body of work that applies statistical learning methods to financial portfolio construction. Previous studies have demonstrated the effectiveness of Lasso and its variants in selecting sparse subsets of assets to replicate index returns. Wu, Y. Yang, and H. Liu (2014b) introduced the nonnegative Lasso for constrained index tracking, while Xia, Y. Yang, and H. Yang (2023) and Y. Yang and Wu (2016) extended this framework by applying nonconvex penalties such as SCAD and Adaptive Lasso to improve selection consistency and reduce estimation bias. These contributions have laid the groundwork for practical applications in ETF design (Y. Liu, L. Li, and H. Yang 2024; L. Shu, Shi, and Tian 2020), and this dissertation aims to further refine and empirically validate these methods.

1.7 Research Question and Key Contributions

This dissertation addresses a key challenge in passive investment: replicating a market index with a reduced set of assets. The central research question is: How can penalized regression methods be used to select a small subset of assets that best replicate the performance of the S&P 500 index? The study contributes in three main ways. First, it implements and compares a range of penalized regression techniques, all under nonnegativity constraints to reflect realistic investment conditions. Second, it empirically evaluates the resulting portfolios using tracking error and multiple financial performance measures. Third, it provides practical insights into how these methods can guide the construction of cost-effective, efficient passive investment products.

1.8 Methods of Analysis Employed and Distinction from Existing Literature

This dissertation employs a combination of advanced econometric and machine learning methodologies to address the index tracking problem under realistic investment constraints. The core approach utilizes penalized linear regression techniques — including Ridge, Lasso, Elastic Net, and Adaptive Lasso — all estimated under non-negativity constraints to reflect the long-only investment strategies typical of ETF construction.

A central methodological contribution of this study is the integration of Distance Correlation Sure Independence Screening (DC-SIS) as a preliminary dimensionality reduction step. This technique enables the efficient handling of high-dimensional financial data while capturing both linear and nonlinear dependencies between asset returns and index returns. To ensure robust out-of-sample performance and mitigate look-ahead bias, a rolling-origin cross-validation framework is implemented, explicitly accounting for the temporal dependencies often overlooked in prior studies.

In contrast to much of the existing literature, which tends to rely on static models or neglect serial dependence in financial time series, this dissertation emphasizes dynamic model validation, realistic portfolio constraints, and replicability. As a result, the proposed framework generates stable, parsimonious, and interpretable tracking portfolios, offering practical relevance for ETF design and sparse portfolio management.

1.9 Presentation of the Structure

The dissertation begins by introducing the research question, defining key concepts, reviewing the literature, and presenting the methodological framework. It then describes the economic environment of index tracking, focusing on the S&P 500 index as the empirical benchmark. The econometric methodology follows, detailing data preprocessing, variable screening using Distance Correlation Sure Independence Screening (DC-SIS), and the application of penalized regression models (Ridge, Lasso, Elastic Net, Adaptive Lasso) under non-negativity constraints, combined with rolling-origin cross-validation. The empirical results are then presented, covering exploratory analysis, variable selection, portfolio construction, and performance evaluation across multiple financial metrics. The dissertation concludes with a summary of findings, discussion of limitations, proposals for future research, and practical recommendations for ETF design and asset management.

2 Economic Environment

2.1 Index Tracking and ETFs

Exchange-Traded Funds (ETFs) are investment vehicles that aim to replicate the performance of market indices like the S&P 500. ETFs achieve broad diversification with relatively low transaction costs, making them attractive for both institutional and retail investors. Replication strategies for ETFs typically fall into two categories: full replication and partial replication. Full replication involves holding every index constituent in proportion to its index weight, minimizing tracking error but increasing transaction costs and portfolio complexity. Partial replication, by contrast, seeks to approximate index performance using only a subset of constituents, thereby reducing costs and simplifying management, while controlling tracking error.

Because index replication using only a subset of assets creates a variable selection problem under high-dimensionality, statistical learning methods offer natural solutions. Partial replication motivates the use of sparse regression models that can automatically select the most influential subset of assets while shrinking insignificant positions to zero.

Sparse regression models incorporating nonnegative and adaptive penalties have proven highly effective in this context. For example, time-weighted nonnegative Lasso models have been developed to account for time-varying market conditions (Qian Chen et al. 2022), while group-based penalties such as the nonnegative group bridge have been proposed to address the natural clustering and correlation structures present in financial data (Y. Liu, L. Li, and H. Yang 2024). Furthermore, non-convex penalties such as the minimax concave penalty (MCP) have been applied in conjunction with nonnegativity constraints to improve estimation accuracy and variable selection consistency in high-dimensional index tracking settings (X. Li and Y. Yang 2021).

These advanced models provide a data-driven approach to constructing tractable, cost-efficient, and economically viable replication portfolios that meet real-world investment constraints while achieving robust performance.

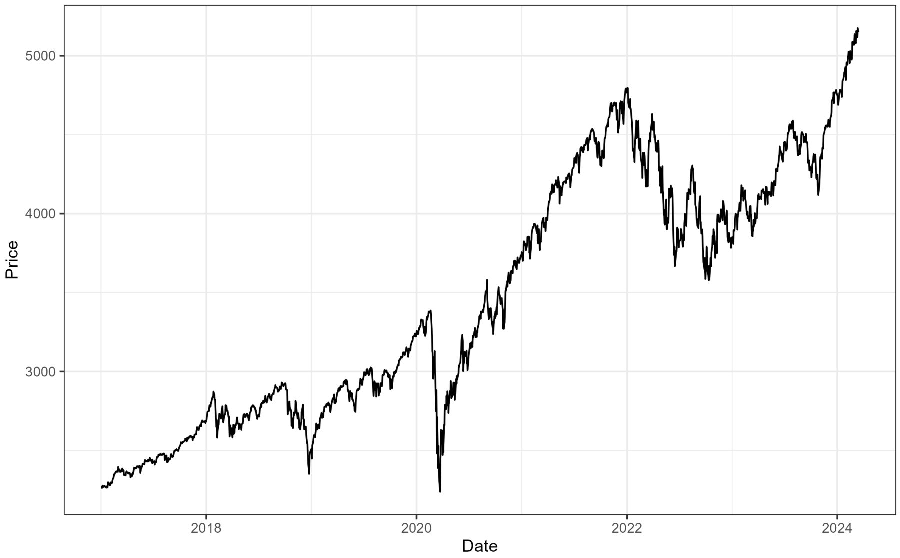
2.2 S&P 500 Index and Asset Selection Framework

The empirical component of this dissertation focuses on the S&P 500 index, which is one of the most widely recognized benchmarks for large-cap U.S. equities. The index includes companies across a broad spectrum of industries and is subject to regular rebalancing based on its inclusion criteria, resulting in both opportunities and complexities for index replication using a subset of constituents.

The S&P 500 is a float-adjusted, market-capitalization-weighted index. Each constituent's weight reflects its share of the total float-adjusted market capitalization. The index level at any time t is computed as the aggregate float-adjusted market capitalization of its constituents divided by a divisor that accounts for corporate actions such as stock splits, share issuances, and constituent changes. As a result, larger firms exert a disproportionate influence on the overall index performance, a factor that directly shapes the asset selection framework applied in this dissertation (S&P Dow Jones Indices 2023).

The empirical analysis relies on historical daily price data for both the S&P 500 index and its constituents, compiled by my dissertation supervisor, covering the period from January 3, 2017, to March 14, 2024. This time window captures a variety of market conditions, including sustained growth, periods of elevated volatility, and sharp market corrections, providing a comprehensive environment for evaluating index tracking models.

Figure 2.1: S&P 500 Price



Source: Yahoo Finance.

Figure 2.1 presents the evolution of the S&P 500 index during the sample period. The index shows long-term growth interrupted by significant downturns such as the COVID-19 shock in early 2020, followed by a rapid recovery and upward trend into 2024. These fluctuations illustrate the importance of selecting representative subsets of stocks that capture the index's behavior under varying market conditions.

Figure 2.2: Overlapping Prices of All Stocks

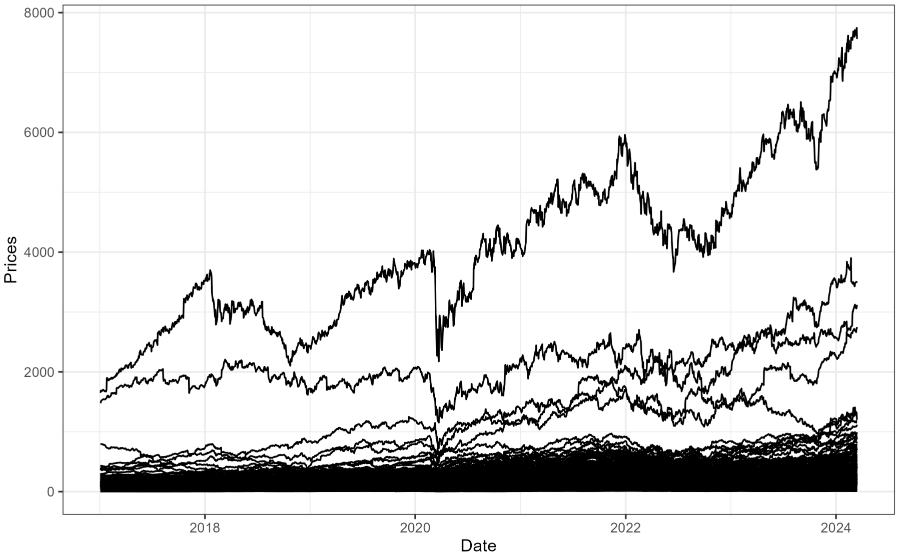


Figure 2.2 illustrates the diversity in stock price trajectories across the different constituents, reflecting considerable heterogeneity in growth paths, sectors, and volatility levels. This cross-sectional variation directly motivates the use of sophisticated asset selection techniques for index replication (H. Shu, Wu, and Y. Yang 2020; Qian Chen et al. 2022).

2.3 Sparse Penalized Regressions for Index Tracking

The asset selection methodology used in this dissertation relies on daily returns of the S&P 500 constituent stocks from January 3, 2017, to March 14, 2024. These returns serve as predictors in regression models designed to approximate the S&P 500 index return using only a subset of its constituents.

In high-dimensional settings where the number of potential predictors can far exceed the number of observations, traditional estimation techniques such as ordinary least squares become unstable. Penalized regression techniques, by introducing shrinkage penalties, effectively mitigate overfitting, multicollinearity, and enhance model interpretability in such settings.

Among these methods, the Least Absolute Shrinkage and Selection Operator (Lasso), introduced by Tibshirani (1996b), is one of the most widely used techniques, applying an ℓ_1-norm penalty to encourage sparsity in the estimated coefficients. Extensions of the Lasso such as the Elastic Net and Adaptive Lasso have further improved model performance, particularly when predictor variables are highly correlated or when oracle properties are desired.

In the context of index tracking, nonnegative constraints are especially important since portfolio weights are typically restricted to be nonnegative to reflect long-only investment strategies and regulatory constraints. To address this, modified versions of penalized regression methods incorporating nonnegativity constraints have been proposed and applied effectively in financial applications (Wu and Y. Yang 2014; Y. Yang and Wu 2016).

Penalized regression models simultaneously select relevant variables and estimate their coefficients, generating sparse replication portfolios that balance tracking accuracy with practical manageability.

2.4 Conclusion

A comprehensive econometric framework was developed to address the index tracking problem through sparse asset selection. This approach integrates advanced data preprocessing, robust variable screening via Distance Correlation Sure Independence Screening (DC-SIS), and penalized regression models subject to non-negativity constraints, allowing for the identification of optimal subsets of S&P 500 constituents. The incorporation of rolling-origin cross-validation further enhances the reliability of the model calibration by accounting for temporal dependencies and mitigating look-ahead bias. Together, these methodological choices provide a solid foundation for constructing parsimonious, interpretable, and practically implementable index-tracking portfolios. The empirical implementation of this framework is now carried out using the S&P 500 constituent data, with the objective of evaluating the asset selection outcomes and the tracking performance of the constructed portfolios relative to the benchmark index.

3 Econometric Methodology

In this chapter, we present the econometric methodology developed to address the index tracking problem using sparse asset selection techniques. The objective is to construct parsimonious portfolios that replicate the performance of the S&P 500 index while holding only a subset of its constituent stocks. To achieve this, we rely on penalized regression models, which impose regularization constraints that promote sparsity and mitigate overfitting, even in high-dimensional settings where the number of predictors can be large relative to the sample size. Specifically, Ridge, Lasso, Elastic Net, and Adaptive Lasso regressions are employed, all subject to non-negativity constraints consistent with practical investment restrictions. These models are calibrated using rolling-origin cross-validation to ensure robustness against temporal dependencies and look-ahead bias. The chapter also details the data preprocessing steps, variable screening procedures, and hyperparameter tuning strategies that together form the foundation for subsequent empirical analysis.

This study employs a regression-based econometric framework to address the index tracking problem, focusing on replicating the S&P 500 using a sparse subset of its constituents. The problem is formulated as a linear regression model in which index returns are regressed on the returns of candidate constituent stocks. The objective is to select a parsimonious portfolio, a small set of stocks and corresponding weights, that closely replicates the index's performance.

To achieve sparsity and mitigate overfitting, we apply penalized regression techniques that impose constraints on model complexity by shrinking coefficient estimates toward zero. This regularization reduces variance and effectively excludes less relevant variables by shrinking their coefficients to (or near) zero. Specifically, we employ four shrinkage methods: Ridge, Lasso, Elastic Net, and Adaptive Lasso, each offering varying balances between shrinkage and variable selection. All regressions are subject to non-negativity constraints (prohibiting short-selling), consistent with practical index replication strategies. Model tuning (i.e., selection of penalty parameters) is performed via cross-validation on a rolling time window to ensure out-of-sample validity and guard against lookahead bias, as detailed below.

All analyses were conducted in the R statistical environment, utilizing specialized packages for data handling and model estimation. Financial data retrieval and management were facilitated by R's data libraries, with the `xts` and `zoo` packages employed for time-series operations, and `PerformanceAnalytics` for portfolio performance evaluation.

Penalized regressions were implemented using the `glmnet` package, with modifications to enforce non-negativity constraints through parameter restrictions. Variable pre-screening was performed using the Sure Independence Screening (SIS) procedure based on distance correlation (DC-SIS), implemented via the `VariableScreening` package. Standard econometric diagnostics, including stationarity tests from the `urca` package, were applied throughout to validate model assumptions. This integrated software approach ensures methodological rigor, reproducibility, and academic transparency.

3.1 Exploratory and Descriptive Analysis

Preliminary data analysis before applying penalized regressions

3.1.1 Data Retrieval

For this study, two datasets were used. The first, provided by my dissertation supervisor, contains daily prices for all S&P 500 constituent stocks from January 2, 2017, to March 14, 2024. The second dataset, retrieved from Yahoo Finance, includes daily closing prices of the S&P 500 index over the same period.

Although both datasets cover the same overall date range, they differ in the specific dates included. The index dataset contains only trading days, excluding weekends and official U.S. stock market holidays. In contrast, the asset-level dataset includes 68 non-trading days (such as New Year's Day, Independence Day, and other holidays) when the market was closed and no price movement occurred. Since these days are absent from the index data and not relevant for return-based analysis, I removed them from the asset dataset. These non-trading days correspond to official U.S. market holidays as defined by the NYSE holiday calendar.

3.1.2 Missing Values

We identified missing values in both the S&P 500 index and its constituent assets. For each instance of missing data, we conducted a systematic assessment to evaluate data availability and investigate the underlying causes, such as corporate events or structural corporate changes.

3.1.3 Return Calculation

After cleaning the data, we transformed the price series into returns. Returns, rather than raw prices, are preferable because they exhibit desirable statistical properties such as (approximate) stationarity and variance stabilization, which are important for many econometric and statistical models.

We calculated daily logarithmic (log) returns using the formula:

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

where P_t is the closing price on day t. Logarithmic returns are additive over time and allow for easier aggregation and modeling.

3.1.4 Outlier Detection and Adjustment

Boudt, Cornelissen, and Croux (2013) propose a robust multivariate method to detect outliers in financial returns. The method identifies extreme returns based on Mahalanobis distances computed from the Minimum Covariance Determinant (MCD) estimator of location and scatter. Observations exceeding the cutoff were winsorized by replacing them with the corresponding threshold value to limit their influence while preserving the time series structure. Specifically, winsorization was applied according to:

$$x_t^* = \begin{cases} L, & \text{if } x_t < L \\ x_t, & \text{if } L \leq x_t \leq U \\ U, & \text{if } x_t > U \end{cases}$$

where x_t denotes the observed return at time t, and L and U represent the lower and upper bounds determined by the robust Mahalanobis distance cutoff.

3.1.5 Stationarity Testing

We assess stationarity using a combination of unit root and stationarity tests, following the methodology taught by Professor Sevi (2024). Specifically, we apply the Augmented Dickey-Fuller (ADF), Phillips-Perron (PP), and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests to exploit their complementary hypotheses and improve robustness.

The ADF test is conducted sequentially to account for deterministic components, starting with a model including both a constant and a linear trend. If the unit root hypothesis is not rejected, we progressively simplify the specification by removing the trend and then the constant. Lag selection is performed using the Akaike Information Criterion (AIC). The PP test is applied in parallel under identical specifications, adjusting for serial correlation and heteroskedasticity. The KPSS test complements the unit root tests by directly testing for stationarity in both level and trend. Joint interpretation of these tests allows for a more reliable diagnosis of the series' stationarity properties.

3.1.6 Descriptive Statistics

After visualizing the data, descriptive statistics including mean, median, standard deviation, skewness, and kurtosis were calculated to assess central tendency, dispersion, and distributional characteristics. The Shapiro-Wilk test was applied to individual variables to evaluate univariate normality, with the understanding that large sample sizes may yield statistically significant results for minor deviations. Since the primary goal was variable selection using penalized regression methods (Ridge, Lasso, Elastic Net, Adaptive Lasso), which are robust to non-normality (Hastie, Tibshirani, and Friedman 2009), deviations from normality were not considered problematic for the analysis.

3.1.7 Correlation Structure

Given the non-normality and potential nonlinearities of financial returns, we employed Spearman rank correlation to capture monotonic relationships between individual stocks and the index. To further account for both linear and nonlinear dependencies, we applied Distance Correlation Sure Independence Screening (DC-SIS), providing a more comprehensive variable screening. The selected subset was then used as input for penalized regression models, which offer robustness against multicollinearity.

3.1.8 Autocorrelation Analysis

The presence of autocorrelation in the returns of each asset and of the S&P 500 was initially assessed using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). The ACF provided insight into linear dependencies across lags, while the PACF helped isolate direct lag-specific relationships by controlling for intermediate lags. Identifying significant autocorrelations at specific lags can reveal temporal dependencies that may bias parameter estimation and variable selection in penalized regression models.

To formally test for joint serial dependence across multiple lags, the Ljung-Box portmanteau test was applied, evaluating the null hypothesis that autocorrelations up to lag K are jointly zero. The choice of K was informed by a combination of the sample size ($K \approx \sqrt{N}$) and visual inspection of the ACF and PACF plots. Significant Ljung-Box statistics indicated the presence of serial correlation in some series.

The Ljung-Box statistic is defined as:

$$Q_K = T(T+2)\sum_{k=1}^{K} \frac{1}{T-k}\hat{\rho}_k^2$$

where T is the sample size, K is the number of lags, and $\hat{\rho}_k$ is the sample autocorrelation at lag k. Under the null hypothesis $H_0 : \rho_1 = \rho_2 = \cdots = \rho_K = 0$, Q_K asymptotically follows a chi-squared distribution with K degrees of freedom:

$$Q_K \overset{a}{\sim} \chi^2(K).$$

This formulation follows the approach taught by Professor Yayi (2024).

Recognizing that even weak autocorrelation can distort model evaluation in time series data, a rolling-window cross-validation procedure was employed. This approach preserves the temporal ordering of observations, reduces look-ahead bias, and accommodates potential time dependencies that may arise in financial returns, even under weak-form market efficiency.

3.1.9 Conclusion

The exploratory and descriptive analysis produced a clean, reliable, and properly transformed dataset, addressing common issues in financial time series such as missing values, non-trading days, outliers, non-stationarity, and autocorrelation. Through rigorous preprocessing—including outlier adjustment, stationarity testing, and thorough evaluation of correlation and autocorrelation structures—we ensured that the resulting return series are well-suited for high-dimensional variable selection. With the data now prepared and key distributional properties established, we proceed to the core stage of the analysis: applying penalized regression models for variable selection and index replication.

3.2 Variable Selection

3.2.1 Cross-Validation Framework

To mitigate look-ahead bias and account for serial dependence in the return series, we employed a rolling-origin cross-validation framework applied consistently to both the variable screening stage (DC-SIS) and the subsequent penalized regression models, ensuring strict preservation of the temporal ordering throughout. The procedure was implemented using rolling windows, where each training window comprised 504 consecutive observations (approximately two years of daily returns), followed by a 21-observation out-of-sample test window, corresponding to one month of daily returns. After each test window, the rolling window advanced by 21 observations (i.e., a skip of 20), while maintaining a constant training window length across all iterations.

3.2.2 Sure Independence Screening Based on Distance Correlation

Prior to penalized regression, we employed Sure Independence Screening based on Distance Correlation (DC-SIS) to reduce dimensionality while preserving relevant predictors. Distance correlation, introduced by Székely, Rizzo, and Bakirov (2007), measures dependence between random variables and can detect both linear and nonlinear associations, making it particularly well-suited for financial return data where such complex dependencies often occur.

Given n observations, let $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ denote the index returns (response variable), and for each predictor $j = 1, \ldots, p$, let $X_j = (X_{j1}, \ldots, X_{jn})^\top \in \mathbb{R}^n$ denote the stock returns.

The population distance correlation between X_j and Y is defined as

$$\mathcal{R}(X_j, Y) = \frac{\mathcal{V}(X_j, Y)}{\sqrt{\mathcal{V}(X_j, X_j)\mathcal{V}(Y, Y)}},$$

where $\mathcal{V}(X_j, Y)$ denotes the distance covariance.

In practice, given the sample $\{(X_{j1}, Y_1), \ldots, (X_{jn}, Y_n)\}$, we compute pairwise Euclidean distance matrices $A_{kl} = |X_{jk} - X_{jl}|$ and $B_{kl} = |Y_k - Y_l|$ for $k, l = 1, \ldots, n$. These distance matrices are double-centered to remove location effects and obtain

$$\tilde{A}_{kl} = A_{kl} - \bar{A}_{k\cdot} - \bar{A}_{\cdot l} + \bar{A}_{\cdot\cdot},$$

and analogously for $\tilde{B}_{kl}$. The sample distance covariance is then estimated as

$$\hat{\mathcal{V}}^2(X_j, Y) = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} \tilde{A}_{kl} \tilde{B}_{kl},$$

and the corresponding sample distance correlation is

$$\hat{\mathcal{R}}(X_j, Y) = \frac{\hat{\mathcal{V}}(X_j, Y)}{\sqrt{\hat{\mathcal{V}}(X_j, X_j)\hat{\mathcal{V}}(Y, Y)}}.$$

We calculated $\hat{\mathcal{R}}(X_j, Y)$ for each predictor X_j and ranked all variables accordingly. The top d variables with the highest sample distance correlations were retained for subsequent penalized regression modeling. This screening procedure allows for effective dimensionality reduction while capturing potentially complex dependencies between individual stocks and the index.

To ensure consistency and avoid data leakage, the DC-SIS method was applied within each training window of the rolling-origin cross-validation procedure described in Section 3.2.1. Specifically, for each rolling window, distance correlations were calculated using only the observations available in the training set used to fit the penalized regression models.

For comparison, penalized regressions were first conducted on the full set of predictors without preliminary screening, and then repeated after applying DC-SIS.

3.2.3 Penalized Regressions

With clean and stationary data, we selected the most relevant variables applying four penalized regression methods: Ridge, Lasso, Elastic Net, and Adaptive Lasso, following the methodology introduced by Professor Darné (2024). These approaches address multicollinearity, control model complexity, and perform automatic variable selection in high-dimensional financial data. Each method applies a distinct form of regularization, balancing bias and variance to improve prediction accuracy while mitigating overfitting risks.

Prior to estimation, all predictor variables X_j were standardized (centered and scaled to unit variance), while the response variable Y was centered. This preprocessing ensures scale invariance of the penalty terms, removes the intercept, and avoids spurious effects arising from differing units or magnitudes across variables (Darné 2024).

To reflect regulatory constraints and practical considerations in ETF portfolio construction, non-negativity constraints were imposed on all regression coefficients.

We modeled the index returns y_t as a linear combination of p asset returns $x_{j,t}$:

$$y_t = \beta_0 + \sum_{j=1}^{p} \beta_j x_{j,t} + \varepsilon_t,$$

subject to $\beta_j \geq 0$ for all j. Adaptive Lasso, which assigns data-driven penalty weights, was included to enhance variable selection consistency.

The general penalized regression problem can be written as:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} P(\beta_j) \right\} \quad \text{s.t. } \beta_j \geq 0, \ \forall j,$$

where $P(\beta_j)$ denotes the penalty function specific to each method, and $\lambda \geq 0$ is the regularization parameter controlling the degree of shrinkage.

The estimation of these penalized models was performed within the rolling-origin cross-validation framework described in Section 3.2.1, which mitigates look-ahead bias and accounts for autocorrelation in returns. In each rolling window, hyperparameters were selected by minimizing the root mean squared error (RMSE) computed on the out-of-sample test sets. Specifically, for the penalized regression models, the elastic net mixing parameter $\alpha \in [0, 1]$ and the regularization parameter $\lambda > 0$ were jointly optimized over a prespecified grid of candidate values. The final grid consists of a logarithmically spaced range for λ from 10^{-7} to 10^{-2} and α values from 0.05 to 0.95 in increments of 0.05; this grid was manually refined based on preliminary results to better concentrate the search around regions near the estimated minima. For each hyperparameter combination, a penalized regression model was estimated on the training set and evaluated on the corresponding out-of-sample test set. Model performance was assessed based on RMSE, and the hyperparameter combination yielding the lowest average out-of-sample RMSE across all cross-validation folds was selected.

3.2.3.1 Ridge Regression (L2 penalty)

Proposed by Hoerl and Kennard (1970), Ridge regression addresses multicollinearity by applying an ℓ_2-norm penalty that shrinks the regression coefficients towards zero. Specifically, Ridge regression estimates the coefficients by solving the following optimization problem:

$$\hat{\beta}_{\text{Ridge}} = \arg\min_{\beta \geq 0} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

The penalty function is therefore defined as $P(\beta_j) = \beta_j^2$. Ridge regression is particularly effective in handling multicollinearity, as it distributes shrinkage across correlated predictors.

In the unconstrained version, Ridge does not shrink any coefficient exactly to zero, and thus does not perform variable selection. However, when non-negativity constraints are imposed, as in the present ETF portfolio context, some coefficients may be set exactly to zero if their unconstrained estimates would have been negative.

3.2.3.2 Lasso Regression (L1 penalty)

Introduced by Tibshirani (1996a), Lasso regression applies an ℓ_1-norm penalty that induces sparsity by shrinking some coefficients exactly to zero, thus simultaneously performing estimation and variable selection. The Lasso estimator solves:

$$\hat{\beta}_{\text{Lasso}} = \arg\min_{\beta \geq 0} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where the penalty function is $P(\beta_j) = |\beta_j|$. This formulation leads to automatic variable selection and model simplification. Nevertheless, Lasso may behave suboptimally in the presence of highly correlated predictors, as it tends to select only one variable from a group of correlated variables, potentially excluding relevant predictors due to its selection bias.

3.2.3.3 Elastic Net Regression (L1 + L2 penalty)

To overcome the limitations of Lasso in the presence of correlated predictors, Zou and Hastie (2005) introduced the Elastic Net, which combines the ℓ_1 and ℓ_2 penalties. The Elastic Net estimator solves:

$$\hat{\beta}_{\text{EN}} = \arg\min_{\beta \geq 0} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j| \right) \right\},$$

where $P(\beta_j) = \alpha\beta_j^2 + (1-\alpha)|\beta_j|$, and the hyperparameter $\alpha \in [0,1]$ governs the convex combination of Ridge and Lasso penalties. When $\alpha = 0$, the Elastic Net reduces to the Lasso; when $\alpha = 1$, it reduces to Ridge regression. The Elastic Net exhibits the grouping effect, meaning that it tends to select entire groups of highly correlated variables together—a feature particularly relevant in financial data (Darné 2024).

In the empirical analysis, two Elastic Net configurations were estimated. The first model was implemented by setting $\alpha = 0.5$, which represents an equal weighting between the Ridge and Lasso penalties. For the second, the hyperparameter α was selected using grid search over the candidate value range from 0.05 to 0.95 (in 0.05 increments). For each value of α, the regularization parameter λ was jointly optimized using grid search within each rolling window. This approach allowed the model to adaptively balance shrinkage and sparsity based on the data characteristics in each estimation window.

3.2.3.4 Adaptive Lasso Regression

While Lasso may introduce estimation bias for large coefficients, Adaptive Lasso, introduced by Zou (2006), corrects for this by assigning data-driven weights to each coefficient. The Adaptive Lasso estimator solves the following optimization problem:

$$\hat{\beta}_{\text{aLasso}} = \arg\min_{\beta \geq 0} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \omega_j |\beta_j| \right\},$$

where the adaptive weights are defined as $\omega_j = \frac{1}{|\hat{\beta}_j^*|^\gamma}$, with $\hat{\beta}_j^*$ denoting initial consistent estimates, typically obtained from Ridge or Lasso regressions, and $\gamma > 0$ controlling the weight decay. In this study, the initial estimates $\hat{\beta}_j^*$ were computed via Ridge regression, with the weight decay parameter set to $\gamma = 1$, resulting in $\omega_j = \frac{1}{|\hat{\beta}_j^*| + \varepsilon}$, where a small constant $\varepsilon = 10^{-5}$ was added to avoid division by zero and ensure numerical stability. By incorporating these data-driven weights, Adaptive Lasso enjoys the so-called oracle property: it can correctly identify the true model asymptotically while reducing bias for large coefficients (Darné 2024).

3.2.4 Conclusion

In this study, we developed a comprehensive variable selection methodology to identify the most relevant S&P 500 stocks for replicating the index using penalized regression techniques. By implementing a robust rolling-origin cross-validation framework to mitigate look-ahead bias and account for serial dependence in the return series, we applied Sure Independence Screening based on Distance Correlation (DC-SIS) to reduce dimensionality and retain the most significant predictors. Ten regressions using four penalized regression methods—Ridge, Lasso, Elastic Net (with both fixed and grid-search values for α), and Adaptive Lasso—were employed to select relevant stocks, with and without DC-SIS preselection. The inclusion of DC-SIS ensured that only the most relevant stocks were selected, improving model interpretability and regression efficiency while capturing the complex dependencies in the data. This methodology allowed us to identify key assets from a high-dimensional set of stock returns, offering a balanced trade-off between model complexity and predictive performance. The 10 regressions resulted in 10 distinct sets of selected stocks, with non-negative coefficients, that form a robust basis for the next phase of ETF construction and index tracking.

3.3 Performance of Constructed Portfolios Relative to the Stock Index

3.3.1 Portfolio Construction

In the next phase of the study, we transition from variable selection to portfolio construction, with the goal of creating ETF portfolios that replicate the performance of the S&P 500 index, using the relevant stocks selected by each regression in the previous section. The following six steps outline the methodology for constructing these portfolios under constant non-negative weights for each of the 10 variable selection regressions.

3.3.1.1 Penalized Regression to Estimate Asset Exposures

The portfolio construction process began with the application of penalized regression techniques to estimate the exposure of the ETF to each selected asset. Specifically, we regressed the ETF log-returns r_t^{ETF} on the log-returns of the individual assets $r_{i,t}$. This was performed using penalized linear models such as Ridge, Lasso, Elastic Net, or Adaptive Lasso, where the regression model was given by:

$$r_t^{\text{ETF}} = \sum_{i=1}^{N} w_i \cdot r_{i,t} + \varepsilon_t,$$

where w_i represented the regression coefficients (i.e., the exposure of the ETF to each asset), N was the total number of assets, and ε_t was the error term. The penalization encouraged sparsity, selecting only a subset of the assets that were most relevant for replicating the ETF's returns. This process was repeated for each of the 10 variable selection regressions, producing 10 distinct sets of regression coefficients corresponding to the 10 different ETF portfolios.

3.3.1.2 Normalization of Regression Coefficients

After estimating the regression coefficients w_i for each model, we normalized them to ensure that the total weight of each ETF portfolio summed to one. This normalization was crucial to guarantee valid portfolio allocation weights. The normalized weights were computed as:

$$w_i^{\text{norm}} = \frac{w_i}{\sum_{j \in S} w_j},$$

where S was the set of selected assets for each regression model, with $w_j \neq 0$. This ensured that the weights were proportional to the magnitude of the coefficients, while maintaining the constraint that the total weight across the selected assets summed to one for each ETF portfolio.

3.3.1.3 Computing ETF Returns

After determining the normalized portfolio weights for each of the 10 regression models, the next step involved computing the reconstructed log-return for each ETF. The ETF's return at each time t was the weighted sum of the returns of the selected assets:

$$\hat{r}_t^{\text{ETF}} = \sum_{i \in S} w_i^{\text{norm}} \cdot r_{i,t}.$$

This weighted sum represented the performance of the ETF at each time point for each of the 10 portfolios, where each asset contributed according to its exposure to the ETF, as determined by the regression coefficients.

3.3.1.4 Price Reconstruction of the ETF

The price of each ETF was then reconstructed using the cumulative sum of the ETF's returns. Starting with an initial price $P_0^{\text{ETF}} = 100$, the price at time t was computed using the standard log-return to price reconstruction formula:

$$P_t^{\text{ETF}} = P_0^{\text{ETF}} \cdot \exp\left(\sum_{s=1}^{t} \hat{r}_s^{\text{ETF}}\right).$$

Setting an initial price of 100 facilitates comparison with the S&P 500 and allows for a straightforward interpretation of performance in percentage terms. This widely used normalization in financial research serves as a common benchmark for comparing alternative strategies and helps make performance more intuitive for investors (Elton et al. 2014). By using this method, each ETF's price evolved over time, reflecting compounded growth based on the returns of the selected assets from each of the 10 regression models.

3.3.1.5 Calculating Investment in Each Asset

At each time t, the monetary investment in each asset for each ETF was calculated. This was determined by multiplying the normalized weight of the asset w_i^{norm} by the ETF's price at time t:

$$\text{Investment}_{i,t} = w_i^{\text{norm}} \cdot P_t^{\text{ETF}}.$$

This step ensured that the amount of capital allocated to each asset in each ETF was proportional to its weight in the portfolio at each time point.

3.3.1.6 Calculating the Number of Shares to Purchase

Finally, the number of shares of each asset to purchase for each ETF portfolio was computed by dividing the investment in each asset by its price at time t. The number of shares $q_{i,t}$ to be purchased at each time point was given by:

$$q_{i,t} = \frac{\text{Investment}_{i,t}}{P_{i,t}} = \frac{w_i^{\text{norm}} \cdot P_t^{\text{ETF}}}{P_{i,t}}.$$

This ensured that the portfolio was structured such that each asset was purchased in the correct proportion based on its weight in the ETF for each of the 10 portfolios. The number of shares purchased adjusted over time with the ETF's price.

3.3.1.7 Conclusion

By following these six steps, the portfolio construction methodology created 10 ETFs, each designed to track the performance of the S&P 500 index. The process began with penalized regression to identify the relevant assets and estimate their exposures to the ETF for each regression model. These coefficients were then normalized to ensure valid allocation, and the ETF's return and price were calculated based on these weights. The monetary investments in each asset for each ETF were determined, and the number of shares to purchase was computed. Throughout this process, the weights of the assets remained constant, and no rebalancing was performed after the initial estimation. This methodology provided a robust framework for constructing tracking portfolios, emphasizing minimizing tracking error while maintaining interpretability and stability.

3.3.2 Performance Evaluation

The evaluation of the sparse index-tracking portfolios relies on several performance measures designed to assess replication accuracy, tracking efficiency, systematic risk exposure, and overall portfolio performance. Each indicator has been selected for its specific relevance in replicating the performance of an index using a reduced subset of assets.

3.3.2.1 Tracking Error

The Tracking Error measures the standard deviation of the difference between the ETF's returns and the benchmark's (S&P 500) returns. As the primary metric for assessing replication accuracy, it quantifies how closely the ETF replicates the performance of the benchmark over the evaluation period (Grinold and Kahn 2000). Formally, the Tracking Error (TE) is defined as:

$$TE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_{p,t} - r_{b,t})^2},$$

where T represents the total number of observations, $r_{p,t}$ denotes the ETF return at time t, and $r_{b,t}$ denotes the benchmark return (S&P 500) at the same time.

3.3.2.2 Active Return

The Active Return measures the average return of the ETF in excess of the benchmark's return. It is calculated as the difference between the average return of the ETF ($\overline{r_p}$) and that of the S&P 500 ($\overline{r_b}$). This indicates how much the ETF has outperformed or underperformed the benchmark:

$$AR = \overline{r_p} - \overline{r_b}.$$

This formula directly compares the ETF's return to the benchmark, highlighting the performance differential.

3.3.2.3 Information Ratio

The Information Ratio assesses the efficiency of the ETF in generating excess returns relative to its Tracking Error. It is the ratio of the average active return to the Tracking Error, providing a measure of how effectively the ETF delivers risk-adjusted returns above the benchmark:

$$IR = \frac{\overline{r_p} - \overline{r_b}}{TE},$$

where $\overline{r_p}$ and $\overline{r_b}$ represent the mean returns of the ETF and S&P 500 benchmark, respectively. A higher Information Ratio indicates better risk-adjusted returns.

3.3.2.4 Correlation

The Correlation measures the strength and direction of the linear relationship between the ETF's returns and the benchmark's returns. The Pearson correlation coefficient (ρ) quantifies how closely the ETF's movements align with those of the S&P 500, with a value close to 1 indicating high correlation:

$$\rho = \frac{\sum_{t=1}^{T}(r_{p,t} - \overline{r_p})(r_{b,t} - \overline{r_b})}{\sqrt{\sum_{t=1}^{T}(r_{p,t} - \overline{r_p})^2} \cdot \sqrt{\sum_{t=1}^{T}(r_{b,t} - \overline{r_b})^2}},$$

as originally introduced by Pearson (Pearson 1895). This formula measures how closely the ETF moves in relation to the benchmark.

3.3.2.5 Beta

The Beta coefficient measures the ETF's sensitivity to systematic market movements. It is calculated by performing a linear regression of the ETF's returns on the returns of the S&P 500, following the Capital Asset Pricing Model (CAPM) (Sharpe 1964). The formula for Beta is:

$$\beta = \frac{\text{Cov}(r_p, r_b)}{\text{Var}(r_b)},$$

where:

- $\text{Cov}(r_p, r_b)$ is the covariance between the ETF's returns (r_p) and the benchmark's returns (r_b),
- $\text{Var}(r_b)$ is the variance of the benchmark's returns (r_b).

Beta indicates the degree to which the ETF's returns are expected to change in response to changes in the benchmark, providing insight into its exposure to systematic risk.

3.3.2.6 Jensen's Alpha

Jensen's Alpha measures the ETF's excess return relative to the expected return, based on its exposure to systematic market risk, as measured by Beta. Specifically, it compares the ETF's actual return to the return that would be expected based on its Beta and the benchmark's return.

In our analysis, we use the daily risk-free rate from 3-month Treasury bills (T-Bills) to calculate both Jensen's Alpha and Beta, as it is widely accepted in financial research as a benchmark for a risk-free investment (Fama and French 2004). The 3-month T-Bill is considered nearly risk-free because it is backed by the U.S. government, ensuring minimal default risk. This makes it an appropriate reference point for determining excess returns relative to the S&P 500 index.

The formula for Jensen's Alpha, α, is:

$$\alpha = \overline{r_p} - \left(r_f + \beta\left(\overline{r_b} - r_f\right)\right),$$

where $\overline{r_p}$ is the average return of the ETF, r_f is the risk-free rate, β is the ETF's Beta, and $\overline{r_b}$ is the average return of the S&P 500. A positive α indicates that the ETF has outperformed the expected return, adding value beyond its exposure to the S&P 500, while a negative α suggests underperformance relative to the benchmark.

4 Presentation of Data and Application Results

This chapter presents the empirical implementation of the previously introduced methodologies, combining an exploratory analysis of the data with the construction and evaluation of index-tracking portfolios. The analysis begins with a detailed examination of the S&P 500 index and its constituents, addressing data preprocessing steps such as non-trading days, missing values, and the transformation of prices into logarithmic returns. The statistical properties of these returns are then analyzed, including their distributional characteristics, stationarity, correlation structures, and autocorrelation patterns. These empirical features motivate the use of penalized regression methods for variable selection, implemented via Ridge, Lasso, Elastic Net, and Adaptive Lasso. The results of these procedures are presented below, both with and without preliminary screening using Distance Correlation Sure Independence Screening (DC-SIS). Based on the selected variables, index-tracking portfolios are constructed, and their performance is evaluated using multiple financial metrics (including tracking error, information ratio, active return, Jensen's alpha, beta, and correlation) to enable rigorous comparison with the benchmark index.

4.1 Exploratory and Descriptive Analysis

4.1.1 Data Retrieval

To ensure proper alignment between the asset-level dataset and the S&P 500 index data, all non-trading days were identified and removed from the asset dataset. These dates correspond to official U.S. stock market holidays when the New York Stock Exchange (NYSE) was closed and no trading activity occurred. Since these days are absent from the index dataset and not relevant for return-based analysis, they were excluded to maintain consistency in the time dimension across both datasets. The full list of these non-trading days, along with their corresponding holiday names, is provided in table A.1.

4.1.2 Missing Values

The reasons for missing data among certain S&P 500 companies are summarized in table A.2. Most of the missing data stem from structural corporate events such as recent Initial Public Offerings (e.g., Uber, Airbnb, Moderna), spin-offs from larger conglomerates (e.g., Carrier Global, Corteva, GE Healthcare Technologies), or corporate restructurings and mergers (e.g., Ingersoll Rand, DuPont de Nemours, Amcor). These events resulted in the creation of new entities or significant changes in reporting, leading to limited or unavailable historical financial data prior to key dates, typically between 2017 and 2023. In light of the absence of sufficient historical data, particularly for the years preceding these formation or listing events, these companies are excluded from the analysis to preserve the robustness and consistency of the study.

4.1.3 Stationarity Testing

The stationarity of the S&P 500 returns is assessed using complementary unit root and stationarity tests (see table A.3). Across all specifications of the Augmented Dickey-Fuller (ADF) test (trend and constant, constant only, and none), the test statistics are substantially lower than their corresponding 1% critical values, leading to a consistent rejection of the unit root null hypothesis at the 1% significance level. Similarly, the Phillips-Perron (PP) test confirms stationarity under both specifications considered (trend and constant, constant only), with highly negative test statistics relative to the 1% critical values.

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, which directly tests the null hypothesis of stationarity, yields test statistics well below the respective 1% critical values in both specifications, leading to a failure to reject stationarity at the 1% level.

Taken together, the combined evidence from the ADF, PP, and KPSS tests strongly supports the conclusion that the S&P 500 returns are stationary in levels. This validates the assumption of stationarity for subsequent analyses such as model estimation and variable selection.

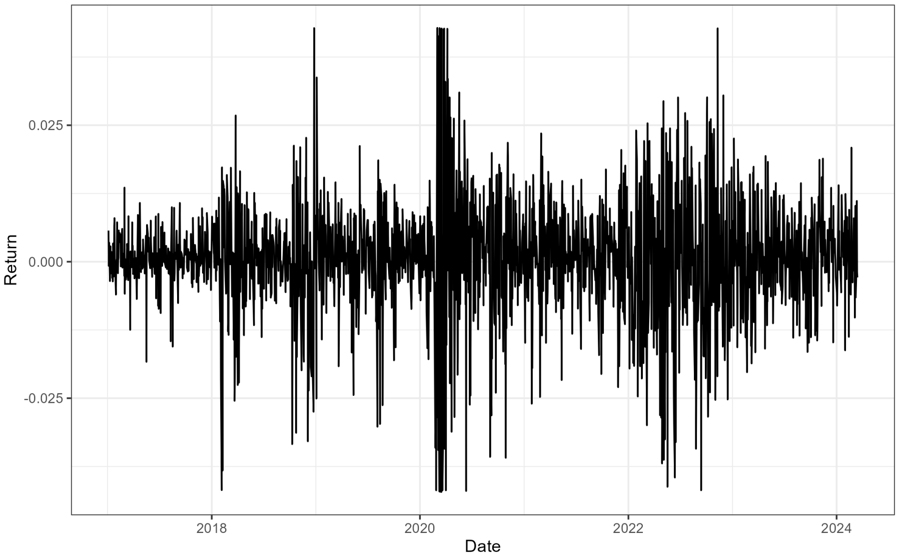
Regarding the individual constituents of the S&P 500 index, unit root and stationarity tests were applied to the returns of all 484 stocks (see table A.4). The results indicate that, under all specifications of the ADF test, all 484 stock return series are stationary at the 1% significance level. The PP test confirms stationarity for all 484 series under both the trend and constant, and constant only specifications. The KPSS test, which tests for stationarity, supports these results for the vast majority of series: 482 stocks out of 484 fail to reject stationarity under both the trend and constant, and constant only specifications at the 1% level.

Overall, the evidence suggests that both the aggregate index return and the vast majority of its constituents can be treated as stationary processes in levels, which supports the validity of applying stationary-based econometric models in the subsequent stages of the analysis.

4.1.4 Data Visualizations

To provide an initial overview of the return dynamics of the index and its constituents, this section presents time series plots of the daily returns for the S&P 500 and its individual components over the period from January 4, 2017 to March 14, 2024, highlighting key features of the data that motivate the subsequent empirical analysis.

Figure 4.1: Logarithmic Returns of the S&P 500



The time series plot of the S&P 500 returns in Figure 4.1 illustrates the daily return fluctuations of the index over the sample period. The returns generally fluctuate around zero, with no significant drift over short horizons, which is typical for equity index returns. Periods of elevated volatility are clearly observable, particularly around early 2020, likely corresponding to the onset of the COVID-19 pandemic, where the market experienced sharp swings both upwards and downwards. Following this period, volatility gradually stabilizes but remains punctuated by occasional spikes, indicating intermittent episodes of market stress. The return series exhibits key characteristics commonly observed in financial markets, such as volatility clustering and the occurrence of extreme returns that may reflect departures from normality.

Figure 4.2: Overlapping Returns of All Stocks

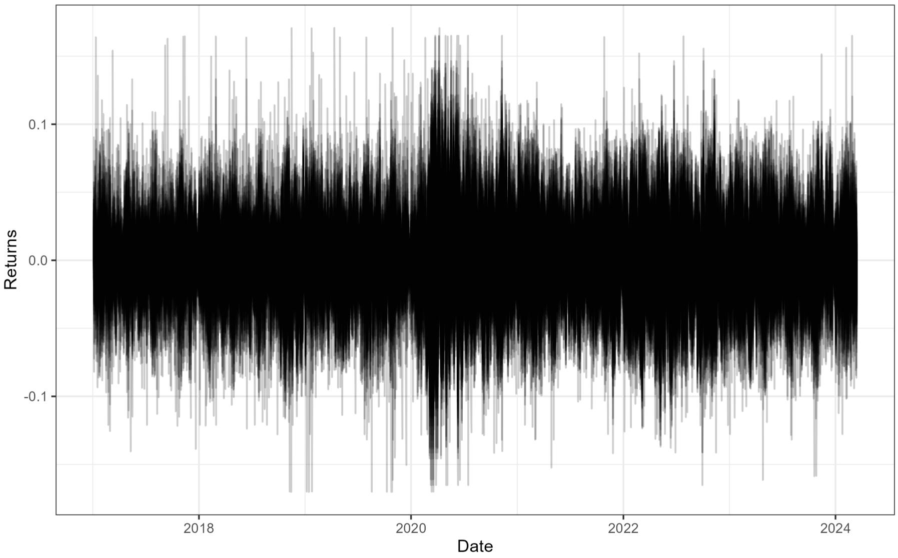


Figure 4.2 presents the overlapping returns of all S&P 500 constituent stocks over the same period. The chart reveals substantially greater dispersion in returns compared to the aggregate index, reflecting both higher variance and the presence of extreme observations among individual stocks. Significant cross-sectional variation is evident, with some stocks experiencing extreme positive or negative returns well beyond the range observed in the index. As with the index, a spike in volatility is observed around 2020, but with even more pronounced extremes among certain constituents. This pattern highlights the diversification effect inherent in index construction: while individual stocks are exposed to substantial specific risk, aggregating a large number of assets into the index effectively mitigates much of this stock-specific variability, resulting in lower overall volatility for the index. The high dispersion and specific risk among individual constituents also underscore the motivation for asset selection methodologies, such as penalized regression techniques, which aim to identify smaller subsets of stocks that can approximate the performance of the full index.

To complement the visual inspection, the next section presents descriptive statistics of both the index and its constituents to further characterize the distributional properties of returns.

4.1.5 Descriptive statistics

Table A.5 in the appendix summarizes the daily log returns of the S&P 500 index over the period from January 4, 2017 to March 14, 2024, covering 1,810 observations without missing data. The mean daily return is 0.000485, while the median, at 0.000732, is slightly higher, suggesting a mild asymmetry in the return distribution. In terms of variability, the standard deviation amounts to 0.010824, indicating moderate daily fluctuations in index performance. Furthermore, the minimum and maximum daily returns, recorded at -4.22% and 4.28% respectively, capture episodes of substantial market stress, most notably during the COVID-19 crisis. Additionally, the distribution exhibits slight negative skewness (-0.400), indicating a tendency for larger negative returns, while a kurtosis of 3.057 reflects fat tails and an increased likelihood of extreme outcomes relative to the normal distribution.

To complement the analysis of the aggregate index, figures A.1 and A.2 in the appendix provide further insights into the cross-sectional behavior of the index constituents. As shown in figure A.1, the vast majority of stocks exhibit mean daily returns tightly clustered around zero, typically ranging between -0.0005 and 0.0015. Only a few stocks display moderately higher average returns, and notably, no extreme outliers are observed over the sample period.

In contrast, figure A.2 highlights considerable variation in return volatility across individual stocks. While many stocks exhibit daily volatilities between 1% and 2% (i.e., 0.01 to 0.02), a significant proportion display higher volatility levels. In particular, several stocks exhibit volatilities exceeding 3%, with a few reaching above 4%. Thus, these results underscore the presence of substantial idiosyncratic risk at the individual stock level, which stands in sharp contrast to the lower volatility observed at the index level due to diversification effects.

Taken together, the descriptive statistics reveal several well-established empirical features of equity returns: low average returns, pronounced heterogeneity in volatility, mild negative skewness, and excess kurtosis. Consequently, these characteristics motivate the application of penalized regression methods that can effectively account for both systematic and idiosyncratic risks in constructing sparse index-tracking portfolios, as discussed in the following empirical analysis.

In addition to the summary statistics, the normality of the return distributions was examined. The Shapiro-Wilk test was applied to the daily returns of the S&P 500 index, yielding a test statistic of $W = 0.94293$ with a p-value below 2.2×10^{-16}, which strongly rejects the null hypothesis of normality at the 1% significance level. Similarly, univariate Shapiro-Wilk tests were conducted for each of the 484 individual constituent stocks. None of the return series satisfied the normality assumption at the 1% level. These results indicate that both the index and its constituent returns exhibit significant departures from normality, consistent with the skewness, excess kurtosis, and fat tails observed in the descriptive statistics. Nevertheless, as penalized regression methods are robust to non-normality (Hastie, Tibshirani, and Friedman 2009), these deviations do not affect the validity of the subsequent variable selection procedures.

4.1.6 Correlation Structure

Pairwise analysis among S&P 500 constituents revealed 795 asset pairs exhibiting strong association, with absolute Spearman rank correlations exceeding 0.7 and p-values below 0.5%. This substantial degree of monotonic comovement suggests notable interdependence among constituents, underscoring the need for penalized regression techniques capable of effectively addressing multicollinearity and complex variable relationships during the variable screening and asset selection process.

4.1.7 Autocorrelation Analysis

The analysis of the autocorrelation (ACF) and partial autocorrelation (PACF) functions of the daily returns of the S&P 500 and its constituent stocks reveals that the index returns exhibit limited serial dependence, with most autocorrelations falling within the 95% confidence bounds (figure A.3 and figure A.4). However, isolated significant spikes at lower lags suggest mild short-term autocorrelation that may reflect transient market frictions or delayed information diffusion. The PACF of the S&P 500 similarly indicates that any direct dependencies are concentrated at a few short lags. In contrast, the superimposed ACFs and PACFs of the individual constituent stocks display substantial heterogeneity (figure A.5 and figure A.6): while many stocks exhibit negligible or near-zero autocorrelation, others display pronounced autocorrelations and partial autocorrelations at various lags, indicating that certain assets may experience stronger short-term dependencies, potentially due to firm-specific factors, liquidity effects, or microstructure noise.

To formally assess the joint presence of autocorrelation, the Ljung-Box test was applied to the daily returns of both the S&P 500 index and its constituents. The test consistently indicates the presence of joint serial dependence, with p-values frequently falling below the 0.1 threshold across multiple series. This suggests that, even though many individual autocorrelations may appear small in magnitude, their combined contribution leads to statistically significant serial dependence. These findings highlight the importance of employing rolling-window cross-validation procedures that preserve the temporal ordering of the data, in order to mitigate the potential biases introduced by serial correlation when applying penalized regression techniques for index tracking and asset selection.

4.2 Variable Selection

This section presents the empirical results of the variable selection procedures implemented to construct sparse index-tracking portfolios. As described in Section 3.2, penalized regressions were applied within a rolling-origin cross-validation framework, both with and without preliminary variable screening via Distance Correlation Sure Independence Screening (DC-SIS). The analysis focuses on evaluating the sparsity, stability, and complexity achieved by each penalized method.

The hyperparameters selected through cross-validation for each model are reported in Table 4.1.

Table 4.1: Estimated Hyperparameters for Regularization Models

Model	α	λ
Ridge	0	0.000614
Lasso	1	0.00000102
Elastic Net ($\alpha = 0.5$)	0.5	0.00000260
Elastic Net	0.35	0.00000368
Adaptive Lasso	1	0.00156
Ridge (DC-SIS)	0	0.000774
Lasso (DC-SIS)	1	0.0000475
Elastic Net (DC-SIS) ($\alpha = 0.5$)	0.5	0.0000955
Elastic Net (DC-SIS)	0.4	0.000120
Adaptive Lasso (DC-SIS)	1	0.000614

As expected, Ridge regression corresponds to $\alpha = 0$, enforcing pure L_2 regularization, while Lasso corresponds to $\alpha = 1$, enforcing pure L_1 penalization that directly promotes sparsity. Elastic Net was estimated in two versions: one with a fixed $\alpha = 0.5$ and another with α optimized via grid search. The data-driven Elastic Net consistently selected intermediate α values (e.g., 0.35 and 0.4), reflecting the presence of correlated predictors and supporting a balance between shrinkage and sparsity. Adaptive Lasso, which incorporates data-driven weights, exhibited higher λ values, consistent with its bias-correction mechanism that aggressively penalizes smaller coefficients while allowing stronger predictors to remain influential.

The application of DC-SIS notably influenced these hyperparameter choices. When dimensionality was reduced via pre-screening, higher λ values were generally selected, indicating that less shrinkage was necessary after filtering out weakly associated variables. This highlights the complementary role of DC-SIS in simplifying the estimation task and enhancing the stability of penalized regressions.

Table 4.2 summarizes the number of variables selected by each model.

Table 4.2: Six Sets of Selected Stocks

Set	**Number of Stocks**	**Model**
1	223	Ridge
2	198	Lasso Elastic Net ($\alpha = 0.5$) Elastic Net
3	143	Adaptive Lasso
4	194	Ridge (DC-SIS)
5	165	Lasso (DC-SIS) Elastic Net (DC-SIS) ($\alpha = 0.5$) Elastic Net (DC-SIS)
6	138	Adaptive Lasso (DC-SIS)

As anticipated, Ridge regression retains the largest number of assets (223 without DC-SIS and 194 with DC-SIS), since its L_2 penalty typically shrinks coefficients continuously rather than eliminating them. However, under the non-negativity constraint imposed in this study, Ridge regression does reduce several coefficients exactly to zero when their unconstrained estimates would have been negative. In contrast, Lasso, Elastic Net, and Adaptive Lasso produce substantially more parsimonious solutions. Adaptive Lasso, in particular, selects the sparsest portfolios, retaining 143 variables without DC-SIS and 138 variables with DC-SIS. This aligns with the oracle property of Adaptive Lasso, which promotes consistent identification of relevant predictors while fully eliminating irrelevant ones.

The stability of these selections across rolling windows is illustrated in figures A.7 to A.11. Ridge regression (figure A.7) exhibits fluctuations in the number of non-zero coefficients, driven by its continuous shrinkage in conjunction with the non-negativity constraint. In contrast, Lasso (figure A.8), Elastic Net with both fixed and data-driven α values (figure A.9), and Adaptive Lasso (figure A.10) demonstrate considerably greater stability. Lasso and both Elastic Net variants consistently select approximately 198 variables, while Adaptive Lasso consistently retains around 143 variables. The introduction of DC-SIS further enhances this stability across all models, as seen in figure A.11, resulting in more consistent variable selection and the sparsest solution being achieved by Adaptive Lasso with DC-SIS (138 variables).

While several models select identical sets of variables, none yield identical estimated coefficients. Specifically, Lasso, Elastic Net ($\alpha = 0.5$), and Elastic Net (data-driven α) consistently select the same variables and constitute one group. Similarly, Lasso (DC-SIS), Elastic Net (DC-SIS) ($\alpha = 0.5$), and Elastic Net (DC-SIS) form a second group with identical selections. Consequently, six distinct sets of selected variables emerge across all models. Nevertheless, due to differences in penalty structure and parameter estimation, each model assigns unique coefficients to the variables it selects.

The full lists of selected assets, along with their corresponding estimated coefficients for each of the ten models, are available in the accompanying code repository. This ensures full transparency, reproducibility, and allows for detailed examination of the variable selection outcomes.

In summary, these findings underscore that combining DC-SIS with Adaptive Lasso delivers the most parsimonious, stable, and interpretable asset selection, while maintaining strong explanatory power for replicating index returns. The convergence of selected variables across rolling windows suggests that certain large-cap stocks consistently account for the majority of index fluctuations, making them well-suited for sparse ETF replication strategies.

4.3 Performance of Constructed Portfolios Relative to the Stock Index

This section analyzes the performance of the constructed index-tracking portfolios relative to the S&P 500 benchmark, following the methodology described in Section 3.3. The evaluation focuses primarily on tracking error, which is the central measure of replication accuracy, while also considering information ratio, active return, Jensen's alpha, correlation, and beta. The models are compared both with and without the application of Distance Correlation Sure Independence Screening (DC-SIS).

The first performance indicators considered are the tracking error and information ratio, which provide a direct assessment of replication accuracy and risk-adjusted efficiency. Table 4.3 reports these two key metrics for each constructed portfolio.

Table 4.3: Tracking Accuracy and Efficiency

Model	**Tracking Error**	**Information Ratio**
S&P500	NA	NA
Ridge	0.01997719	1.66349554
Lasso	0.01980500	1.76708437
Elastic Net ($\alpha = 0.5$)	0.01980904	1.76745688
Elastic Net	0.01980896	1.76762003
Adaptive Lasso	0.02005503	1.81964261
Ridge (DC-SIS)	0.03231044	0.06653447
Lasso (DC-SIS)	0.03140548	0.16244691
Elastic Net (DC-SIS) ($\alpha = 0.5$)	0.03148074	0.15994617
Elastic Net (DC-SIS)	0.03154830	0.15843034
Adaptive Lasso (DC-SIS)	0.03162787	0.12548465

The Lasso and Elastic Net models achieve the highest replication precision, both recording tracking errors around 0.0198. Adaptive Lasso follows closely at 0.0201, while Ridge achieves a slightly lower tracking error of 0.0200. These results highlight the advantage of sparsity-inducing methods that focus on selecting the most informative assets while excluding redundant variables, although Ridge's shrinkage approach allows for marginally lower tracking error by retaining a larger set of assets.

In terms of risk-adjusted replication efficiency, Adaptive Lasso stands out with the highest information ratio of 1.82, indicating superior excess return generation relative to its tracking error. Lasso and Elastic Net also perform well, with values near 1.77, while Ridge lags behind at 1.66. The ability of Adaptive Lasso to assign adaptive penalties allows it to better balance replication accuracy and return generation.

In contrast, the models incorporating DC-SIS experience a clear decline in performance, as tracking errors rise to approximately 0.031–0.032, while information ratios fall sharply below 0.17. Although DC-SIS achieves substantial dimensionality reduction, it does so at the cost of excluding important explanatory variables, reducing replication precision.

The evaluation continues with the analysis of the active returns and Jensen's alphas, which provide further insight into the portfolios' capacity to generate returns beyond simple replication. These results are presented in table 4.4.

Table 4.4: Raw and Risk-Adjusted Returns

Model	**Active Return**	**Jensen's Alpha**
S&P500	0.000000000	0.000000000
Ridge	0.033231969	0.029206682
Lasso	0.034997108	0.030855882
Elastic Net ($\alpha = 0.5$)	0.035011625	0.030862452
Elastic Net	0.035014722	0.030865601
Adaptive Lasso	0.036492986	0.032159435
Ridge (DC-SIS)	0.002149758	0.003130131
Lasso (DC-SIS)	0.005101723	0.005720242
Elastic Net (DC-SIS) ($\alpha = 0.5$)	0.005035223	0.005652718
Elastic Net (DC-SIS)	0.004998208	0.005600954
Adaptive Lasso (DC-SIS)	0.003968813	0.004839948

Adaptive Lasso generates the strongest return performance, delivering an active return of 3.65% and a Jensen's alpha of 3.22%. The Lasso and Elastic Net models follow closely, producing active returns of approximately 3.50% and alphas near 3.09%. Ridge performs more weakly, with an active return of 3.32% and an alpha of 2.92%.

For models estimated after DC-SIS filtering, active returns drop substantially, ranging from 0.21% to 0.51%, while Jensen's alphas lie between 0.31% and 0.57%. The reduced return performance reflects the limited capacity of these highly sparse models to capture the index's return-generating process after the aggressive elimination of explanatory variables.

Although Adaptive Lasso offers the highest returns, these gains are accompanied by a marginal increase in tracking error compared to the Lasso and Elastic Net.

Finally, table 4.5 reports the correlation and beta coefficients, which measure the extent of co-movement with the S&P 500 and sensitivity to market fluctuations.

Table 4.5: Systematic Risk Exposure

Model	**Correlation**	**Beta**
S&P500	1.0000000	1.0000000
Ridge	0.9941262	1.0310050
Lasso	0.9942692	1.0318873
Elastic Net ($\alpha = 0.5$)	0.9942692	1.0319485
Elastic Net	0.9942692	1.0319481
Adaptive Lasso	0.9941739	1.0333607
Ridge (DC-SIS)	0.9824978	0.9925234
Lasso (DC-SIS)	0.9835157	0.9952901
Elastic Net (DC-SIS) ($\alpha = 0.5$)	0.9834387	0.9952979
Elastic Net (DC-SIS)	0.9833724	0.9954113
Adaptive Lasso (DC-SIS)	0.9832339	0.9933483

All models demonstrate strong co-movement with the S&P 500 index. The non-DC-SIS models achieve correlations of approximately 0.994, while the DC-SIS models exhibit slightly lower correlations near 0.983, consistent with their higher tracking errors.

The beta coefficients indicate mild market overexposure for the non-DC-SIS models, with values slightly above one. For instance, Adaptive Lasso records a beta of 1.033. This modest overexposure contributes to higher active returns. In contrast, DC-SIS models display beta values close to unity, consistent with their lower return performance but not sufficient to offset their diminished replication accuracy.

When replication precision is prioritized, as is standard in index tracking, Lasso and Elastic Net deliver the most accurate replication of the S&P 500. Adaptive Lasso offers stronger active returns and superior risk-adjusted performance, though at the cost of a minor increase in tracking error. Ridge, despite achieving slightly lower tracking error than Adaptive Lasso, underperforms overall due to weaker return generation. The DC-SIS models, while achieving extreme sparsity, exhibit significantly higher tracking errors and lower returns, confirming that aggressive dimensionality reduction undermines replication quality. In practice, where minimizing tracking error is paramount, Lasso and Elastic Net are the most effective modeling approaches. Adaptive Lasso remains a strong alternative when modestly relaxing replication precision is acceptable in exchange for improved return efficiency.

4.4 Conclusion

Penalized regression techniques prove highly effective for constructing sparse index-tracking portfolios on the S&P 500. Variable selection plays a critical role, with Adaptive Lasso consistently identifying smaller and stable subsets of stocks that capture the essential dynamics of the index. Tracking error, as the primary measure of replication accuracy, shows that Lasso and Elastic Net achieve the lowest deviations from the benchmark, while Adaptive Lasso offers an optimal compromise between parsimony and performance. The application of DC-SIS further simplifies the models by reducing dimensionality but at the cost of slightly higher tracking errors, reflecting the inherent trade-off between simplicity and precision. Overall, these results highlight the practical relevance of penalized regression methods—particularly Adaptive Lasso and Elastic Net—for building efficient and realistic index-tracking portfolios that respect non-negativity constraints and practical investment considerations.

5 Conclusion & Discussion

This dissertation investigated the construction of sparse index-tracking portfolios using penalized regression techniques, applied to the S&P 500 index over the period 2017 to 2024. The objective was to replicate the index's performance while selecting only a limited subset of its constituents, thereby reducing transaction costs, improving portfolio manageability, and enhancing interpretability without compromising replication accuracy.

The empirical analysis demonstrated that penalized regression models, particularly Lasso, Elastic Net, and Adaptive Lasso, are highly effective for this purpose. Among these, Adaptive Lasso offered the most favorable trade-off between model sparsity and replication performance. Although Lasso and Elastic Net achieved slightly lower tracking errors, Adaptive Lasso delivered superior risk-adjusted performance, as evidenced by its highest Information Ratio. This reflects its ability to penalize weaker predictors while retaining key constituents that explain the majority of index variability.

Ridge regression achieved a marginally lower tracking error but required the inclusion of a substantially larger number of assets. While this allowed for slightly closer replication, it came at the cost of reduced portfolio simplicity and potentially higher transaction costs. In contrast, Adaptive Lasso achieved competitive replication performance while limiting the number of selected stocks to approximately 140, offering a more practical and cost-efficient approach for index replication.

The integration of Distance Correlation Sure Independence Screening (DC-SIS) further contributed to dimensionality reduction and improved stability in variable selection across rolling estimation windows. However, its application introduced a modest increase in tracking error, highlighting the inherent trade-off between model simplicity and replication precision. This suggests that variable screening can effectively support model stability but must be applied with care to avoid excluding important explanatory variables.

In addition to tracking error, complementary performance measures confirmed the robustness of the constructed portfolios. High correlations and betas close to unity demonstrated that the models successfully captured the index's systematic risk exposure. Furthermore, positive Jensen's alpha values indicated the potential for modest excess returns beyond simple replication, even after controlling for systematic risk.

One of the main challenges encountered during the research process was the need to acquire substantial new knowledge. At the beginning of this dissertation, I was not familiar with penalized regression methods or the DC-SIS approach. A significant part of the work was therefore dedicated to studying both the theoretical foundations and the practical implementation of these advanced techniques. Overcoming this initial learning curve proved essential for conducting a rigorous and comprehensive analysis.

Despite the encouraging results, several limitations of the study should be acknowledged. The analysis was restricted to the S&P 500 index and relied exclusively on historical return data, without incorporating macroeconomic indicators, firm-level fundamentals, or sector-specific factors that may enhance predictive accuracy. Moreover, the models were estimated under a static, non-rebalancing framework, while real-world index tracking typically requires periodic rebalancing to adjust for changes in index composition. Additionally, transaction costs, liquidity constraints, bid-ask spreads, and tax considerations were not explicitly incorporated, though they play a critical role in actual portfolio management.

These limitations suggest several promising directions for future research. Extending the methodology to other indices, international markets, or alternative asset classes would allow for testing its broader applicability. Incorporating additional explanatory variables such as firm fundamentals, sector exposures, or macroeconomic factors could further improve model robustness. Exploring dynamic rebalancing strategies would better reflect real-world trading environments, allowing for a more realistic assessment of turnover and implementation costs. Finally, complementing penalized regression with advanced machine learning methods, while ensuring interpretability, may yield further gains in predictive accuracy and practical relevance.

In conclusion, this dissertation demonstrates that penalized regression methods, particularly Adaptive Lasso and Elastic Net, offer powerful and practical tools for constructing sparse index-tracking portfolios. When combined with rigorous variable screening and robust cross-validation frameworks, these approaches provide asset managers with effective solutions to balance replication precision, portfolio simplicity, and operational efficiency in the design of passive investment products.

Bibliography

Boudt, Kris, Jeroen Cornelissen, and Christophe Croux (2013). “Robust portfolio optimization with systematic errors in mean and covariance estimates”. In: *Journal of Empirical Finance* 20, pp. 26–39.

Chen, Qian et al. (2022). “Time-weighted nonnegative lasso index-tracking model”. In: *North American Journal of Economics and Finance* 59, p. 101603. DOI: 10.1016/j.najef.2021.101603.

Chen, Qingyu and et al. (2022). “A Time-weighted Nonnegative LASSO Index Tracking Model under Market Constraints”. In: *North American Journal of Economics and Finance* 62, p. 101774.

Darné, Olivier (2024). *Machine Learning, Régressions Pénalisées et Sélection de Variables.* Course Notes, Université de Nantes.

Elton, Edwin J. et al. (2014). *Modern Portfolio Theory and Investment Analysis.* 9th. Hoboken, NJ: Wiley.

Fama, Eugene F. and Kenneth R. French (2004). “The Capital Asset Pricing Model: Theory and Evidence”. In: *The Journal of Economic Perspectives* 18.3, pp. 25–46.

Grinold, Richard C. and Ronald N. Kahn (2000). *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk.* New York: McGraw-Hill.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd. Springer.

Hoerl, Arthur E. and Robert W. Kennard (1970). “Ridge Regression: Biased Estimation for Nonorthogonal Problems”. In: *Technometrics* 12.1, pp. 55–67. DOI: 10.1080/00401706.1970.10488634.

Li, Xiang and Yuehan Yang (2021). “Index tracking via nonnegative minimax concave penalty”. In: *Statistics and Probability Letters* 170, p. 109009. DOI: 10.1016/j.spl.2020.109009.

Liu, Yonghui, Lin Li, and Hu Yang (2024). “Nonnegative group bridge and application in financial index tracking”. In: *Statistical Papers* 65, pp. 681–700. DOI: 10.1007/s00362-023-01463-3.

Pearson, Karl (1895). “Notes on Regression and Inheritance in the Case of Two Parents”. In: *Proceedings of the Royal Society of London* 58, pp. 240–242.

S&P Dow Jones Indices (2023). *S&P 500 Index Methodology.* https://www.spglobal.com/spdji/en/documents/methodologies/methodology-sp-500.pdf. S&P Global.

Sevi, Benoît (2024). “Séries temporelles univariées – Chapitre 2”. In: Course notes, M1 ECAP, Nantes Université.

Sharpe, William F. (1964). “Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk”. In: *The Journal of Finance* 19.3, pp. 425–442.

Shu, Haiwei, Lan Wu, and Yuehan Yang (2020). “Adaptive elastic net for high-dimensional index tracking”. In: *Quantitative Finance* 20.2, pp. 237–252. DOI: 10.1080/14697688.2019.1655814.

Shu, Lianjie, Fangquan Shi, and Guoliang Tian (2020). “High-dimensional index tracking based on the adaptive elastic net”. In: *Quantitative Finance* 20.9, pp. 1513–1530. DOI: 10.1080/14697688.2020.1737328.

Shu, Yiqian, Liuhui Zhang, and Ziye Yan (2020). “Sparse index tracking based on lasso-type optimization”. In: *Quantitative Finance* 20.4, pp. 653–670. DOI: 10.1080/14697688.2019.1654611.

Silva, Julio Cezar Soares and Adiel Teixeira de Almeida Filho (2023). “A Systematic Literature Review on Solution Approaches for the Index Tracking Problem in the Last Decade”. In: *IMA Journal of Management Mathematics* 34.2, pp. 224–248. DOI: 10.1093/imaman/dpad007. URL: https://doi.org/10.1093/imaman/dpad007.

Silva-Filho, Diego et al. (2023). “A multi-objective view of index tracking”. In: *Journal of Mathematical Modeling*. arXiv preprint arXiv:2303.14085. URL: https://arxiv.org/abs/2303.14085.

Székely, Gábor J., Maria L. Rizzo, and Nail K. Bakirov (2007). “Measuring and Testing Dependence by Correlation of Distances”. In: *The Annals of Statistics* 35.6, pp. 2769–2794. DOI: 10.1214/009053607000000505.

The Vanguard Group (2024). *Our History*. Accessed April 2025. URL: https://corporate.vanguard.com/content/corporatesite/us/en/corp/who-we-are/sets-us-apart/our-history.html.

Tibshirani, Robert (1996a). “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

— (1996b). “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Wu, Lan and Yuehan Yang (2014). “Nonnegative elastic net for high-dimensional regression models and its applications in index tracking”. In: *Applied Mathematics and Computation* 236, pp. 72–81. DOI: 10.1016/j.amc.2014.03.057.

Wu, Lan, Yuehan Yang, and Hanzhong Liu (2014a). “Nonnegative-lasso and application in index tracking”. In: *Computational Statistics & Data Analysis* 70, pp. 116–126.

— (2014b). “Nonnegative-lasso and application in index tracking”. In: *Computational Statistics and Data Analysis* 70, pp. 116–126. DOI: 10.1016/j.csda.2013.08.012.

Xia, Siwei, Yuehan Yang, and Hu Yang (2023). “High-dimensional sparse portfolio selection with nonnegative constraint”. In: *Applied Mathematics and Computation* 443, p. 127766. DOI: 10.1016/j.amc.2022.127766.

Yang, Yuehan and Lan Wu (2016). “Nonnegative adaptive lasso for ultra-high dimensional regression models and a two-stage method applied in financial modeling”. In: *Journal of Statistical Planning and Inference* 174, pp. 52–67. DOI: 10.1016/j.jspi.2016.01.011.

Yayi, Eric (2024). “Évaluation d’actifs - Chapitre 2: Efficience des marchés financiers”. In: Course notes, M1 ECAP, Nantes Université.

Zou, Hui (2006). “The Adaptive Lasso and Its Oracle Properties”. In: *Journal of the American Statistical Association* 101.476, pp. 1418–1429. DOI: 10.1198/016214506000000735.

Zou, Hui and Trevor Hastie (2005). “Regularization and Variable Selection via the Elastic Net”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

Appendices

Table A.1: U.S. Market Non-Trading Days (2017–2024)

Date	**Holiday**
2017-01-02	New Year's Day (observed)
2017-01-16	Martin Luther King Jr. Day
2017-02-20	Presidents' Day
2017-04-14	Good Friday
2017-05-29	Memorial Day
2017-07-04	Independence Day
2017-09-04	Labor Day
2017-11-23	Thanksgiving Day
2017-12-25	Christmas Day
2018-01-01	New Year's Day
2018-01-15	Martin Luther King Jr. Day
2018-02-19	Presidents' Day
2018-03-30	Good Friday
2018-05-28	Memorial Day
2018-07-04	Independence Day
2018-09-03	Labor Day
2018-11-22	Thanksgiving Day
2018-12-05	National Day of Mourning
2018-12-25	Christmas Day
2019-01-01	New Year's Day
2019-01-21	Martin Luther King Jr. Day
2019-02-18	Presidents' Day
2019-04-19	Good Friday
2019-05-27	Memorial Day
2019-07-04	Independence Day
2019-09-02	Labor Day
2019-11-28	Thanksgiving Day
2019-12-25	Christmas Day
2020-01-01	New Year's Day
2020-01-20	Martin Luther King Jr. Day
2020-02-17	Presidents' Day
2020-04-10	Good Friday
2020-05-25	Memorial Day
2020-07-03	Independence Day (observed)

Date	**Holiday**
2020-09-07	Labor Day
2020-11-26	Thanksgiving Day
2020-12-25	Christmas Day
2021-01-01	New Year's Day
2021-01-18	Martin Luther King Jr. Day
2021-02-15	Presidents' Day
2021-04-02	Good Friday
2021-05-31	Memorial Day
2021-07-05	Independence Day (observed)
2021-09-06	Labor Day
2021-11-25	Thanksgiving Day
2021-12-24	Christmas Day (observed)
2022-01-17	Martin Luther King Jr. Day
2022-02-21	Presidents' Day
2022-04-15	Good Friday
2022-05-30	Memorial Day
2022-06-20	Juneteenth (observed)
2022-07-04	Independence Day
2022-09-05	Labor Day
2022-11-24	Thanksgiving Day
2022-12-26	Christmas Day (observed)
2023-01-02	New Year's Day (observed)
2023-01-16	Martin Luther King Jr. Day
2023-02-20	Presidents' Day
2023-04-07	Good Friday
2023-05-29	Memorial Day
2023-06-19	Juneteenth
2023-07-04	Independence Day
2023-09-04	Labor Day
2023-11-23	Thanksgiving Day
2023-12-25	Christmas Day
2024-01-01	New Year's Day
2024-01-15	Martin Luther King Jr. Day
2024-02-19	Presidents' Day

Note: Holidays marked with "(observed)" indicate that the official holiday fell on a weekend and was observed by market closure on the nearest weekday, following standard U.S. financial market practice.

Table A.2: Reasons for Missing Data and Key Dates for S&P 500 Companies

Company	**Reason for Missing Data**	**Year**
Uber Technologies	Initial Public Offering (IPO)	2019
Ingersoll Rand	Data starts after 2017 restructuring and spin-merger	2017
Airbnb A	Initial Public Offering (IPO)	2020
DuPont de Nemours	Formed after DowDuPont breakup	2017
Carrier Global	Spin-off from United Technologies	2020
Constellation Energy	Spin-off from Exelon	2022
Dow Ord Shs	Spin-off from DowDuPont breakup	2019
GE Healthcare Technologies	Spin-off from General Electric	2023
Corteva	Spin-off from DowDuPont breakup	2019
Kenvue	Spin-off from Johnson & Johnson	2023
Moderna	Initial Public Offering (IPO)	2018
Otis Worldwide	Spin-off from United Technologies	2020
Vici Pptys	Spin-off from Caesars Entertainment	2018
Invitation Homes	Initial Public Offering (IPO)	2017
Veralto	Spin-off from Danaher Corporation	2023
Amcor	Listed on NYSE after Bemis acquisition	2019
Dayforce	IPO as Ceridian HCM, rebranded to Dayforce	2018, 2023
Fox A	Created after Disney acquired 21st Century Fox assets	2019
Fox B	Created after Disney acquired 21st Century Fox assets	2019

Table A.3: Stationarity Tests for S&P 500 Returns

ADF Test	**T+C**	**C**	**None**
Test Statistic	-30.7	-30.7	-30.6
Critical Value	-3.96	-3.43	-2.58
Stationarity	TRUE	TRUE	TRUE
PP Test	**T+C**	**C**	**-**
Test Statistic	-46.5	-46.5	-
Critical Value	-3.97	-3.44	-
Stationarity	TRUE	TRUE	-
KPSS Test	**T+C**	**C**	**-**
Test Statistic	0.0431	0.0428	-
Critical Value	0.216	0.739	-
Stationarity	TRUE	TRUE	-

Note: T+C = trend and constant; C = constant only; None = no trend or constant.

Table A.4: Stationarity Tests for Stocks Returns

ADF Test	**T+C**	**C**	**None**
Stationarity	484	484	484
PP Test	**T+C**	**C**	**-**
Stationarity	484	484	-
KPSS Test	**T+C**	**C**	**-**
Stationarity	482	482	-

Note: T+C = trend and constant; C = constant only; None = no trend or constant.

Table A.5: Summary Statistics of S&P 500 Returns

Statistic	**Value**
Number of Observations (nobs)	1810
Missing Values (NAs)	0
Minimum	-0.042184
Maximum	0.042838
1st Quartile	-0.003842
3rd Quartile	0.006177
Mean	0.000485
Median	0.000732
Sum	0.878130
Standard Error of Mean (SE Mean)	0.000254
Lower 95% CI Mean (LCL Mean)	-0.000014
Upper 95% CI Mean (UCL Mean)	0.000984
Variance	0.000117
Standard Deviation (Stdev)	0.010824
Skewness	-0.400422
Kurtosis	3.056640

Figure A.1: Mean of Daily Log Returns for S&P 500 Constituents

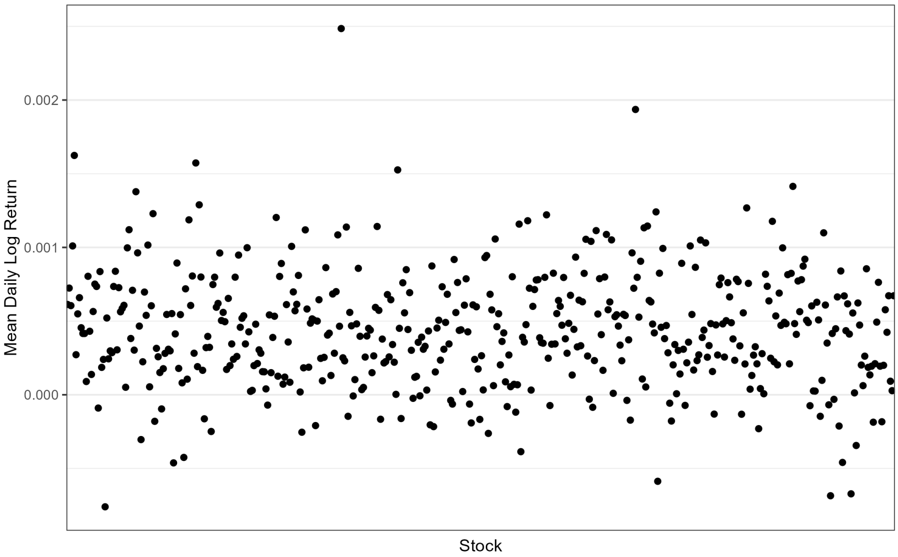


Figure A.2: Volatility of Daily Log Returns for S&P 500 Constituents

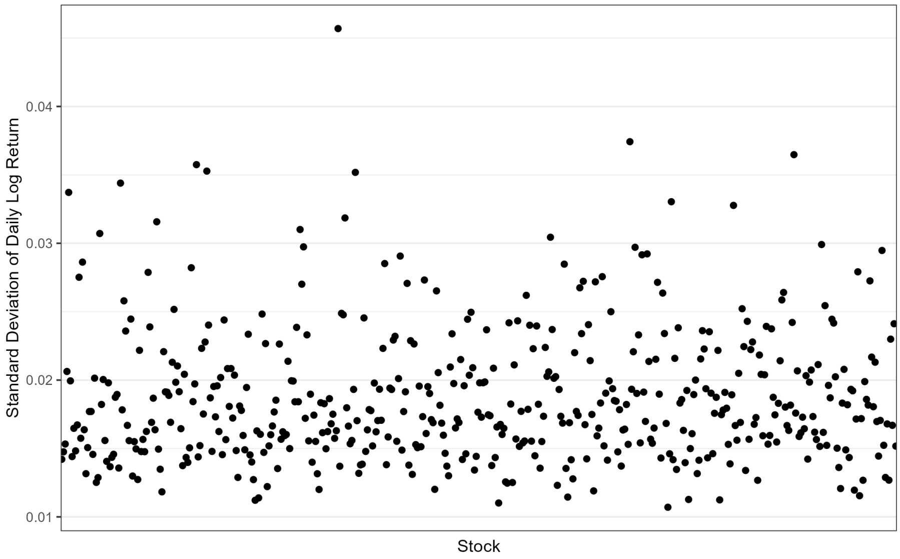


Figure A.3: ACF of S&P 500 Return

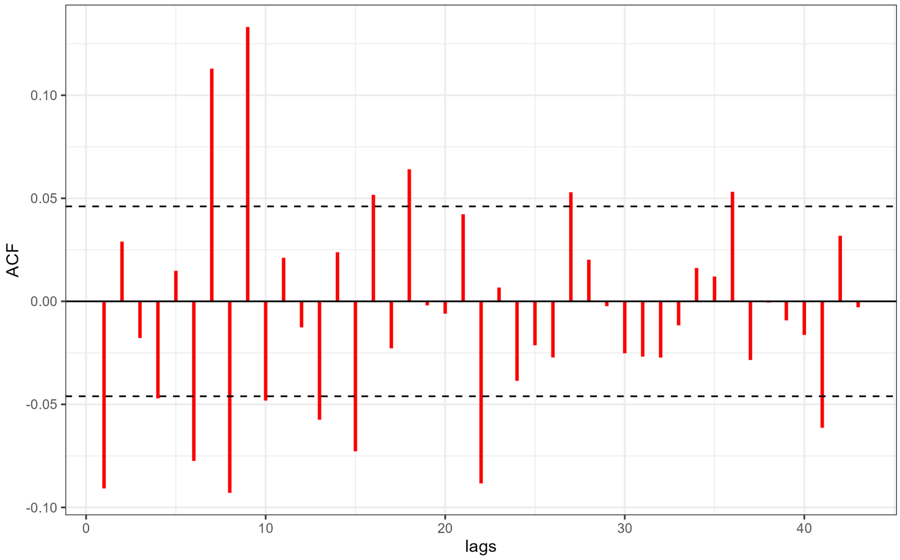


Figure A.4: PACF of S&P 500 Return

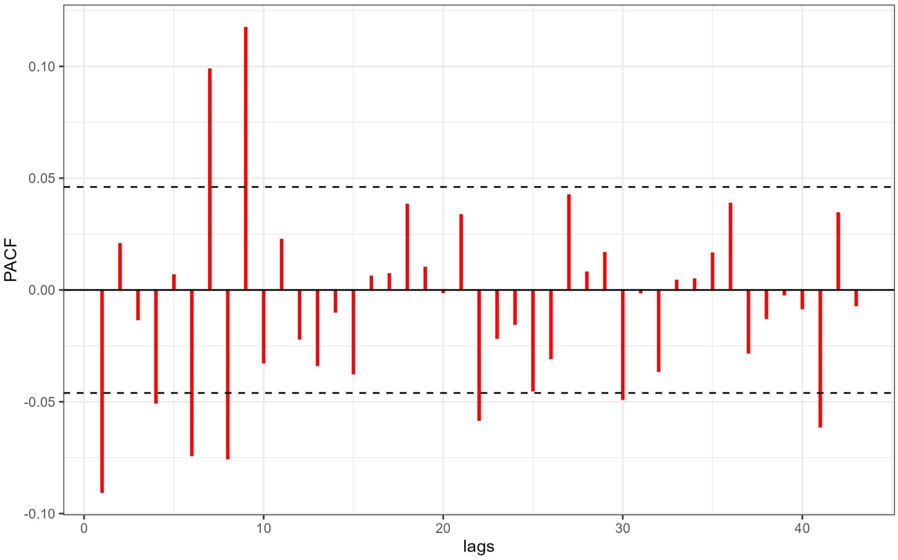


Figure A.5: Superimposed ACFs of All Stock Returns

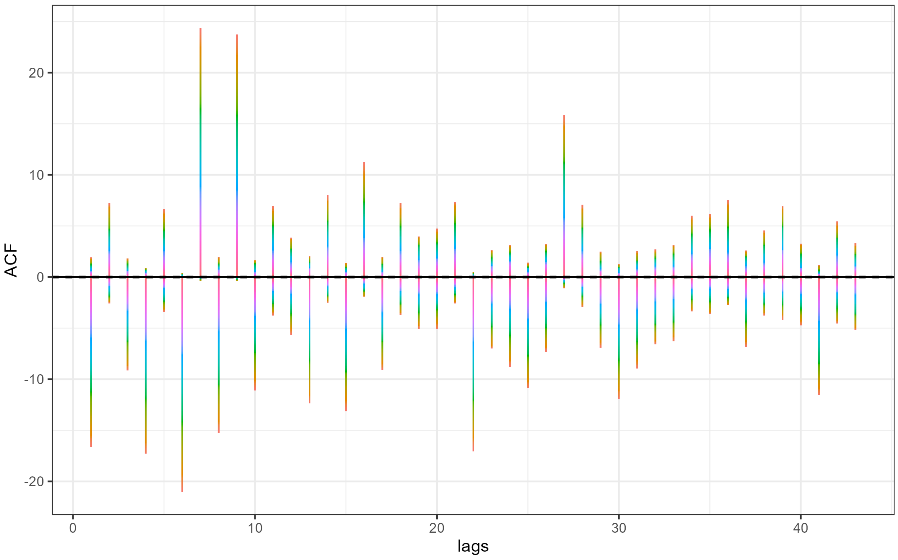


Figure A.6: Superimposed PACFs of All Stock Returns

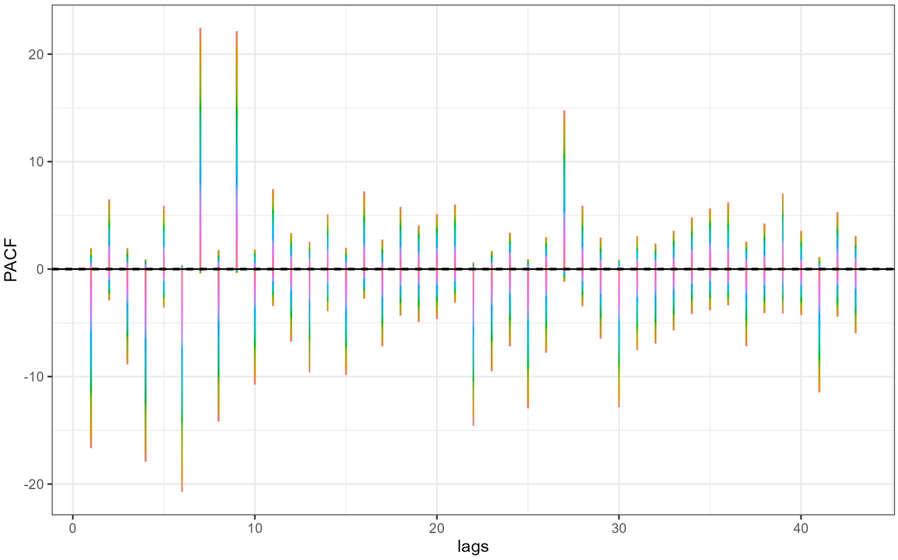


Figure A.7: Cross-Validation RMSE Curve for Ridge Regression Models

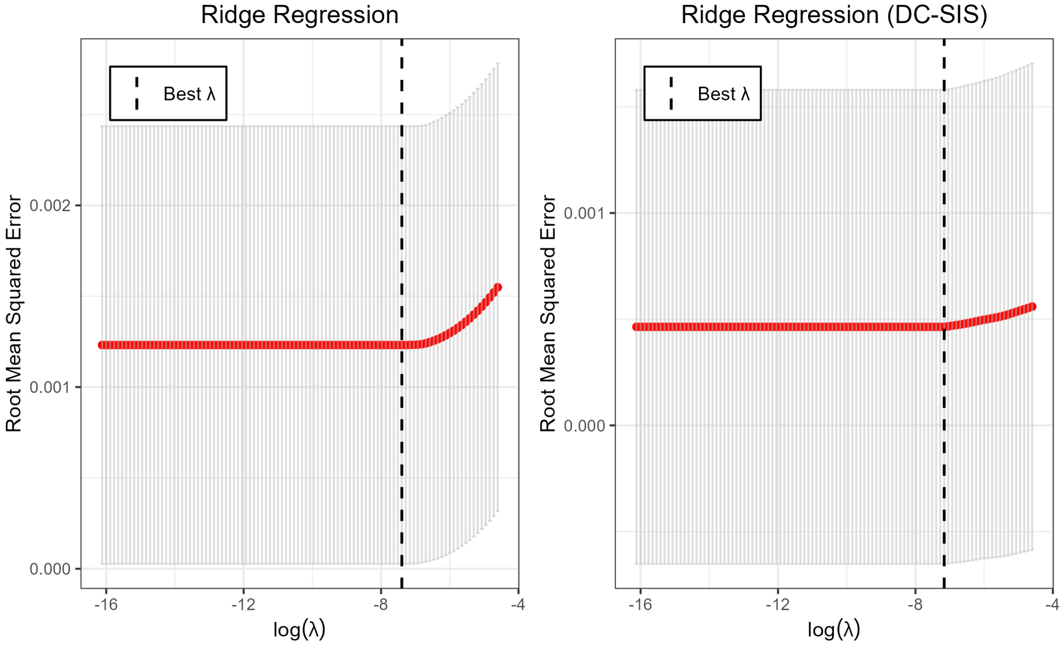


Figure A.8: Cross-Validation RMSE Curve for Lasso Regression Models

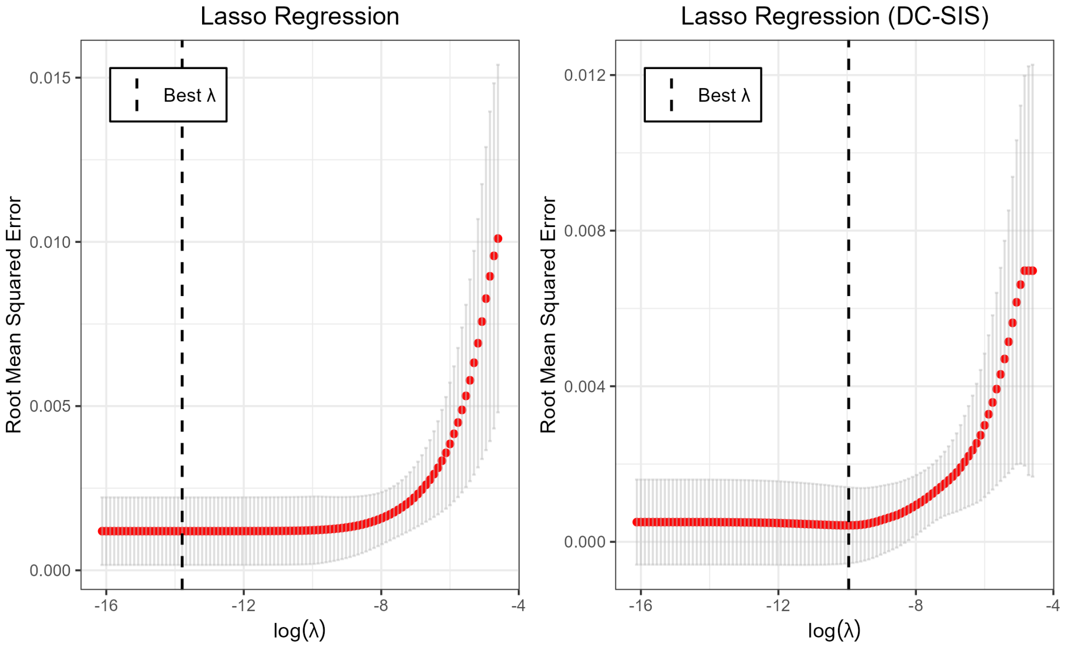


Figure A.9: Cross-Validation RMSE Curve for Elastic Net Regression Models ($\alpha = 0.5$)

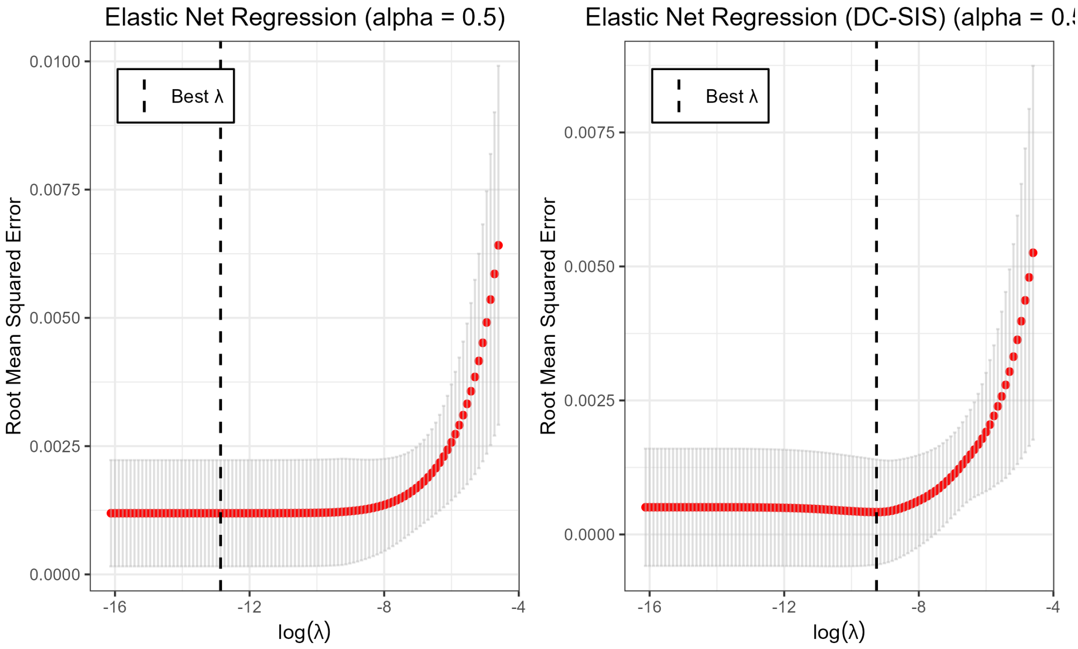


Figure A.10: Cross-Validation RMSE Curve for Elastic Net Regression Models

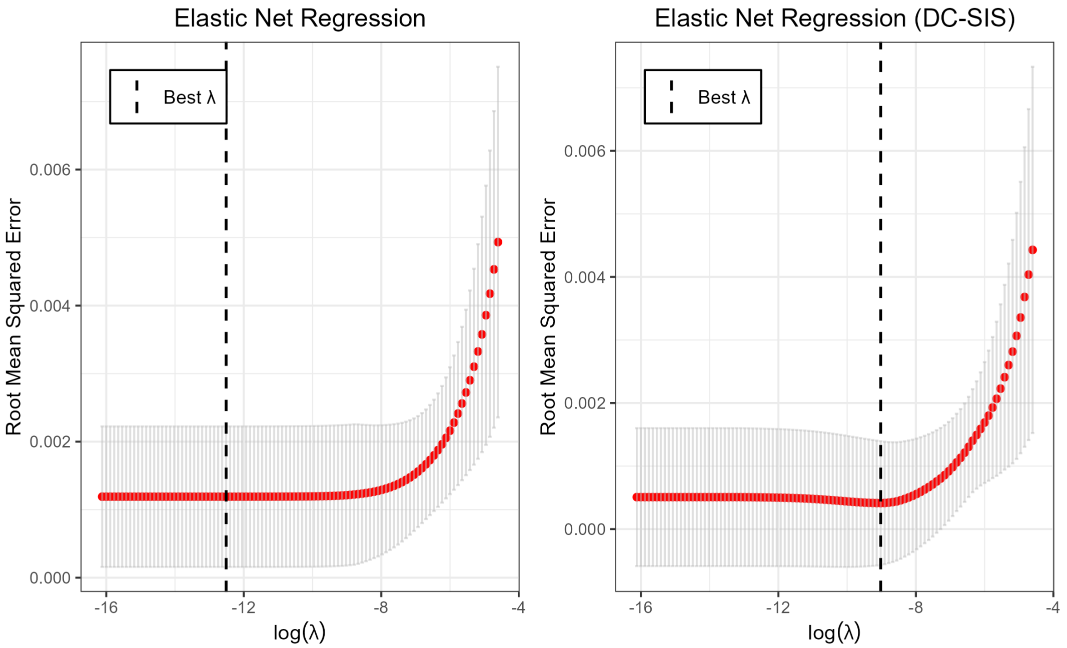


Figure A.11: Cross-Validation RMSE Curve for Adaptive Lasso Regression Models

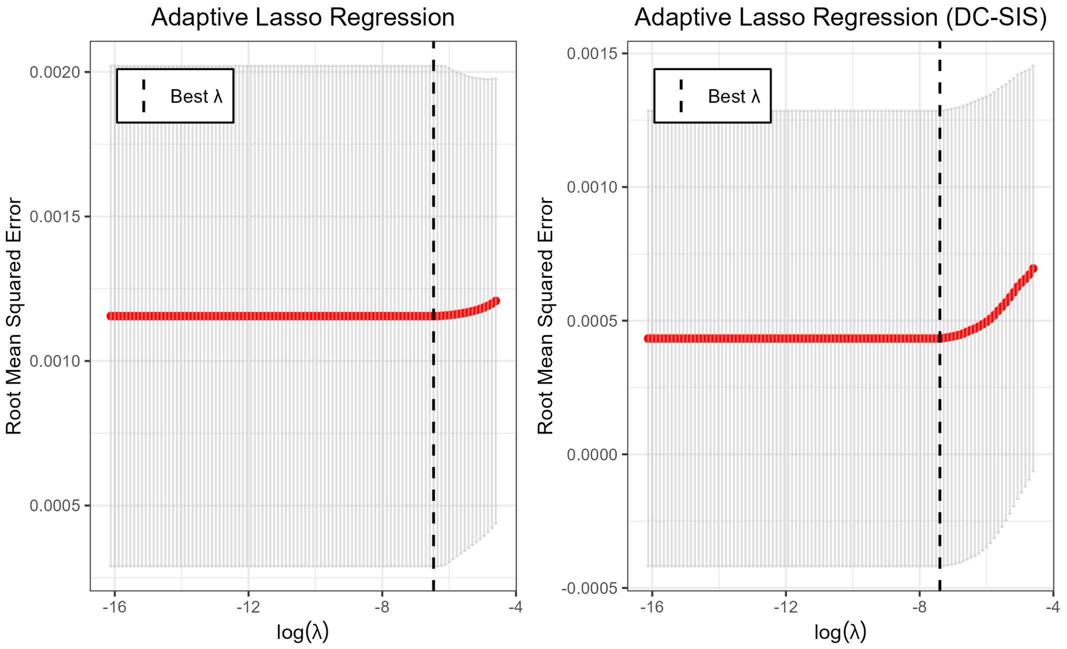


Table of Contents