UNIVERSITY OF TURKU

Data-Driven Equity Selection with PCA and Minimum Variance Optimization

Accounting and Finance

Master's thesis

Author:

Samuli Valtonen

17.12.2025

Turku

Master's thesis

Subject: Finance
Author: Samuli Valtonen
Title: Data-Driven Equity Selection with PCA and Minimum Variance Optimization
Supervisor(s): Professor Jan Pfister & Postdoctoral researcher Annika Hasselblad
Number of pages: 81 pages + appendices 1 page
Date: 17.12.2025

The topic of this thesis is the development of a quantitative equity strategy that combines Principal Component Analysis (PCA) with a minimum-variance portfolio. The purpose of the study is to examine whether a strategy based on rolling PCA-based stock selection and variance-minimizing allocation can provide a more efficient and adaptive alternative to a traditional market-capitalization-weighted index. The thesis also investigates how PCA-derived factors can be transformed into scoring rules that guide investment decisions and how the strategy performs in different market conditions.

The literature review presents the theoretical foundations of financial markets, including the efficient market hypothesis, the random walk model, pricing theories and behavioral finance. The theoretical framework further discusses the basics of principal component analysis and Markowitz's portfolio theory in more detail, as well as earlier research on the use of PCA in stock selection and risk management.

The empirical part of the study is conducted using daily data on S&P 500 constituents for the period 2010–2024. The strategy applies rolling one-year windows, from which PCA-based selection scores are constructed and the weights of a minimum-variance portfolio are solved. The performance of the strategy is evaluated over the full sample period using several return and risk measures. The results show that, over the sample period, the PCA-based strategy clearly outperforms the S&P 500 index in terms of cumulative return, achieves higher risk-adjusted returns and exhibits smaller maximum drawdowns. However, the excess return is clearly regime dependent and concentrated in certain periods, and rolling alpha and the Sharpe ratio are not consistently positive. Therefore, the results cannot be generalized to other markets or time periods without further evidence.

Key words: principal component analysis, minimum variance optimization, Markowitz portfolio theory, quantitative finance.

Table of contents

PICTURES

TABLES

1 Introduction

Financial markets are inherently complex, influenced by macroeconomic conditions, investor behavior, and structural inefficiencies. Investors have long sought strategies that generate consistent excess returns, yet achieving this goal remains an elusive challenge. Markets tend to self-correct, with profitable opportunities becoming arbitraged away over time, making it difficult for any single approach to maintain long-term outperformance (Joubert, 2005).

Traditional investment strategies, such as value investing, growth investing, and index investing, have been widely used by institutional and retail investors. The value investing framework, pioneered by Graham and Dodd (1934) and later refined by Fama and French (1993), suggests that stocks trading below their intrinsic value tend to outperform in the long run. Growth investing, on the other hand, focuses on companies with strong revenue expansion and innovation, even if their valuations appear stretched. Index investing, popularized by Bogle (1975), emphasizes passive exposure to broad market indices, offering investors a low-cost, diversified approach. While these strategies have yielded significant returns over decades, they have also exhibited periods of prolonged underperformance, particularly during shifts in macroeconomic conditions and increasing institutional competition (Shukla, 2004).

In response, quantitative investment strategies have gained prominence, offering data-driven methods to exploit inefficiencies in financial markets. Techniques such as factor investing, momentum trading, and regime-switching models attempt to capture predictable patterns in asset prices (Dal Pra et al., 2018). Factor investing, for example, seeks to isolate systematic return drivers such as size, value, momentum, and low volatility, which have been empirically shown to provide risk-adjusted excess returns (Jegadeesh & Titman, 1993). Momentum strategies, which capitalize on stocks exhibiting strong past performance, have been particularly successful in certain market conditions. However, these strategies are not immune to diminishing returns—as factors become widely recognized and arbitraged, their excess return potential often erodes (Malaj & Malaj, 2016).

Moreover, portfolio optimization remains a fundamental yet challenging problem in finance. Traditional mean-variance models, introduced by Markowitz (1952), aim to construct portfolios that optimize expected return for a given level of risk. However, they rely on historical estimates of returns and covariances, which may not adequately capture evolving market conditions. Furthermore, these models struggle to account for variations in risk exposure across different economies and asset classes, particularly in a globalized financial environment (Norland & Wilford, 2002). These challenges raise a critical question: how can investment strategies be designed to remain adaptive and robust over time?

One promising solution lies in the integration of Principal Component Analysis (PCA) with Markowitz's portfolio optimization. PCA, a statistical technique widely used in data science and econometrics, offers a dynamic approach to factor selection, extracting key return-driving components from financial datasets without assuming static factor persistence. By reducing dimensionality and isolating principal sources of variance, PCA can enhance portfolio construction by identifying the most influential risk drivers in real-time. (Jolliffe, 2002.) When combined with minimum variance portfolio optimization, PCA-based strategies may help investors construct resilient portfolios that prioritize risk minimization while maintaining exposure to high-return factors.

Over the past decade, PCA has gained traction in finance as a powerful tool for stock selection and market prediction. Researchers have demonstrated its ability to reduce data dimensionality, extract hidden patterns, and enhance predictive modeling for stock prices and returns. PCA has been widely applied in factor modelling, portfolio construction, and risk management, improving portfolio efficiency, risk estimation, and covariance structure analysis. (Tan, 2012.)

Empirical applications of PCA in equity markets point to three broad themes. First, PCA-based methods can compress large universes of correlated stocks into compact portfolios that still capture the main index movements. Second, projecting returns onto a low dimensional component space can improve the conditioning of forecasts and covariance estimates used in portfolio construction. Third, PCA-based factor and covariance models provide a practical way to implement risk first allocation in high dimensional settings. Together, these findings suggest that PCA is not only a dimensionality reduction technique but also a flexible tool for turning noisy return data into investable signals and risk estimates. The detailed evidence on these applications is reviewed in Chapter 2.

Despite this progress, many existing PCA-based strategies either focus on narrow market segments, rely on static estimation windows or use complex modelling choices such as neural networks or latent regime models that are difficult to implement in practice. Less is known about how a simple and transparent combination of rolling PCA-based selection and minimum variance allocation performs on a broad liquid equity index over a long sample. This gap motivates the present thesis. The study examines whether a data-driven PCA-based selection rule, combined with Markowitz's minimum variance portfolio, can deliver robust risk-adjusted performance and effective risk control relative to a market capitalization weighted benchmark.

1.1 Purpose of the study

This study aims to explore whether PCA-enhanced portfolio construction can offer a viable alternative to traditional static factor models. By leveraging data-driven asset selection and robust portfolio allocation methods, this research contributes to the ongoing discourse on quantitative investment strategies, providing a framework designed to adapt to evolving market conditions while mitigating factor decay.

To achieve this, the study integrates Principal Component Analysis (PCA) and Markowitz's portfolio optimization into a unified investment framework. Unlike traditional factor models that rely on predefined characteristics, PCA enables dynamically extract key return-driving factors from market data without assuming their persistence. This adaptability allows for real-time identification of shifting market structures, potentially reducing the risk of factor obsolescence. Additionally, combining PCA with Markowitz's minimum variance optimization might ensure that selected assets are weighted in a way that prioritizes risk minimization while maintaining exposure to the most relevant systematic return drivers.

This research aims to contribute to the development of adaptive, data-driven investment strategies that can withstand changing market conditions. By providing a dynamic alternative to traditional factor-based approaches, this study seeks to offer investors a more resilient framework for portfolio construction and risk management.

The proposed methodology is particularly well-suited for rational investors seeking to maximize returns while effectively controlling risk. PCA might help isolate systematic risk factors, making it an effective tool for investment decision-making, while Markowitz's portfolio theory provides a mathematically sound foundation for optimizing asset allocation. This combination could be particularly relevant in volatile markets, where uncertainty and asset correlations fluctuate rapidly. By integrating these two techniques, the developed strategy aims to help investors manage risks more effectively and make well-informed decisions, particularly when handling large and complex financial datasets.

1.2 Research questions and objectives

The primary objective of this thesis is to develop and evaluate an investment strategy that integrates PCA with Markowitz's minimum variance portfolio optimization. The aim is to assess whether such a strategy can provide a more adaptive, risk-aware, and effective framework for navigating dynamic financial markets compared to traditional benchmark-based approaches.

This thesis is guided by the following primary research question:

Q1: *Does combining PCA-based factor analysis with Markowitz's portfolio theory enable the construction of an efficient and flexible investment strategy?*

To support this main question, the following subquestions are posed:

Q2: *How can the key factors identified through PCA be utilized in investment allocation?*

Q3: *How well does the proposed strategy perform across different economic conditions based on historical backtests?*

This research contributes to both practice and academic literature. For investors, the study presents a systematic approach that integrates modern statistical techniques with classical portfolio theory. This strategy aims to support better-informed decision-making, especially under conditions of heightened market uncertainty and evolving asset correlations. Additionally, it offers a framework for constructing portfolios that adapt dynamically to changing market structures.

From an academic perspective, the research expands understanding of the integration of PCA and Markowitz's theory into investment strategies. It offers insights into analyzing return dynamics and improving risk management. Through historical testing, empirical evidence of the strategy's effectiveness is generated, which could serve as a foundation for further studies, such as the interaction between systematic and idiosyncratic factors in markets. Moreover, the research demonstrates how the analysis of large datasets can simplify complex decision-making processes.

Based on these research questions and previous literature, three hypotheses are formed, which are presented later in this work.

1.3 Significance and contribution

As said earlier, financial markets are increasingly shaped by complexity, volatility, and rapid shifts in investor sentiment and macroeconomic signals. Traditional investment strategies, whether based on fundamental valuation, historical factors, or static allocation models, often fail to adapt to these changing dynamics. In this environment, investors face a dual challenge: capturing meaningful return opportunities while simultaneously managing risk in real time.

This thesis is motivated by the belief that modern portfolio construction should be both data-driven and adaptive. PCA offers a powerful yet underutilized tool for uncovering latent return drivers within noisy financial data. Markowitz's portfolio theory, on the other hand, provides a mathematically rigorous foundation for minimizing risk through optimal diversification. By combining these two methodologies, this research proposes a strategy that systematically adapts to changing market conditions, rather than relying on fixed assumptions about asset behavior or factor persistence. Compared to previous studies, the advantage of this study is its simplicity and transparency. The design uses a small set of components: covariance-based PCA with a limited number of principal components, a long only minimum variance allocation, and simple OLS regression for stock selection. The approach omits macro predictors, regime switches, and deep learning. The aim is a simple and reproducible protocol that stays adaptive through frequent re-estimation and selection.

The significance of this research lies in its practical relevance and methodological integration. For investors, the strategy offers a way to construct portfolios that respond to the actual structure of return correlations, not assumptions or legacy factor models. This is particularly important in today's market, where correlations are time-varying, and macro shocks can quickly render static strategies ineffective. The framework developed here allows for frequent updates, enabling the strategy to remain responsive and robust.

Academically, this work contributes to the evolving discussion around hybrid quantitative strategies, where statistical pattern recognition methods like PCA are used not just for forecasting or dimensionality reduction, but as core tools in asset selection and allocation. (Cao & Wang, 2020; Bufalo et al., 2025; Sun et al., 2020). It builds on existing theory but adapts it to modern market conditions, emphasizing rolling analysis, dynamic scoring, and empirical validation through historical testing. The results may serve as a foundation for further research into machine learning enhanced portfolio models, adaptive factor timing, or data-driven risk regimes.

Ultimately, this research aims to bridge the gap between theoretical elegance and real-world applicability. It responds to a central need in finance today: building strategies that are not only grounded in proven theory but designed to evolve with the market itself.

2 Financial theory and literature review

This chapter introduces the classical financial theories that form the conceptual foundation for this thesis. These theories explain how markets should process information, how asset prices evolve, and how investors are expected to behave under rational assumptions. Understanding these models is essential for evaluating the motivation behind the investment strategy developed in later chapters.

The chapter begins with the Efficient Market Hypothesis (EMH) and Random Walk Theory, which describe the informational efficiency of markets and the unpredictability of asset prices. It then reviews key asset pricing models, including Modern Portfolio Theory (MPT), the Capital Asset Pricing Model (CAPM), and Arbitrage Pricing Theory (APT). These models offer structured ways to assess return expectations, risk, and diversification — all of which are relevant to the construction of a systematic investment strategy.

These classical frameworks also serve as reference points against which the thesis strategy is positioned. While they provide a useful baseline, the proposed approach in this thesis seeks to address some of their limitations through a data-driven and adaptive implementation, discussed in the next chapter.

2.1 Efficient Market Hypothesis

The Efficient Market Hypothesis (EMH) is a central concept in financial economics. It posits that financial markets are informationally efficient, meaning that asset prices at any given time fully reflect all available information (Fama, 1970). Under this assumption, it is impossible to consistently achieve returns that exceed average market returns on a risk-adjusted basis, since new information is quickly incorporated into prices.

EMH is commonly divided into three forms: weak, semi-strong, and strong. The weak form suggests that current prices incorporate all historical price and volume data. According to this view, technical analysis is ineffective because patterns in past prices cannot be used to predict future movements. (Slezak, 2003) The semi-strong form asserts that all publicly available information, including earnings announcements, economic reports, and company news is already reflected in asset prices. As a result, fundamental analysis cannot generate consistent excess returns. The strong form goes further, claiming that all information, both public and private (including insider knowledge), is instantly reflected in prices. This implies that even individuals with privileged access to information cannot achieve superior performance. (Fama, 1970.)

Each form of EMH presents progressively stronger claims about the extent to which information is priced into markets. While the weak and semi-strong forms have received partial empirical support, the strong form has been widely criticized and largely rejected in empirical studies.

The implications of EMH are profound for investment strategies. If markets are truly efficient, then active portfolio management becomes futile, and passive investing is optimal. However, numerous empirical studies and market anomalies have raised questions about the absolute validity of EMH. Behavioral biases, delayed information diffusion, and structural inefficiencies have all been cited as reasons why markets may deviate from perfect efficiency (Shleifer, 2000).

Malkiel (1973) argued that even though short-term anomalies may occur, they are difficult to exploit consistently, reinforcing the general principle that markets are broadly efficient over time. Jensen (1968) provided empirical support for EMH by showing that mutual fund managers, on average, fail to outperform the market after fees and expenses. In contrast, Grossman and Stiglitz (1980) challenged the strong form of market efficiency by arguing that if all information were instantly reflected in prices, there would be no incentive for investors to seek or trade on new information, thereby making perfectly efficient markets theoretically impossible. Furthermore, Rossi and Gunardi (2018) and Slezak (2003) have both highlighted the role of information asymmetry and structural frictions in limiting efficiency in emerging and developed markets alike.

Despite these critiques, EMH remains a foundational theory in finance. It serves as a theoretical benchmark for evaluating market anomalies and for assessing whether a given investment strategy, such as the one proposed in this thesis, provides value beyond what an efficient market would allow.

2.2 Random Walk Theory

Closely related to EMH is the Random Walk Theory, which suggests that stock prices move randomly and changes in asset prices are independent and identically distributed over time. In essence, this theory argues that future price movements cannot be predicted based on past information, since price changes follow a stochastic process akin to a random walk (Malkiel, 1973; Kendall, 1953).

Kendall (1953) was among the first to show empirically that price changes in financial markets behave in a non-systematic and unpredictable manner. His findings laid the groundwork for what later became known as Random Walk Theory. According to Bodie et al. (2024), the Efficient Market Hypothesis implies that security prices behave as though they follow a random walk process, meaning that changes in prices are unpredictable and do not follow discernible patterns. This randomness is not due to irrational behavior but to the fact that rational investors quickly incorporate new information into prices before others have time to act on it.

The theory holds that successive price changes are uncorrelated and that the direction of any future change is as likely to be up as it is to be down. In other words, according to the theory, it is impossible to predict that if the stock price rises the previous day, it will also rise the following day (Chitenderu et al., 2014). There is a strong conceptual link between the Efficient Market Hypothesis and the Random Walk Theory. Both suggest that asset prices reflect all available information and that forecasting future returns based solely on past data is ineffective. Bodie et al. (2024) emphasize that price changes being unpredictable does not imply irrationality, rather, it reflects rational investors reacting promptly to new information.

However, several empirical studies have challenged the strict validity of the random walk hypothesis. For instance, Lo and MacKinlay (1988) found evidence of short-term return predictability in some markets, suggesting that prices may deviate from a purely random process. Disanaike (1997) also observed anomalies such as overreaction to market events, implying that irrational behavior may occasionally distort price dynamics. Additionally, market efficiency and the presence of random walk behavior may vary by region, as tighter regulation and better information infrastructure in developed markets are thought to enhance randomness in price formation. Such empirical challenges have motivated the exploration of alternative, data-driven strategies capable of capturing structure in return dynamics—particularly when traditional models fall short.

Samuelson (1965) provided the formal theoretical foundation for this idea by proving that if prices are properly anticipated based on all available information, they must fluctuate randomly. His martingale-based framework reinforces the view that returns cannot be systematically predicted and that any attempt to forecast prices based on past trends will fail in an efficient market.

In the context of this thesis, Random Walk Theory supports the motivation for adopting adaptive, data-driven approaches. Understanding the limits of predictability in price movements underscores the need for strategies that respond dynamically to changing market conditions. The thesis explores whether PCA, as a flexible and non-parametric method, can extract latent return-driving structures from market data in real time and translate them into more robust portfolio allocations than static models. If past prices do not contain persistent predictive power, then static models may underperform. Instead, tools like PCA, which analyze return structures without relying on historical trends, may provide more robust mechanisms for systematic asset allocation.

2.3 Models of stock price formation

This chapter reviews the central theoretical models that explain how stock returns are generated in financial markets. While often referred to as pricing models, their primary focus is on the mechanisms of return formation. Understanding these models is essential, as they establish the baseline of what constitutes “normal” returns and thereby provide the framework against which potential anomalies and abnormal returns can be identified.

The most influential contributions are the Modern Portfolio Theory of Markowitz (1952), the Capital Asset Pricing Model developed by Sharpe (1964), Lintner (1965) and Mossin (1966), and the Arbitrage Pricing Theory of Ross (1976). Together, these models have shaped the foundations of modern finance by formalizing the relationship between risk and return, and by introducing both single- and multi-factor approaches to asset pricing. In the following subsections, their main concepts and implications are discussed in turn.

2.3.1 Modern portfolio theory

Modern Portfolio Theory (MPT) was introduced by Harry Markowitz in his groundbreaking article *“Portfolio Selection”* (1952), which is widely regarded as the foundation of modern financial economics. The theory is based on the assumption that investors are risk-averse and aim to maximize the expected return of their portfolios for a given level of risk, or equivalently, minimize risk for a given expected return. Markowitz demonstrated mathematically that a well-diversified portfolio can provide superior risk–return trade-offs compared to investments in individual securities.

The essential mechanism behind diversification lies in the correlation structure of asset returns. Since securities rarely move in perfect unison, combining assets with less than perfectly correlated returns reduce the variance of the portfolio. Even if correlations are positive, diversification still lowers overall risk (Knüpfer & Puttonen, 2014). In Markowitz's framework, the expected return of a portfolio is simply the weighted average of the expected returns of its component assets.

$$E(R_p) = \sum_{i=1}^{N} w_i E(R_{i)} = \boldsymbol{w^T \mu},$$

where $E(R_p)$ is the expected return of the portfolio and w_i is the portfolio weights. Same can be formed by using matrices, where $\boldsymbol{w^T}$is vector of portfolio weights and $\boldsymbol{\mu}$ is vector of expected asset returns. The risk of the portfolio, however, depends not only on the individual variances of each asset but also on the covariances between them:

$$\sigma_p^2 = \sum_{i=1}^{N} w_i^2 \sigma_i^2 + \sum_{i=1}^{N} \sum_{i \neq j}^{N} w_i w_j \, \sigma_{ij} = \boldsymbol{w^T \Sigma} \mathbf{w}.$$

In formula σ_p^2 is the variance of portfolio, σ_i^2 is the variance of individual asset and σ_{ij} is the covariance between assets i and j. In the matrix form, $\boldsymbol{\Sigma}$ is the covariance matrix of asset returns. These formulas highlight that diversification reduces portfolio variance by combining assets whose returns do not move perfectly together.

Risk in MPT is divided into two components: systematic risk, which stems from market-wide factors and affects all securities, and unsystematic (idiosyncratic) risk, which is specific to individual companies or industries. Systematic risk cannot be eliminated through diversification, while unsystematic risk can be almost completely diversified away as the number of holdings increases (Markowitz, 1952; Knüpfer & Puttonen, 2014). This distinction is fundamental for portfolio theory: investors are not compensated for bearing idiosyncratic risk, since it can be diversified at no cost.

Markowitz showed that by varying the weights of assets, investors can construct a set of efficient portfolios—those that provide the maximum expected return for a given level of risk.

This set is known as the efficient frontier. Portfolios lying below the frontier are inefficient because the same level of risk could achieve a higher return, or conversely, the same return could be achieved with less risk. The efficient frontier formalizes the trade-off between risk and return: higher returns can only be obtained by accepting greater risk, while reducing risk requires sacrificing expected return.

The theory was later extended by Sharpe (1964), who introduced the possibility of combining risky portfolios with a risk-free asset. This extension leads to the Capital Market Line (CML), which represents all combinations of the risk-free asset and the tangency portfolio (the point of tangency between the CML and the efficient frontier). On this line, investors can select their desired balance of risk and return according to their preferences. More risk-averse investors allocate more weight to the risk-free asset, while more risk-tolerant investors can borrow at the risk-free rate and leverage the market portfolio to achieve higher expected returns. In equilibrium, all rational investors will hold some combination of the risk-free asset and the market portfolio, which reflects the aggregate supply of risky assets in the economy (Sharpe, 1964; Knüpfer & Puttonen, 2014).

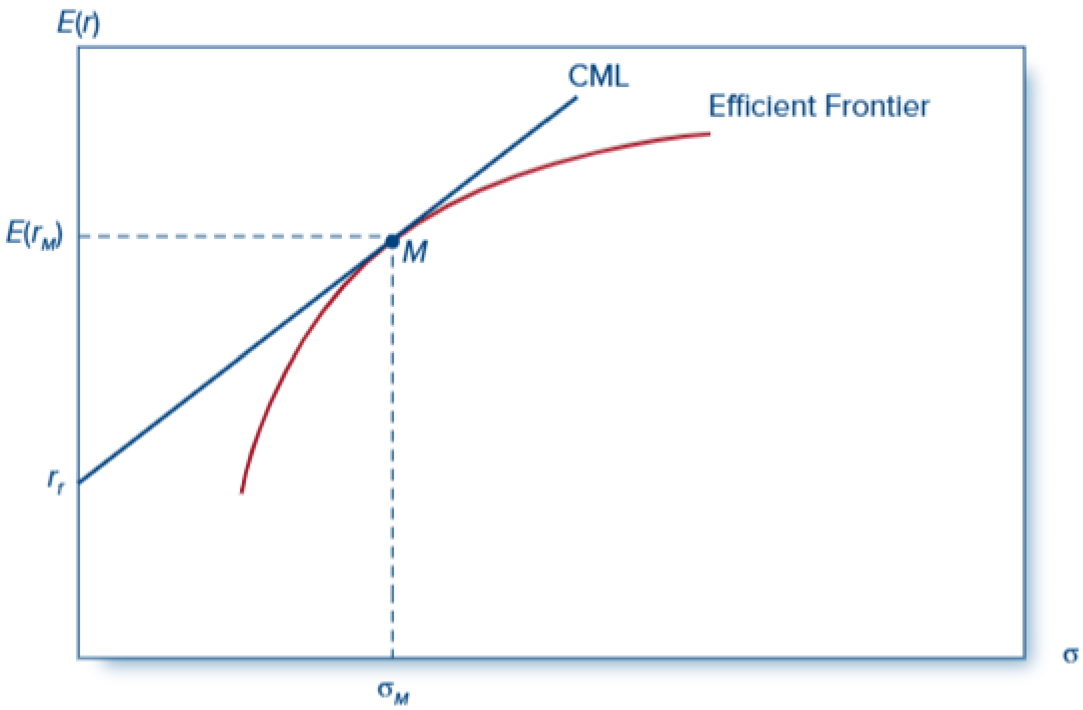

Figure 1 Capital Market Line (Bodie et al., 2024)

An important implication of Markowitz’s framework is the concept of the minimum variance portfolio. This refers to the portfolio that achieves the lowest possible level of risk for a given set of assets. In other words, among all feasible portfolios, the minimum variance portfolio represents the maximum benefit of diversification. The position of this portfolio is determined by the covariance structure of asset returns, since it is precisely the co-movements between securities that allow overall risk to be reduced.

Minimum variance portfolios are particularly relevant for investors who prioritize risk reduction rather than return maximization. They form the lower bound of the efficient frontier and are widely used as benchmarks in empirical research and practical portfolio management. Because their construction relies on the estimation of the covariance matrix, their performance is sensitive to estimation errors, especially in large asset universes.

In this thesis, the concept of the minimum variance portfolio provides the theoretical foundation for the empirical methodology. The detailed optimization problem and its mathematical formulation will be presented later, in the theoretical section.

Although MPT has been highly influential, it has also faced criticism. The model relies on several simplifying assumptions: returns are normally distributed, correlations are stable over time, investors behave rationally, and they all have homogeneous expectations. Empirical evidence has shown, however, that asset returns are not normally distributed but instead display heavy tails and volatility clustering (Mandelbrot, 1963; Fama, 1965; Cont, 2001). Correlations between assets have also been shown to vary over time and to increase sharply during market crises, which undermines the stability of diversification benefits (Longin & Solnik, 2001). Furthermore, behavioral finance research has demonstrated that investors are not always fully rational and that preferences are shaped by heuristics and biases, contradicting the assumption of homogeneous expectations (Kahneman & Tversky, 1979; Shiller, 2003). Despite these limitations, MPT remains a cornerstone of modern finance, as its principles of diversification and risk minimization form the basis of subsequent developments such as the Capital Asset Pricing Model and Arbitrage Pricing Theory (Elton & Gruber, 1997).

2.3.2 Capital Asset Pricing Model

The Capital Asset Pricing Model (CAPM) was developed independently by Sharpe (1964), Lintner (1965) and Mossin (1966) as an extension of portfolio theory. The model links an individual asset's required return to its exposure to market-wide risk and formalizes the idea that only systematic risk is compensated in equilibrium, while idiosyncratic risk can be diversified away in well-constructed portfolios. In the CAPM setting, investors hold combinations of a risk-free asset and the market portfolio, which implies a linear relation between expected return and a single measure of systematic risk.

In its basic form, the expected excess return on asset i is proportional to its beta with respect to the market portfolio. The expected return relation is

$$E(R_i) = r_f + \beta_i\big(E(r_m) - r_f\big),$$

where r_f denotes the risk-free rate, $E(R_m)$ is the expected return on the market portfolio, and β_i measures the sensitivity of asset i to market movements (Sharpe, 1964; Lintner, 1965; Mossin, 1966). The beta coefficient is defined as

$$\beta_i = \frac{cov(r_i, r_m)}{var(r_m)}$$

so that $\beta_i > 1$ indicates that the asset tends to move more than the market on average, whereas $\beta_i < 1$ indicates lower sensitivity. In empirical applications the model is often implemented in regression form:

$$r_i - r_f = \alpha_i + \beta_i\big(r_m - r_f\big) + \varepsilon_i,$$

where α_i measures abnormal performance and ε_i is an idiosyncratic disturbance with $E[\varepsilon_i] = 0$ and $cov\big(\varepsilon_i, r_m - r_f\big) = 0$. Under the CAPM in equilibrium, the expected value of α_i is zero.

Graphically, the relation between expected return and beta is illustrated by the Security Market Line (SML). The SML plots the expected return of assets and portfolios as a linear function of their beta, starting from the intercept r_f and rising with slope $E(r_m) - r_f$. Securities priced fairly lie on the line, observations above (below) the line imply positive (negative) abnormal performance for the level of market risk borne. This idea underlies the widely used Jensen's alpha performance measure in active management (Jensen, 1968). In practice, the market portfolio is proxied by a broad equity index, and the risk-free rate by a short-term government bill, which allows the model to be implemented for both individual securities and diversified portfolios.

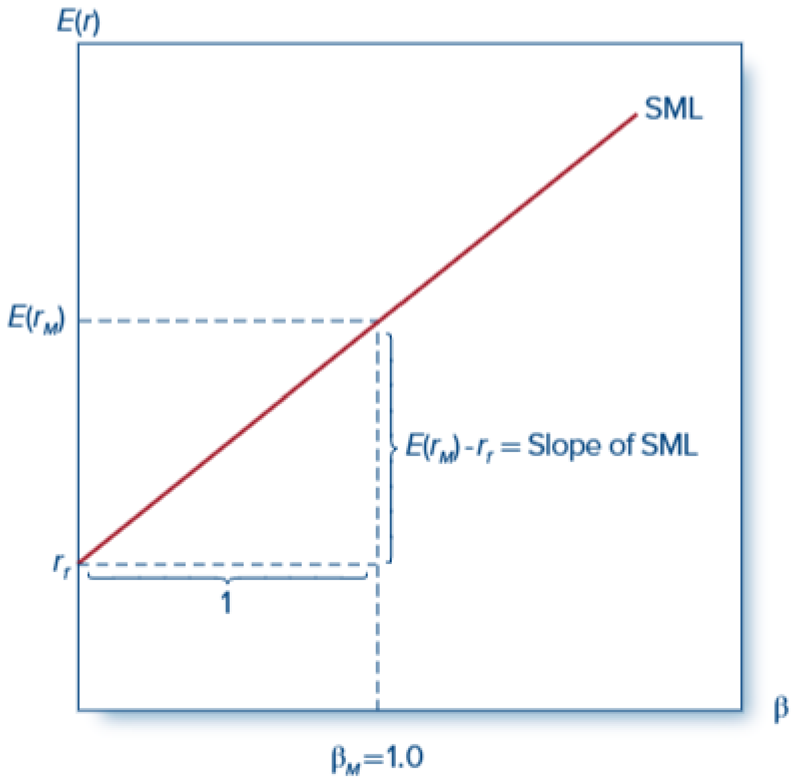

Figure 2 The security market line (Bodie et al., 2024)

Historically, the Sharpe–Lintner formulation assumed risk-free borrowing and lending. Black (1972) relaxed this by developing the zero-beta CAPM, which replaces the risk-free asset with a zero-beta portfolio yet preserves the core linear pricing relation. Together, these variants form the classic single-factor view that underpins much of modern asset pricing.

CAPM rests on a set of simplifying assumptions familiar from mean–variance analysis: investors optimize over a single horizon using mean and variance, borrowing and lending at the risk-free rate is possible, markets are frictionless, and investors have homogeneous expectations about expected returns and the covariance matrix (Sharpe, 1964; Lintner, 1965; Mossin, 1966). Roll (1977) further noted that an exact test of the model is complicated by the unobservability of the true market portfolio, which in principle should include all forms of wealth.

Empirical evidence has nevertheless documented systematic deviations from the model's predictions. Early studies found that the realized trade-off between beta and average return is flatter than the CAPM implies (Black, Jensen & Scholes, 1972; Fama & MacBeth, 1973). Moreover, several cross-sectional patterns are not fully captured by a single beta factor, including the value effect (Stattman, 1980; Rosenberg, Reid & Lanstein, 1985), the size effect (Banz, 1981), earnings-price effects (Basu, 1977), momentum (Jegadeesh & Titman, 1993) and the low-volatility/low-beta anomaly (e.g. Haugen & Baker, 1991; Ang, Hodrick, Xing & Zhang, 2006). A broad review concludes that once size and value are considered, the simple beta–return relation largely disappears (Fama & French, 1992; Fama & French, 2004). These findings point either to missing risk factors or to departures from the model's behavioral and informational assumptions.

The literature has responded along two main lines. One extends the theory to allow richer sources of systematic risk, as in Merton's Intertemporal CAPM (ICAPM), where investors care about future investment opportunities and multiple state variables can become priced (Merton, 1973), and in consumption-based models (CCAPM) that link expected returns to covariation with consumption growth (Lucas, 1978; Breeden, 1979; see also Hansen & Singleton, 1982). In an alternative yet complementary direction, Arbitrage Pricing Theory (APT) provides a no-arbitrage foundation for multifactor linear pricing without specifying the market portfolio (Ross, 1976). Empirically, these ideas coalesced in Fama and French's (1993) three-factor model (market, size, value), later extended with momentum (Carhart, 1997), which better matches observed return patterns than the single-factor CAPM.

Despite these limitations and extensions, CAPM remains a central benchmark because it ties the risk–return trade-off of portfolio theory to a tractable, market-based measure of systematic risk and provides a transparent baseline for applications such as the cost of equity and performance evaluation (Fama & French, 2004). In this thesis the CAPM serves as the conceptual bridge from diversification to factor-based thinking: common (systematic) risk drives required returns, while idiosyncratic components average out in diversified portfolios. Methodologically, rather than assuming a single factor, this thesis focusses on accurately modeling co-movements in returns. The covariance structure is estimated directly and extracts latent common components using principal component analysis, which are then used in constructing portfolios.

2.3.3 Arbitrage Pricing Theory

Arbitrage Pricing Theory (APT), introduced by Ross (1976), provides an alternative linear asset-pricing framework to the single-factor CAPM by allowing multiple sources of systematic risk to determine expected returns. The core idea is that in well-functioning markets with no-arbitrage, the expected return on a sufficiently diversified asset or portfolio must be a linear function of its exposures to a small number of common risk factors. Unlike CAPM, APT does not require the market portfolio to be mean–variance efficient or even observed; instead, it relies on a no-arbitrage condition in economies where asset returns admit an approximate factor structure (Ross, 1976; Chamberlain & Rothschild, 1983).

In the APT representation, the one-period return on asset i at time t can be written as:

$$R_{i,t} = E(R_i) + \sum_{k=1}^{K} \beta_{i,k} f_{k,t} + \varepsilon_{i,t}$$

Here, $E(R_i)$ is the expected return of asset i, $\beta_{i,k}$ is its loading on factor k, $f_{k,t}$ is the realization of factor k common to all assets in period t, and $\varepsilon_{i,t}$ is an idiosyncratic component with $E(\varepsilon_{i,t}) = 0$ that diversifies away in large portfolios. Under no-arbitrage, expected returns satisfy a linear pricing relation:

$$E(R_i) = r_f + \sum_{k=1}^{K} \beta_{i,k} \lambda_k ,$$

with r_f the risk-free rate and λ_k the factor risk premia (Ross, 1976; Chamberlain & Rothschild, 1983; Lehmann & Modest, 1988). This means that only common components of risk command compensation, whereas asset-specific risk does not, because it can be diversified.

A useful way to understand APT logic is through a simple arbitrage-free argument. Suppose that two well-diversified portfolios, A and B, have identical factor loadings (the same $\beta_{i,k}$ for all $k = 1, \dots, K$), but different expected returns. An investor could short the portfolio with the lower expected return and long the portfolio with the higher expected return. Since the factor exposures cancel each other out, the resulting position is (approximately) factor-neutral and therefore almost risk-free but offers a positive expected return – an arbitrage. In competitive markets, such opportunities cannot persist. Therefore, portfolios with the same beta vector must have the same expected return, which implies a linear pricing relationship in betas with factor-specific risk prices λ_k. (Ross, 1976) This argument underlies both the exact APT in limited markets with perfect diversification and the approximate APT, which becomes increasingly valid as the cross-section of assets increases (Chamberlain & Rothschild, 1983).

Relative to CAPM, APT is more general in its assumptions. It does not require investors to be mean–variance optimizers, does not assume a specific return distribution, and does not rely on the efficiency or observability of a single market portfolio (Roll, 1977). The key assumptions are:

1. asset returns admit a linear factor structure,
2. idiosyncratic risks are diversifiable in large portfolios, and
3. there are no arbitrage opportunities.

Under these conditions, expected returns are linear in factor loadings. The cost of this generality is that theory does not pin down which factors are relevant or how many should be included, that is left to empirical identification. (Ross, 1976.)

In empirical work, factor identification proceeds along three complementary routes. First, macroeconomic or otherwise observable factors are selected on economic grounds. A classic example is Chen, Roll and Ross (1986), who proxy systematic risk using industrial production growth, unexpected inflation and changes in expected inflation, a default-risk premium, and the term-structure spread, several of these variables carry significant prices of risk in cross-sectional tests and help explain stock returns. Second, statistical or latent factors are extracted directly from returns using principal component analysis or related methods. Connor and Korajczyk (1988) develop estimation and testing procedures for large cross-sections and show that a small number of latent factors capture most covariation without imposing an economic label ex ante. This route is widely used in modern risk models, the trade-off is interpretability. Third, characteristics- or style-based factors are constructed from empirical regularities in average returns, such as size, value, and momentum. Factor-mimicking portfolios based on these characteristics are used as priced factors in multifactor models (Fama and French, 1993; Carhart, 1997). While not derived from macroeconomic primitives, these factors are APT-consistent if they proxy for pervasive risks that command premia.

Estimation typically follows a two-step logic. Time-series regressions estimate each asset's beta vector against a chosen factor set

$$\left(r_{i,t} - r_{f,t}\right) = \alpha_i + \sum_{k=1}^{K} \beta_{i,k}\, f_{k,t} + u_{i,t},$$

where $f_{k,t}$ are factor returns and $u_{i,t}$ is an idiosyncratic disturbance. Cross-sectional regressions then relate average returns to the estimated betas to recover factor prices λ_k. The Fama–MacBeth (1973) approach operationalizes this by running cross-sectional regressions period-by-period and averaging the resulting λ estimates, while controlling for sampling error. When factors are estimated and not observed (e.g., PCA factors), errors in the variables can weaken beta estimates and distort λ estimates. Shanken (1992) analyzes the joint distribution of factors and beta estimates and provides corrections for standard errors. Practical implementations must also choose the number of factors, information criteria for approximate factor models (e.g. Bai–Ng criteria) are commonly used in large cross-sections to determine K.

Factor-mimicking portfolios provide a bridge between theory and implementation. Given either macro variables or statistical factors, one can form tradable portfolios whose returns replicate the behavior of the underlying factors (Grinold & Kahn, 2007). In a K-factor setting, the return covariance matrix of N assets, $\boldsymbol{\Sigma}$, can be written in the familiar factor form

$$\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{\Sigma}_f\boldsymbol{B}^T + \boldsymbol{D},$$

where $\boldsymbol{B}$ is the $N \times K$ matrix of betas, $\boldsymbol{\Sigma}_f$ is the $K \times K$ factor covariance matrix, and $\boldsymbol{D}$ is a diagonal matrix of idiosyncratic variances. (Cochrane, 2005.) This structure both motivates and enables practical risk modeling: factor variances and covariances are typically more stable and parsimonious to estimate than the full $N \times N$ covariance matrix, and the decomposition clarifies how much portfolio risk comes from each factor versus idiosyncratic components. In high-dimensional settings, statistical factors obtained via PCA offer a data-driven basis for $\mathbf{B}$ and $\boldsymbol{\Sigma}_f$, while shrinkage of $\boldsymbol{D}$ and $\boldsymbol{\Sigma}_f$ can improve out-of-sample stability. (Connor & Korajczyk, 1988; Bai & Ng, 2002.)

The empirical evidence generally supports a multifactor view over a single-factor description. Early studies such as Roll and Ross (1980) find multiple pervasive factors in stock returns, rejecting a one-factor explanation. Macroeconomic tests document that economic shocks are priced in the cross-section, while characteristic and latent-factor models explain patterns—such as size and value premia—that the CAPM leaves unexplained. At the same time, APT tests are joint tests of an approximate linear factor structure and the chosen factor set, conclusions can be sample-dependent, and the proliferation of candidate factors raises concerns about model selection, errors-in-variables and overfitting (Shanken, 1992; Daniel & Titman, 1997). Conditional or time-varying specifications (e.g. allowing factor premium or betas to vary with instruments such as dividend yields or volatility) can improve fit but add complexity and estimation risk.

Applications are extensive. In portfolio construction and risk control, factor models provide parsimonious and often more stable covariance estimates and a natural decomposition of portfolio risk into factor contributions. In a K-factor world, the efficient frontier can be spanned by a small set of factor-mimicking portfolios plus the risk-free asset this simplifies allocation and mitigates estimation error in high-dimensional settings. For risk attribution, portfolio variance, $Var(r_p)$, can be decomposed into the sum of factor contributions $\beta_p^T \boldsymbol{\Sigma_f} \beta_p$ and idiosyncratic components $w^T \boldsymbol{D} w$, clarifying which economic exposures drive total risk,

$$Var(r_p) = \beta_p^T \boldsymbol{\Sigma}_f \beta_p + w^T \boldsymbol{D} w,$$

where $\beta_p = \boldsymbol{B}^T w$. In performance evaluation, multifactor benchmarks distinguish between returns due to rewarded factor exposures and true alpha (Connor and Korajczyk, 1986). In corporate finance and valuation, multifactor implementations of APT offer an alternative to CAPM for the cost of equity by aligning expected returns with a firm's exposures to priced macro or style factors rather than to the market alone, sector- or region-specific factors (e.g. term-structure or currency exposures) can be incorporated when relevant. (Carhart, 1997.)

International versions of APT extend the framework to multiple countries and currencies, recognizing that global and local factors jointly drive returns. Studies such as Solnik (1983) and subsequent work explore whether exchange-rate risk, global term-structure shifts, and regional business-cycle indicators are priced. For global portfolios, separating global from country-specific factors and modeling currency exposures are standard APT-consistent practices in risk models used by international asset managers.

A natural bridge between APT and empirical implementation is provided by principal component analysis (PCA). In large cross-sections, excess returns admit an approximate factor structure, and PCA recovers the dominant directions of common variation without imposing economic labels ex ante. In this thesis latent-factor viewpoint is adopted primarily for stock selection: PCA is used to summarize common co-movements in returns and to translate the resulting loadings into asset-level scores that guide the construction of the investable set. After selection, I estimate the return covariance matrix for the chosen assets and solve a minimum-variance optimization to obtain portfolio weights. In this way, PCA acts as a data-driven conduit for APT-consistent common components, while the optimization step manages overall portfolio risk through the empirically estimated covariance matrix, detailed modeling choices and diagnostics are presented in the methodology section.

2.4 Behavioral Finance

Behavioral finance departs from traditional asset-pricing by foregrounding how real human behavior influences markets. The approach is motivated by empirical episodes where prices move contrary to the predictions of fully rational, frictionless models. In addition to the presence of non-rational traders, a key argument is that arbitrage is only weakly effective once trading costs, financing and short-sale limits, and model misspecification are acknowledged (Bodie et al., 2024). Drawing on cognitive psychology, this section surveys recurrent decision shortcuts and biases observed in investor choices (Barberis et al., 2003). Heuristics mean simplified rules used to make judgments under uncertainty. The most relevant phenomena for financial markets include representativeness, conservatism, overconfidence, and the preference patterns characterized by prospect theory (Kahneman & Tversky, 1979).

Understanding these mechanisms matters for the idea behind this work. If beliefs and trading constraints shape prices, then co-movements and return regularities can be episodic: they emerge, strengthen, and fade as investor psychology and market frictions vary over time. Such dynamics imply that empirical analysis should allow for shifting common components rather than assume a fixed set of stable drivers, and it should pay particular attention to downside and concentration risks that arise when sentiment becomes one-sided or liquidity thins.

These considerations have observable counterparts. Periods of elevated investor sentiment or attention often coincide with episodic co-movements, for example, style rotations, momentum surges and subsequent crashes, or value rebounds (Barberis et al., 2003). Because arbitrage is constrained, such patterns can persist temporarily rather than vanish on discovery (Shleifer & Vishny, 1997; Merton, 1987). Consistent with this, the strength of common components in equity returns varies over time, tending to rise during sentiment-driven episodes and to recede as conditions normalize (Ritter, 2003).

2.4.1 Cognitive Psychology

Cognitive psychology concerns the processes by which people perceive, form beliefs, and make choices under uncertainty (Ritter, 2003). A defining observation is that errors are systematic rather than purely random: individuals rely on mental shortcuts and stable preference patterns that simplify decision making but also distort it in predictable ways (Barberis & Thaler, 2003). These patterns matter for financial markets because asset-pricing models typically presume coherent preferences and Bayesian belief updating, when these conditions fail, prices may deviate from rational-expectations benchmarks for extended periods (Bodie et al., 2024).

A natural starting point is the role of heuristics in judgment. Heuristics are simplified procedures that economize attention and computational effort (Tversky & Kahneman, 1974). In everyday life they enable fast, workable decisions, in markets they can bias probability assessments and reactions to news (Ritter, 2003). Tversky and Kahneman (1974) describe three foundational heuristics, representativeness, availability, and anchoring, that have repeatedly been documented empirically. Representativeness leads people to evaluate events by their similarity to salient prototypes. When applied to time series, this induces extrapolation from short-run patterns and neglect of base rates: a stock that has performed unusually well is perceived as belonging to a "strong performer" category and is expected to keep doing so, even when such streaks are largely noise (Ritter, 2003). Availability biases the assessment toward vivid, recent, or easily recalled information, such as front-page news, dramatic earnings surprises, and extreme returns, while less visible but diagnostically significant information is ignored. (Tversky & Kahneman, 1974). Anchoring refers to the disproportionate influence of initial reference values on subsequent judgments. In finance, prior price levels, round numbers, or recent trading ranges serve as anchors, and adjustments to new information are typically insufficient, yielding slow drift toward fundamental values (Tversky & Kahneman, 1974; Ritter, 2003).

The interaction of these heuristics with belief dynamics generates a characteristic pattern of under- and overreaction. In behavioral finance, conservatism refers to slow updating of beliefs, meaning that investors react too little to new information and stick to their previous views, which delays price reactions (Bodie et al., 2024). As signals accumulate, investors can then assess trends based on representativeness. The same forces that delay the initial update may later cause an overreaction when the pattern becomes clear enough to be considered "representative" of the new system (Ritter, 2003). Such sequences are consistent with the empirical coexistence of medium-horizon momentum and longer-horizon reversals without invoking time-varying risk alone (Barberis & Thaler, 2003).

Overconfidence further reinforces this dynamic. When making assessments, people tend to overestimate the accuracy of their private information and the validity of their predictions. In such cases, they attribute success to their own skills and failure to chance (Barberis & Thaler, 2003). In financial markets, these trends increase trading and risk-taking and undermine moderate, evidence-based decision-making. Barber and Odean (2001) document that higher turnover is associated with lower performance at the investor level, a result consistent with overconfidence eroding risk-adjusted returns.

Prospect theory provides a complementary account of preferences under risk. Instead of evaluating outcomes solely in terms of final wealth levels, people assess gains and losses relative to a reference point, typically the status quo or a salient benchmark. The value function is concave in gains and convex in losses, with a steeper slope for losses than for equal-sized gains, loss aversion (Kahneman & Tversky, 1979). Decision weights distort objective probabilities, overestimating small probabilities and underestimating moderate probabilities (Kahneman & Tversky, 1979). Together, these characteristics point to context-dependent risk-taking attitudes. Individuals are generally risk-averse when it comes to gains but risk-seeking when it comes to losses. This results in several well-documented patterns of market behavior, such as the disposition effect (it is easier to realize gains than losses), demand for lottery-like payoffs, and asymmetric reactions to positive versus negative surprises (Odean, 1998; Ritter, 2003).

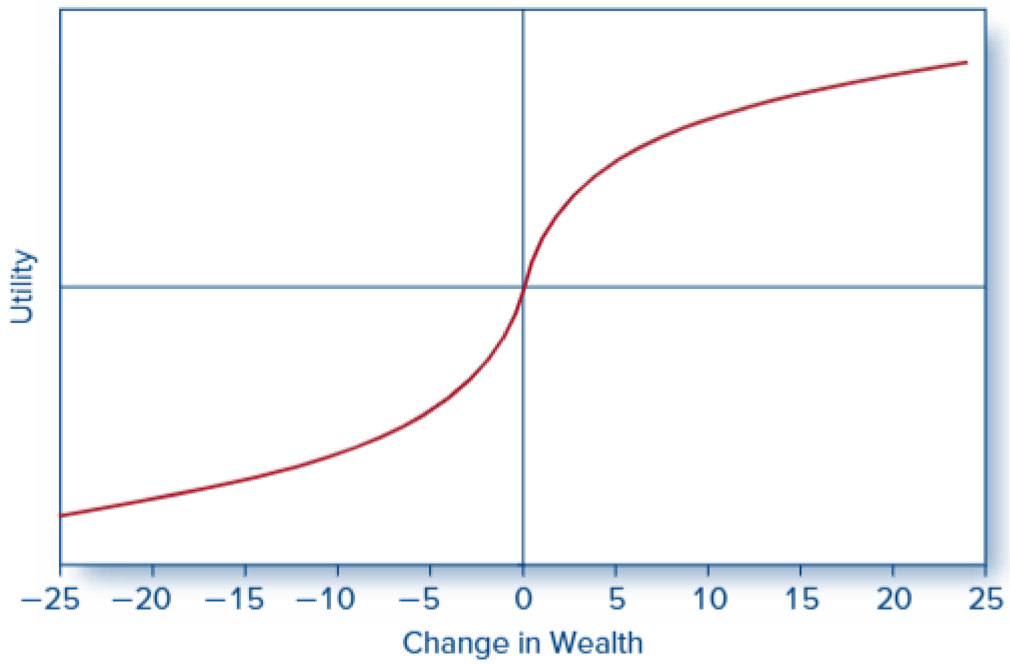

Figure 3 Utility Function under Prospect Theory (Bodie et al., 2024)

Limited attention and limited processing capacity are common background factors that influence how the above-mentioned mechanisms appear in the market. Information is constantly coming in, but investors cannot pay attention to everything at once, instead, they sort information based on its importance, timing, and presentation (Barberis & Thaler, 2003). When information is complex, fragmented, or arrives outside of concentration time, price changes are slow and incomplete. When information is clear or coordinated across different channels, reactions are rapid but may be too strong (Ritter, 2003). These information dynamics help explain phenomena such as post-earnings-announcement drift, underreactions to complex announcements, and periodic increases in volatility around significant events (Barberis & Thaler, 2003).

An important implication of the cognitive perspective is that beliefs and preferences vary across investors and over time, rather than being homogeneous and stationary (Barberis & Thaler, 2003). The variation is driven by differences in experience, knowledge, and cognitive constraints. Variation over time arises when significant narratives change, when media cycles highlight different risks, and as feedback from market prices affects reference points and confidence (Ritter, 2003). Empirically related phenomena include medium-term momentum in stock returns (Jegadeesh & Titman, 1993), slower price adjustment when investor attention is limited (Da et al., 2011), and lottery-demand bias associated with lower subsequent average returns (Bali et al., 2011). The cross-sectional and temporal aggregation of these heterogeneous, changing assessments produces common patterns in prices that are neither purely idiosyncratic nor fixed in form. Some periods show strong common trends related to attention-grabbing themes or categories, while other periods show more fragmented coverage, with individual information dominating. (Barberis & Thaler, 2003.)

Cognitive biases also play a role in institutional contexts, which can either reinforce or mitigate their effects. Professional investors must face reporting cycles, benchmark indices, and tax-related issues that affect their attention and risk-taking. Even if individual professionals are aware of biases, organizational incentives can reinforce extrapolation, for example by following recently successful styles, or encourage passivity, such as unwillingness to realize losses (Ritter, 2003). These institutional layers are not arbitrage disputes themselves but rather relate to the limitations of arbitrage. However, they are channels through which psychological regularities are expressed in trading prices. Identifying these channels is essential for interpreting apparent pricing errors and understanding why the same biases can produce different market outcomes at different times and for different investor groups. (Barberis & Thaler, 2003.)

In summary, it can be said that the cognitive psychology perspective emphasizes structured deviations from a rational reference point: probability estimates are modified by representativeness, availability, and anchoring, belief updating is conservative and followed by extrapolation, confidence is often excessive, and preferences reflect loss aversion and probability weighting. These deviations produce predictable patterns of underreaction and overestimation, attention-driven asymmetries, and context-dependent risk-taking. (Kahneman & Tversky, 1979; Ritter, 2003). The following subsection considers why competitive markets do not immediately eliminate such effects through arbitrage, focusing on the practical constraints faced by would-be arbitrageurs.

2.4.2 Limits to arbitrage

In theory, arbitrage is risk-free but in reality, it demands capital and deliberate risk-bearing (Merton, 1987). Professional arbitrageurs typically minimize exposure and therefore avoid the most volatile environments, which makes correction slower exactly when volatility is high (Shleifer & Vishny, 1997). Efficient markets reflect fundamental factors, and there is no real arbitrage. Behavioral economics does not dispute that pricing errors, when they occur, attract rational traders to exploit them. However, it questions when and to what extent this occurs in practice. Strategies aimed at eliminating mispricing can be risky and costly, so their returns may fall short of what theory suggests. The key sources of risk and cost include fundamental risk, noise-trader risk, and implementation costs (Barberis & Thaler, 2003).

In asset pricing, arbitrage is the mechanism that should align market prices with fundamental values. Behavioral finance emphasizes that the limits of arbitrage determine whether rational traders can exploit and eliminate pricing errors and how quickly (Barberis & Thaler, 2003). In theory, pure arbitrage is risk-free but in practice it requires capital, tolerance for interim losses, and access to instruments such as short sales, so correction is neither instantaneous nor costless (Shleifer & Vishny, 1997).

A central constraint is fundamental risk. While a position is open, news can change intrinsic values, and imperfect hedges leave residual exposure. Even selling a close substitute short, for example, company in the same industry, seldom eliminates risk entirely, so an apparently attractive trade can move against the arbitrageur before convergence (Barberis & Thaler, 2003). Since funding is often marked to market and performance is evaluated over a limited time frame, these temporary risks affect position sizes and willingness to initiate trades (Shleifer & Vishny, 1997).

Order flow driven by emotions or impulsive reactions can push prices further away from fundamentals, exposing opposing positions to additional losses in the short term. De Long et al. (1990) show how such non-rational demand can persist and how its volatility makes betting against it risky. If capital providers respond to losses by withdrawing funding, arbitrage capacity shrinks at precisely the wrong moment, prolonging the deviations.

A further set of frictions arises from implementation costs. Identifying opportunities consume research resources. For this reason, implementation involves brokerage fees, bid–ask spreads, and market impact such as collateral related to maintaining positions and general operating costs. Once taxes are included, net returns compress and only the largest mispricings remain attractive (Herschberg, 2012). In addition, many corrections require short selling. Limited lendable supply, borrow fees, and recall risk make overpricing harder to attack than underpricing, slowing the incorporation of negative information (Cochrane, 2011).

These limits imply that mispricings are arbitraged only up to the point where expected gains cover fundamental risk, financing and funding fragility, noise-trader risk, and implementation costs. Prices don’t always match fundamental values, even when markets are competitive. How fast and how fully prices move back toward fundamentals depends on market conditions. When funding for arbitrage is tight and volatility is high, the adjustment typically takes longer (Shleifer & Vishny, 1997).

2.5 Quantitative Finance

Quantitative finance applies probability and statistics to characterize uncertainty in asset prices and guide disciplined decision making in the face of uncertainty. It treats returns, risks, and constraints as measurable objects and builds models that can be verified against data and evaluated outside the sample (Vogl, 2022).

The principle of modelling is pragmatic and adaptive because markets evolve, and the usefulness of any model depends on the prevailing market environment. Methods are evaluated based on how well they hold up in changing conditions and whether their assumptions remain reasonable as conditions change (Lo, 2017).

Data processing is the foundation of every quantitative process. The workflow begins with defining the universe, calendar alignment, and corporate action adjustments, and continues with construction of excess returns, standardization, and careful handling of missing values and outliers. These steps determine what estimators observe and they have a strong influence on out-of-sample behavior and error control (López de Prado, 2018).

The family of methods in quantitative finance is broad. Classical time series models such as ARIMA capture short run dynamics in levels and differences and they provide a baseline for forecasting and diagnostics in financial data that are often close to unit root behavior (Tsay, 2010). Volatility clustering is addressed with GARCH type models, which allow the conditional variance to evolve through time and improve risk measurement relative to constant variance assumptions in many datasets (Tsay, 2010). When there is no closed form solution to a pricing or risk problem, Monte Carlo simulation is used to estimate expectations under the relevant probability measure and to quantify distributional tail risks that are important to investors and risk management (Glasserman, 2004). Derivative pricing and hedging link models to market prices through no-arbitrage logic and numerical methods that translate differential equations into implementable strategies (Wilmott, 2006). Machine learning expands the toolkit with feature extraction, nonlinearity, and flexible function approximation, but its value depends on a research process that controls for data collection, nonstationary, and the small sample nature of many financial problems (López de Prado, 2018).

Risk modelling plays a key role because portfolio decisions depend on second moments and on co-movement of different assets. The sample covariance matrix is unstable when the number of assets is large relative to the estimation window and this instability leads to unreliable optimized weights, which motivates shrinkage and other regularization to improve conditioning (Ledoit & Wolf, 2004). Portfolio construction inherits the quality of the covariance input, and the benefits of diversification are only realized when risk is measured using estimators that are robust to noise (Markowitz, 1952).

The entire workflow can be condensed into a repeatable process. Data is collected and cleaned so that estimators see consistent inputs. Features, forecasts, or risk metrics are estimated in rolling windows. Signals or risk summaries are produced. The portfolio rule maps these inputs to weights under liquidity and risk limits (López de Prado, 2018). Implementation connects portfolio targets to execution algorithms and cost models so that realized returns reflect both alpha generation and trading quality in electronic markets (CFA Institute, 2024). Practical manuals show how to convert model outputs into positions and how to set numerical details that keep daily operations stable (Wilmott, 2006). Industry practice connects this process to concrete rebalancing rules, risk budgets, and compliance checks that guide daily portfolio management (Chincarini and Kim, 2010).

Validation and backtesting standards protect against false discoveries. Reliable research defines data sets that do not contain future information, uses walk forward evaluation, and reports performance reliability metrics in multiple market conditions (López de Prado, 2018). Overfitting is a significant risk in the financial sector, as the number of trials can be large relative to the true signal, so diagnostics are needed to assess how likely it is that reported results are due to chance. Backtesting should evaluate out-of-sample performance based on repeatable skill rather than luck, which is why multiple testing and data mining must be controlled in the design. (Bailey et al., 2017.)

Implementation constraints reduce paper returns and must be modelled explicitly. Realistic assessments account for explicit costs such as commissions and fees and for implicit costs such as bid ask spreads, slippage, and market impact that scale with liquidity and order size in electronic limit order books (CFA Institute, 2024). Trading costs arise from the interaction of orders with the limit order book, and they accumulate through spread, depth, timing, and adverse selection, which means design choices in order placement and routing matter for realized outcomes (Harris, 2003). Market microstructure evidence shows that modern market makers earn the spread by supplying liquidity with fast strategies and that realized performance depends on how orders are routed and timed in this environment (Menkveld, 2013).

Institutional use cases show how the toolkit is deployed in practice. Hedge funds and proprietary trading firms implement systematic strategies that translate signals into trades under risk limits and cost budgets, while dealers and banks use quantitative models for pricing, hedging, and risk reporting across large books of instruments that trade in markets (Wilmott, 2006). Market structure research documents how algorithmic and high frequency trading shape liquidity provision, volatility transmission, and execution quality, which explains why implementation choices are integral to quantitative performance (Menkveld, 2013).

Quantitative finance is therefore a process discipline that links measurement to action through models that are explicit about uncertainty. The approach scales to large datasets and it can be tested and refined, yet it remains constrained by model risk, by the limits of historical information, and by the frictions of real markets, which is why careful validation remains as central as the models themselves (Vogl, 2022).

2.6 Earlier studies

The use of principal component analysis and minimum variance portfolios is common in financial research. In equities they are applied to forecast returns, select stocks, stabilize covariance estimates for risk-first allocation, and to build compact index-tracking or dispersion baskets. This subsection reviews prior studies that are relevant to this thesis.

Principal component analysis has been used directly for stock selection in several equity markets. The common idea is to compress correlated information into a few components and translate them into stock-level signals or compact representative subsets. According to Zhou & Yin (2020), a peer-reviewed study on China's CSI 300 reduces a large, collinear sixty-factor library with kernel principal component analysis and then regresses subsequent returns on the extracted components to obtain coefficients that serve as selection scores. The sample spans 2010–2016 with an out-of-sample test in 2017, estimation is essentially static for extraction with next-period application. The authors report outperformance against the CSI 300 in 2017, with multiple components significant at the one-per-cent level and robustness supported by bootstrap tests. The contribution is to demonstrate that orthogonalized components can be mapped into actionable rankings that outperform simple factor screens when many candidate variables are present.

Yang et al. (2015) propose a PCA-aided selection for the ASX 200 that identifies a small subset of constituents able to span index behavior. Using rolling subperiods to preserve representativeness as correlations change, they show that a portfolio of roughly fifteen stocks can mimic the index closely and that the number needed varies with market conditions.

John et al. (2016) apply PCA to the Ghana Stock Exchange to identify sector-representative names. Starting from thirty-four stocks, their procedure selects twelve that explain most of the cross-sectional variance within sectors. Estimation is static on the sample, and the objective is coverage rather than excess return, yet the results show how components can support stock screening by isolating those names that best span sectoral movements.

Hargreaves & Mani (2015) examine Australian equities by applying PCA to a broad set of company variables and then using a perceptual map of the principal factors to propose winning candidates. The approach is a selection heuristic rather than an optimizer and contains limited out-of-sample benchmarking or explicit risk allocation, but it underscores that principal components can simplify the decision space for practitioners when many inputs are available.

Principal component analysis is also used with other statistical methods to forecast returns and market direction. Ghorbani & Chong (2020) develop a reduced-dimension forecasting method that projects returns onto a principal-component subspace estimated from an exponentially weighted covariance matrix. Using daily data for 150 US stocks across small, mid and large caps together with major indices, they predict one to ten days ahead and compare against moving-average rules and a Gauss–Bayes benchmark. The principal-component filter delivers higher directional accuracy, around 0.78 to 0.80 across cap buckets versus 0.56 to 0.66 for moving-average baselines, and its estimation variance falls as the observation window lengthens, reflecting the stability benefits of rolling estimation. The study interprets the gains as arising from denoising and conditioning improvements when forecasts are made in a low-dimensional PC space.

Waqar et al. (2017) apply principal component analysis as a preprocessing step before linear regression to classify market direction across the London, New York and Karachi exchanges. Input features include open, high, low, current and change, augmented with differ, mid-month trend and monthly trend to induce correlation. After PCA, root-mean-square error drops sharply on LSE (from 16.43 to 1.40) and NYSE (from 36.00 to 1.00), while KSE accuracy worsens (RMSE 0.13 to 1.01), underscoring that component choice and retained dimensionality are critical. The study uses discrete windows rather than a purely rolling setup, and they illustrate the utility of PCA primarily to reduce collinearity and noise before simple linear models.

Cao & Wang (2019) combine principal component analysis and feed forward neural networks to predict changes in stock indices. Financial statement data and trading variables are first compressed into a small orthogonal input set using PCA, after which three back propagation training schemes are compared. The Bayesian regularization network achieves the highest accuracy with smaller prediction errors and less overfitting compared to gradient descent and conjugate gradient variants. The paper also presents a stock-selection overlay driven by the model’s signals and reports that the combined PCA–ANN approach is effective on a sample of companies listed by the authors.

Fan et al. (2019) propose a double-selection Lasso to screen a large factor set, apply PCA to the selected factors, and then fit returns with either ordinary least squares or support-vector regression on the principal components. Using US and Chinese equities (monthly data 2004–2013 for the US training step), they retain ten PC factors after selection. Several factors are statistically significant in OLS, and the non-linear SVR variant improves prediction further. In backtests, the strategy built on PC factors outperforms the Dow Jones Industrial Average and CSI 300, illustrating that combining supervised factor screening with PCA-based regression can yield tradable forecasts in high-dimensional settings.

Principal component analysis is also used to stabilize covariance estimates and to support risk first allocation through minimum variance and related quadratic optimizers. Sun et al. (2020) combine principal component analysis with a risk-aware optimizer in a sector setting. From a library of 91 factors for 76 Chinese energy stocks, they first de-extreme and standardize inputs, then use principal components to orthogonalize the factor set and predict next-period returns with the component scores. Portfolio weights are obtained from quadratic optimization that explicitly penalizes risk. The sample is split into 2013–2016 for model building and 2017–July 2019 for testing, with weekly rebalancing after a seven-period listing buffer. Reported out-of-sample results include an annualized return around 46% and consistent outperformance of the sector proxy, which the authors attribute to noise reduction before optimization and to the explicit treatment of risk in the allocator.

Fan et al. (2013) model the covariance as a low-rank factor part estimated by principal components with a sparse idiosyncratic residual (POET). In minimum variance tests on large equity universes, POET-based covariance yields lower out of sample portfolio volatility and more stable weights than the raw sample covariance. Aït-Sahalia & Xiu (2017) show that a small number of latent factors explain most of the high-frequency stock return covariance, their PCA-based estimator produces lower realized risk for out of sample minimum variance portfolios than naive estimates. Ledoit & Wolf (2004) demonstrate that shrinking noisy sample covariance toward a structured target reduces estimation error and improves minimum variance performance, in practice, combining a PCA factor part with shrinkage or thresholding of the residual further reduces risk and turnover. These findings justify using factor-based, regularized covariance estimators when constructing minimum variance portfolios in high dimensions.

One use case is also to select compact subsets that replicate index behavior or to form baskets for dispersion trades. The purpose is representation and efficient tracking rather than explicit alpha generation. As discussed, earlier Yang et al. (2015) study the ASX 200 and use principal components to identify a small subset of constituents that spans benchmark movements. With rolling subperiods to reflect changing correlations, they report that about fifteen stocks can mimic index behavior closely and that the size of the subset depends on the situation. The approach targets low operational complexity and representation quality rather than return enhancement, and it shows how component structure can guide subset construction.

Bufalo et al. (2025) proposes a hybrid principal component method that factorizes the benchmark and the asset universe to pick a lean S&P 500 tracker. Using a rolling one-year window with weekly updates, they select ten names that jointly explain benchmark variation and show lower tracking error than a conventional optimization based approach in most weeks. The tracking index sometimes generates small positive excess returns compared to the index, while maintaining compactness and avoiding heavy calibration.

Schneider & Stübinger (2020) use principal components to measure each constituent's explanatory power for the S&P 500 and select a subset for a market neutral dispersion strategy. The chosen investment portfolio serves to trade differences between index and constituent volatility. In an out of sample backtest from 2000 to 2017, the strategy generates a financially significant return with a Sharpe ratio of 0.3–0.4 after costs, which shows that PCA-based subset selection can also support volatility-oriented strategies rather than pure tracking.

In summary, the studies show that principal components can be mapped to stock level signals for selection, used as reduced dimension predictors, and employed to stabilize covariance for risk first allocation. Rolling estimation generally improves stability and out-of-sample performance. When using regression-based scores, several studies report better performance than benchmark indices. When using factor-based covariance with shrinkage or sparse residuals, minimum variance portfolios produce lower risk and more stable weights. In index representation settings, principal components enable compact baskets that track benchmarks or support dispersion trades. These findings support the design in this thesis: rolling principal component analysis translated into stock level scores and a minimum variance allocator built on regularized covariance estimates.

3 Theoretical Framework

This section introduces the quantitative methods that form the theoretical basis for empirical work. It focuses on principal component analysis and Markowitz's minimum variance portfolio, which together support data-driven stock selection and risk-controlled allocation. First, principal component analysis is presented as a rotation of the return space. I set out the data structure, the eigen decomposition, and the construction of principal component scores. I will clarify how eigenvectors describe directions of co-movement and how eigenvalues measure their importance. This provides the foundation for the scoring rule that is later used to rank stocks. Second, Markowitz's minimum variance portfolio is introduced as a risk only allocation supported by regularized covariance estimation. I will explain how diversification arises from the covariance structure and how dominant components drive portfolio risk. The section closes with a short bridge that links these concepts to the methodology that follows.

3.1 Principal Component Analysis

The purpose of principal component analysis is to find a set of orthogonal components in multidimensional data that represent its essential structure with as little information loss as possible. In geometric terms, PCA identifies the directions and associated subspaces onto which the projection of the centered data attains the greatest variance. When all components are retained, no information is lost. In practice a small subset captures most of the variation. (Jolliffe and Cadima, 2016.) In simple terms, PCA builds a few composite variables from the original variables that change the most, so that data can be described with fewer numbers while retaining the essential information (Jolliffe, 2002).

The idea traces to Pearson's (1901) closest-fit lines and planes and was formalized by Hotelling (1933) into uncorrelated linear combinations ordered by decreasing variance. Subsequent expositions present PCA as an eigenvalue and eigenvector solution for the sample covariance or correlation matrix and as a special case of the singular value decomposition (Jolliffe, 2002). In finance it is used to forecast returns, select stocks, summarize common co-movements, build tracking portfolios, and to extract data-driven factors that describe return dynamics. Beyond finance, PCA underpins pattern recognition tasks such as face recognition or handwriting analysis, where it compresses inputs while retaining structure that helps discrimination (Jolliffe, 2002). In signal processing it is used to denoise and to separate mixed sources by decorrelating signals and isolating dominant oscillations (Gewers et al., 2021).

The analysis begins with a $\boldsymbol{T} \times \boldsymbol{N}$ data matrix $\boldsymbol{X} \in \mathbb{R}^{T\times \mathrm{N}}$, whose rows index observations and columns index variables. Centering the data by subtracting the sample mean of each series is a common procedure, but it is not mandatory. Omitting centering causes the principal components to maximize variation relative to the origin and may skew the first component toward the mean (Jolliffe, 2002). Since PCA is not invariant to translations, centering changes the components. A covariance or correlation matrix is calculated from the data. In covariance-based PCA, centering is implicit when the sample covariance $\boldsymbol{S}$ is formed, so the eigenvectors correspond to variables cleaned of their mean (Jackson, 1991). The sample covariance is $\boldsymbol{S} = \frac{1}{T-1}\widetilde{\boldsymbol{X}}^{\boldsymbol{T}}\widetilde{\boldsymbol{X}}$, where $\widetilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{1}_T\boldsymbol{\mu}^T$ (Jolliffe, 2002). In this equation, $\widetilde{\boldsymbol{X}}$ denotes the centered data and $\boldsymbol{\mu}$ is vector of sample means. Scaling is a modelling choice, not a requirement (Jackson, 1991). Covariance-based PCA with unstandardized returns preserves the original units and allows variance differences to convey information, which is often desirable in finance when volatility itself is meaningful (Jolliffe & Cadima, 2016). In contrast, correlation-based PCA standardized each series to unit variance so that all series contribute equally to the extraction of components, which is useful when variables are on incomparable scales (Anderson, 1984).

When correlation-based PCA is used, the analysis is equivalent to applying covariance-based PCA to z-scored data (Anderson, 1984). Loadings are often defined as correlations between standardized variables and corresponding component scores. The eigenvalue provides a scale that links the eigenvector coefficients to these correlations, which helps to interpret patterns of common movement (Jolliffe, 2002). With both inputs, rotation decorrelates the new variables, but does not generally make them independent (Jackson, 1991). The total variance is equal to the trace of the input matrix and is preserved under orthogonal change of basis, so the eigenvalues tell how variance is distributed across components (Jolliffe, 2002). In covariance-based PCA, series with high variance usually dominate the first components, which is a desirable feature when the scale itself is informative in terms of risk (Jolliffe & Cadima, 2016).

An eigenvector $\boldsymbol{q}_j$ is a set of weights that defines the direction of co-movement across the original variables. The entries of $\boldsymbol{q}_j$ are the loadings, so large positive values indicate variables that move together according to the model, while large negative values indicate variables that move in the opposite direction. The corresponding eigenvalue λ_j measures how much variability the data has along that direction. It is equal to the variance of the associated scores and indicates the relative importance of the pattern. The sign of $\boldsymbol{q}_j$ and of the score is arbitrary and can be flipped without changing meaning. The orthogonality of different eigenvectors means that the components are uncorrelated (Jolliffe, 2002).

Projecting the data onto $\boldsymbol{q}_j$ gives the score series $\boldsymbol{z}_j$ where each observation's score is its coordinate along the jth pattern. Large positive scores mean alignment with variables that have positive loadings on $\boldsymbol{q}_j$, large negative scores indicate alignment with variables that have negative loadings, and values near zero indicate little relation to that pattern (Jolliffe, 2002). With standardized inputs the scores are dimensionless and comparable across components, whereas with unstandardized inputs the units reflect the original scale and high variance variables tend to influence the first components more (Jackson, 1991). Since the symbols are arbitrary, comparisons are understood through the common sign flip. Their interpretation is based on relative magnitudes and load structures rather than their absolute symbols (Jolliffe, 2002).

Principal directions and their variances are obtained from the spectral decomposition.

$$\boldsymbol{S} = \boldsymbol{Q\Lambda Q}^{\top}, \qquad \boldsymbol{Q}^{\top}\boldsymbol{Q} = \boldsymbol{I}, \qquad \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_N), \qquad \lambda_1 \geq \cdots \geq \lambda_N \geq 0.$$

Here $\boldsymbol{S}$ is the sample covariance matrix, $\boldsymbol{Q} = [\boldsymbol{q_1}, \dots, \boldsymbol{q_N}]$ collects the eigenvectors, and the $\boldsymbol{\Lambda}$ represents the eigenvalues.

To maximize variance, the first weight vector $\boldsymbol{q}_1$ must be:

$$\boldsymbol{q}_1 = \arg\max_{q^{\top}q=1} \boldsymbol{q}^{\top}\boldsymbol{Sq} = \arg\max_{||q||=1} ||\widetilde{\boldsymbol{X}}\boldsymbol{q}||^2,$$

and the jth eigenvector is:

$$\boldsymbol{q}_j = \arg\max_{\substack{q^{\top}q=1 \\ q^{\top}q_l=0,\, l<j}} \boldsymbol{q}^{\top}\boldsymbol{Sq} = \arg\max_{\substack{||q||=1 \\ q^{\top}q_l=0,\, l<j}} ||\widetilde{\boldsymbol{X}}\boldsymbol{q}||^2, \qquad (j = 2, .., N).$$

Score matrix $\boldsymbol{Z} = [z_1, \dots, z_N]$ is computed by projection $\boldsymbol{Z} = \widetilde{\boldsymbol{X}}\boldsymbol{Q}$, with $Cov(\boldsymbol{Z}) = \boldsymbol{\Lambda}$ and $Var(z_j) = \lambda_j$, and the jth score series is $z_j = \widetilde{\boldsymbol{X}}\boldsymbol{q_j}$. Using component scores as regressors is convenient because the scores are orthogonal by construction, which removes multicollinearity and stabilizes coefficient estimates when the original predictors are highly correlated. Replacing many correlated variables with a small number of high-variance scores also reduces dimensionality and acts as a form of regularization, which can improve out-of-sample performance when T is limited relative to N (Jolliffe, 2002).

Explained variance EV summarizes how much of the total variance is observed by the leading components. The explained variance of the first K components is:

$$EV_K = \frac{\sum_{j=1}^{K} \lambda_j}{\sum_{j=1}^{N} \lambda_j}.$$

Plotting λ_j in descending order produces a scree curve. Elbows or steep drops indicate dominant directions, while flat tails suggest a noisy structure (Jolliffe & Cadima, 2016). Retaining only the first K components gives a low-rank reconstruction $\widehat{\boldsymbol{X}} = \boldsymbol{Z_K}\, \boldsymbol{Q}_K^{\top} + \mathbf{1}_T \boldsymbol{\mu}^{\top}$ which minimizes squared reconstruction error among all rank-K linear projections. There $\boldsymbol{Q}_K = [\boldsymbol{q}_1, \dots, \boldsymbol{q}_K]$ and $\boldsymbol{Z}_K = \widetilde{\boldsymbol{X}}\, \boldsymbol{Q}_K$, $\mathbf{1}_T$ is $T \times 1$ vector of ones, and $\boldsymbol{\mu}$ is the vector of data means. The number of components K is typically selected by balancing parsimony and fit. Standard rules retain enough components to reach a target level of explained variance and use the scree plot to identify an elbow where additional components add little information. (Jolliffe, 2002.) In high-dimensional factor settings, information-criterion methods select K by penalizing residual variance in the factor model and are consistent under broad conditions (Bai and Ng, 2002).

Matrix Z contains the principal component scores, each column z_j is the series for component j and each row collects the component coordinates for one observation (Jolliffe, 2002). A single score indicates how strongly the observation expresses a given pattern. The relative magnitudes of the components indicate which patterns dominate the observation, and the interpretation should be read in conjunction with the relative magnitudes of the eigenvalues (Jackson, 1991).

3.2 Markowitz Portfolio Theory

In the Financial Theory chapter Modern Portfolio Theory was introduced as the mean variance foundations of diversification and efficient frontier. This subsection focuses on the computational framework that will be used later. Returns are determined for both objects that summarize their joint distribution. The focus of this section is on issues related to portfolio selection.

Let $\boldsymbol{r}_t \in \mathbb{R}^N$ denote the random vector of asset returns at time t. Its first two moments are summarized by the mean vector $\boldsymbol{\mu} = E[\boldsymbol{r}_t]$ and the covariance matrix $\Sigma = Var(\boldsymbol{r}_t)$. A portfolio is represented by a weight vector $\boldsymbol{w} \in \mathbb{R}^N$, where each element represents the share of capital invested in the asset. The budget constraint requires that the weights sum to one. Short selling and other constraints are modelled using admissible sets for w and these will be presented later.

The one period portfolio return is given by the linear form $R_{p,t} = \boldsymbol{w}^\top \boldsymbol{r}_t$. The expected portfolio return equals $\mu_p = E\left[R_p\right] = \boldsymbol{w}^\top \boldsymbol{\mu}$ and the portfolio variance equals $\sigma_p^2 = Var\left(\boldsymbol{R}_p\right) = \boldsymbol{w}^\top \boldsymbol{\Sigma}\, \boldsymbol{w}$. These identities are the basis for all mean variance results. They provide mapping from asset level quantities to portfolio level moments, and they are the only elements needed to pose the efficient frontier as a quadratic program in the next step. The standard regularity assumptions are maintained for the full rank $\boldsymbol{\Sigma}$ and finite second moments so that these expressions are well defined. Mean variance analysis evaluates portfolios using only the first two moments. This is theoretically complete under multivariate normal returns or quadratic preferences because portfolios can then be ranked by mean and variance alone. In general distributions with skewness or heavy tails higher moments matter, so variance is an incomplete risk summary (Markowitz, 1952). Here, variance is considered as a measure of working risk and the focus on the properties that follow from the quadratic form and its eigenstructure. This setting enables a clear combination of the PCA-based factor view and Markowitz's view. PCA provides information and data-based factorization of return dynamics, while Markowitz's mapping converts this information into portfolio moments.

The minimum variance allocation by the optimization problem that minimizes portfolio variance subject to full investment is characterized by the following optimization program:

$$\min_{\boldsymbol{w} \in R^N} \boldsymbol{w}^\top \boldsymbol{\Sigma}\, \boldsymbol{w}, \quad \text{subject to } \mathbf{1}_N^\top \boldsymbol{w} = 1.$$

Expected returns are not required. The objective is quadratic, and the feasible set is an affine hyperplane. If $\boldsymbol{\Sigma} \succeq 0$ the objective is convex, and if $\boldsymbol{\Sigma} \succ 0$ the problem is strictly convex, and the minimiser is unique. This formulation defines the global minimum variance portfolio on the efficient frontier. (Markowitz, 1952.)

Solving the program by the Lagrangian gives the global minimum variance weights. The Lagrangian is:

$$\mathcal{L}(w,\lambda) = \boldsymbol{w}^{\top}\boldsymbol{\Sigma}\,\boldsymbol{w} - \lambda\,(1_N^{\top}\boldsymbol{w} - 1),$$

where $\lambda \in \mathbb{R}$ is the Lagrange multiplier associated with the budget constraint $1_N^{\top}\boldsymbol{w} = 1$. The first order conditions are $2\boldsymbol{\Sigma}\,\boldsymbol{w} - \lambda\,\mathbf{1}_N = 0$ and $1_N^{\top}\boldsymbol{w} - 1 = 0$. These give $w = \frac{\lambda}{2}\,\boldsymbol{\Sigma}^{-1}\mathbf{1}_N$. Enforcing the budget constraint yields the closed form solution $\mathbf{w}_{GMV} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}_N}{1_N^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{1}_N}$ with variance $\sigma^2_{\text{GMV}} = \frac{1}{\mathbf{1}_N^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{1}_N}$. Here $\mathbf{w}_{GMV}$ denotes the global minimum variance portfolio. The weights are invariant to a positive scalar rescaling of the covariance matrix. The global minimum variance portfolio is located on the left edge of the efficient frontier. If the target return is set at m, the efficient set can be obtained using the programme:

$$\min_{\boldsymbol{w}\in R^N} \boldsymbol{w}^{\top}\boldsymbol{\Sigma}\,\boldsymbol{w} \quad \text{s.t.} \quad 1_N^{\top}\boldsymbol{w} = 1,\ \boldsymbol{w}^{\top}\boldsymbol{\mu} = m.$$

Let $A = \mathbf{1}_N^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{1}_N$, $B = \mathbf{1}_N^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, and $C = \boldsymbol{\mu}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. The minimum variance at target m is

$$\sigma_p^2(m) = \frac{Am^2 - 2Bm + C}{AC - B^2},$$

which shows a parabola in the mean variance plane, with $m = \frac{B}{A}$ and $\sigma^2_{\text{GMV}} = \frac{1}{A}$. (Markowitz, 1952.)

Portfolio risk can also be interpreted in the eigen basis of the covariance. Diagonalising the covariance gives $\boldsymbol{\Sigma} = \boldsymbol{Q}\,\boldsymbol{\Lambda}\,\boldsymbol{Q}^{\top}$ with $\boldsymbol{Q}$ orthonormal and $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_N)$ positive. The portfolio variance becomes $\boldsymbol{w}^{\top}\boldsymbol{\Sigma}\,\boldsymbol{w} = ||\boldsymbol{\Lambda}^{1/2}\,\boldsymbol{Q}^{\top}\boldsymbol{w}||^2 = \sum_{j=1}^{N}\lambda_j\,u_j^2$, where $\boldsymbol{u} = \boldsymbol{Q}^{\top}\boldsymbol{w}$ are portfolio coordinates in the eigenbasis. Inversion acts $\boldsymbol{\Sigma}^{-1} = \boldsymbol{Q}\,\boldsymbol{\Lambda}^{-1}\,\boldsymbol{Q}^{\top}$, so small eigenvalues reinforce instability when calculating $\mathbf{w}_{GMV}$. The conditioning of $\boldsymbol{\Sigma}$ is summarised by $\kappa(\boldsymbol{\Sigma}) = \frac{\lambda_{max}}{\lambda_{min}}$. Expressing risk on this basis corresponds to the PCA section and clarifies why regularization focuses on eigenvalues. (Jollife & Cadima, 2016.)

Estimation error affects the covariance that enters the allocation. In finite samples the sample covariance $\widehat{\boldsymbol{\Sigma}}$ and its smallest eigenvalues are biased downward. Inversion magnifies this error and produces unstable weights when the condition number is large or when dimension is high relative to the sample. These facts motivate eigenvalue-focused regularization such as shrinkage to obtain a well conditioned covariance for Markowitz optimization. A well conditioned estimator improves portfolio risk forecasts and stabilizes allocations. (Ledoit & Wolf, 2004.)

Shrinkage improves the conditioning of the covariance by blending the sample estimate with a structured target. A linear shrinkage estimator takes the form $\Sigma^* = \delta F + (1 - \delta)\,\hat{\Sigma}$, where and $\boldsymbol{F}$ is a well-conditioned target such as a constant correlation model and $\delta \in [0,1]$ is chosen to minimize expected quadratic loss. This reduces dispersion of the sample eigenvalues, preserves positive definiteness, and stabilizes inversion in this kind solutions. In high dimensional settings a nonlinear spectral shrinkage adjusts eigenvalues while keeping sample eigenvectors, which improves conditioning relative to linear shrinkage. (Ledoit & Wolf, 2017.)

In finance, long only constraints reflect common investment rules and frictions. Many mandates prohibit short selling, borrowing is restricted, and the cost of short selling changes the set of feasible trades. Reflecting the minimum variance problem in the simplex produces interpretable weights and avoids leverage, which is consistent with institutional practice. (Cornuéjols & Tütüncü, 2007.) Long only constraints also mitigate estimation error in the sample covariance. The constraints act like an implicit shrinkage that reduces extreme positions and improves out of sample risk control for mean variance portfolios, even though the restrictions are not literally true in terms of the data production process. (Jagannathan & Ma, 2003.) Adding long only constraints augments the budget constraint and restricts the weights to the simplex $\Delta_N = \{w \in R^N\!: \; w \geq 0, \; 1_N^\top \boldsymbol{w} = 1\}$. This gives $\min_{\boldsymbol{w} \in R^N} \boldsymbol{w}^\top \boldsymbol{\Sigma}\, \boldsymbol{w}$ subject to $\mathbf{1}_N^\top \boldsymbol{w} = 1$ and $\boldsymbol{w} \geq \mathbf{0}$. The first order optimality conditions for constrained optimization state that there exist $\lambda \in \mathbb{R}$ and $\gamma \in \mathbb{R}_+^N$ such that $2\Sigma\, w - \lambda\, 1_N - \gamma = 0$, $1_N^\top \boldsymbol{w} = 1$, $w \geq 0$, $\gamma \geq 0$, and $\gamma_i w_i = 0$ for all i. Hence $w_i > 0$ implies $2(\boldsymbol{\Sigma}\mathbf{w})_i = \lambda$ and $w_i = 0$ implies $2(\boldsymbol{\Sigma}\mathbf{w})_i \geq \lambda$. (Boyd & Vandenberghe, 2004.)

Portfolio risk is measured by the volatility $\sigma_p = \sqrt{\boldsymbol{w}^\top \boldsymbol{\Sigma}\, \boldsymbol{w}}$. The marginal contribution is the instantaneous change in portfolio volatility when the weight of asset i is increased slightly while the weights are held fixed. The marginal contribution is determined $\frac{\partial \sigma_p}{\partial w_i} = \frac{(\boldsymbol{\Sigma}\, \boldsymbol{w})_i}{\sigma_p}$. The risk contribution $RC_i = w_i\, \frac{(\Sigma\, w)_i}{\sigma_p}$ expresses the share of total volatility that is attributable to asset i in volatility units. And in percentage terms contribution is $PRC_i = \frac{RC_i}{\sigma_p} = \frac{w_i (\boldsymbol{\Sigma}\, \boldsymbol{w})_i}{\boldsymbol{w}^\top \boldsymbol{\Sigma}\, \boldsymbol{w}}$. Contributions add to portfolio volatility $\sum_{i=1}^{N} RC_i \; = \sigma_p$. A negative RC_i or PRC_i indicates that asset i hedges the rest of the portfolio through negative covariance with the portfolio. In an equal risk contribution allocation, all PRC_i are equal by construction, whereas in the risk only minimum variance case the first order condition implies equal marginal contributions for the assets that are active, not equal PRC_i. These identities allow interpretation of allocations in risk units and are standard in risk budgeting. (Roncalli, 2013).

These calculations are based on assumptions and acceptable constraints, the validity of which determines when the minimum variance problem has a well-defined solution. With $\boldsymbol{\Sigma} \succ 0$ the unconstrained programme has a unique minimiser. If $\boldsymbol{\Sigma} \succeq 0$ the objective is convex and a minimizer exists, but it may not be unique when the budget hyperplane intersects a flat direction. If $\boldsymbol{\Sigma}$ is singular the program under the budget constraint is ill posed and such cases are not treated in this work. Under long only constraints the feasible set is the simplex, which is compact, so a minimizer exists for any symmetric positive semidefinite $\boldsymbol{\Sigma}$ and uniqueness holds when the restriction of $\boldsymbol{\Sigma}$ to the active support is positive definite. (Boyd and Vandenberghe, 2004.)

4 Data and methodology

This chapter describes the data, its collection, and the empirical procedures used in strategy construction and evaluation. It first formulates the research hypotheses, based on the previous literature and the research questions introduced in Chapter 1. The goal of the study is to develop an adaptive, data-driven equity strategy that combines principal component analysis and Markowitz portfolio selection in accordance with the hypotheses presented. The strategy is tested on S&P 500 constituents during the research period, and its performance is evaluated using pre specified return and risk metrics. All selections are made in advance and applied consistently. The study is conducted using R software The S&P 500 is a float-adjusted, market-capitalization-weighted index of 500 leading U.S. companies and is widely regarded as the primary gauge of large-cap U.S. equities, by design it covers roughly 80% of available U.S. market capitalization, which makes it a representative universe and benchmark for this study. The index was introduced by Standard & Poor's in 1957 and has since become one of the most tracked equity indices globally, reinforcing its suitability for empirical evaluation and implementation realism. (S&P Global, 2025)

4.1 Research hypotheses

To systematically evaluate the effectiveness of the proposed investment strategy, this study formulates three hypotheses. Each hypothesis corresponds directly to one of the research questions and tests a core feature of the strategy: the use of PCA in asset allocation, its integration with portfolio optimization and its performance across different market regimes. Based on the earlier studies presented in this thesis, the hypotheses reflect gaps identified in the existing literature regarding the use of PCA in stock selection and allocation, the combination of PCA with minimum variance optimization and the risk-adjusted performance of such strategies across different market conditions. First, PCA-based risk and covariance management techniques appear to improve portfolio risk profiles, but prior applications often focus on single markets, specific sectors, short evaluation windows or the risk model in isolation and therefore do not provide a clear answer on whether a corresponding PCA–Markowitz approach actually outperforms a simple market index (Sun et al., 2020; Fan et al., 2013; Ledoit and Wolf, 2004). Second, the literature on PCA-based stock selection shows that principal components compress information and can support the selection process, yet it remains unclear whether component information can be used at all to construct systematic scoring rules that guide stock selection and portfolio weights in a consistent way (Zhou and

Yin, 2020; Yang et al., 2016; Cao and Wang, 2020). Third, strategies that combine PCA and minimum variance or related allocation rules have delivered promising or even exceptionally strong risk-adjusted returns in some studies, but the evidence is typically restricted to particular markets, sectors or limited samples, which leaves open whether any potential advantage persists across different market conditions or over longer horizons relative to established benchmark indices (Sun et al., 2020; Fan et al., 2019; Bufalo et al., 2025). These uncertainties and research gaps motivate the formulation of three hypotheses that focus on the efficiency of the strategy, the usability of PCA components as selection rules and the risk-adjusted return profile of the strategy in different market environments.

The first hypothesis corresponds to the primary research question (Q1), which examines whether combining PCA-based factor analysis with Markowitz portfolio theory can be used to construct an efficient and flexible investment strategy that improves on a market capitalization weighted benchmark.

> H1: *Combining PCA-based factor selection with Markowitz's minimum variance optimization yields a more efficient and flexible investment strategy than a market-cap-weighted benchmark portfolio.*

The second hypothesis addresses subquestion Q2, which explores how PCA extracted components can be used in investment allocation. PCA generates asset level loadings that can, in principle, be converted into scores for ranking and selecting stocks, but it is not clear from the existing literature whether such scores can be defined in a systematic and repeatable way that directly guides portfolio construction.

> H2: *The principal components extracted using PCA can be systematically translated into asset scores that guide stock selection and portfolio allocation.*

The third hypothesis relates to subquestion Q3 and examines how the strategy performs across different market environments. A data-driven, adaptive method is expected to outperform a benchmark index in terms of risk-adjusted performance, particularly during volatile periods. By isolating the most influential return-driving factors, PCA is expected to reduce noise and improve risk-adjusted returns.

> H3: *A portfolio constructed using PCA-based factor selection and Markowitz optimization achieves a higher risk-adjusted return than a market-cap-weighted benchmark portfolio.*

These hypotheses will be tested using historical S&P 500 data, applying PCA to extract key risk factors and incorporating them into a Markowitz optimized portfolio. The empirical analysis will assess the performance, stability and robustness of the strategy under different economic conditions.

4.2 Data

The period selected for the study is 4 January 2010 to 31 December 2024. The period covers several market conditions. It includes long upswings, sharp declines, and subsequent recoveries. It is close enough to the present to keep the daily data consistent. The length provides sufficient observations for rolling estimation. It improves comparability across subperiods and reduces the risk of placing too much weight on a single market episode.

As mentioned, the investment universe is the S&P 500. Index constituents change over time. Index membership is handled point in time at each rebalancing. This mitigates survivorship bias. Constituent identifiers and index status were obtained from Refinitiv, and index company changes were reconciled on this basis. Daily adjusted closing prices were obtained from Yahoo Finance. Adjusted prices reflect corporate actions. All time series were aligned to a common trading calendar before return calculation. This section does not present the full list of 500 constituents. Representative examples across sectors include Apple, Microsoft, Amazon, Alphabet, NVIDIA, Berkshire Hathaway, Johnson & Johnson, JPMorgan Chase, Exxon Mobil, and Meta Platforms. The figure below shows the development of the index during the period in question.

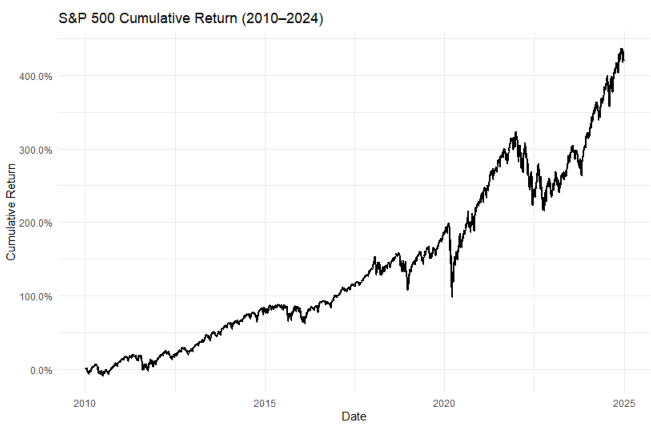

Figure 4 S&P 500 Return 2010–2024

During 2010–2024 the S&P 500 index shows a strong upward trend with several clear declines. The early part of the sample reflects recovery from the global financial crisis and a long expansion. In 2011, as macroeconomic uncertainty increased, a brief correction occurs after which the rise continues. The years 2013–2017 are characterized by steady appreciation supported by broad earnings growth. In late 2018 there is a sharp decline, followed by a recovery in 2019. The steepest negative move is the COVID-19 shock in early 2020. The index falls quickly and then recovers at an unusual pace as policy support and improving expectations take hold. The rise continues throughout 2021. In 2022 the index declines again during a tightening cycle and high inflation. In 2023–2024 the markets recover again.

Daily share returns are calculated from adjusted closing prices for each company included in the index. The simple return for share i on trading day t is:

$$r_{t,i} = \frac{P_{t,i}^{\text{adj}}}{P_{t-1,i}^{\text{adj}}} - 1.$$

Adjusted prices reflect corporate actions such as stock splits and cash dividends. Yahoo Finance does not provide data for many delisted companies, so complete histories are not available for all historical S&P 500 members. The empirical analysis uses a working universe of about 650 companies. In total, the index has included more than 800 distinct companies over its history. All return series are aligned to a common trading calendar before return calculation. Missing or non-finite prices are treated as NA values. Prices equal to zero or negative are treated as errors and set to NA. One day returns are set to NA when either the current or the previous adjusted close is missing.

Later in the calculation, NA values are treated within each rolling window with a minimum data coverage requirement. Assets that do not meet this requirement are excluded from that window. For the remaining assets the matrices used for principal component analysis, and the regressions are formed from complete case dates inside the window, and dates with missing values are dropped. NA values are not forward filled or interpolated.

4.3 Methodology

4.3.1 Overview

The strategy is developed based on the literature presented earlier. The strategy is fully data-driven and walk forward. It forms a factor model from recent co-movement in stock returns using principal component analysis. The component scores are then linked to short horizon returns with a simple OLS regression model that is estimated separately for each stock. The regression output is used to construct a scoring rule that selects a set of 10 stocks. Capital is allocated by solving a minimum variance optimization. The objective is to test whether information in common return movements can be translated into short horizon portfolio choices.

The process runs on a fixed schedule. The portfolio is rebalanced every 14 calendar days. On each rebalance date the method forms a lookback window of recent daily returns that covers about one trading year. The portfolio is formed on that date for the next 14 days, so the rebalance dates always reflect the most recent available data. Hyperparameters are set out of sample before the test and do not change over time.

Factors are created with principal component analysis applied to a shrinkage estimate of the return covariance in the lookback window. The number of components is chosen by a cumulative explained variance rule with an upper bound. Component signs are oriented to maintain continuity across windows. The first component is aligned to move in the same direction as a simple market proxy. These choices stabilize the interpretation of the components through time.

For each stock the method estimates a linear relation between its 14-day cumulative return and the contemporaneous component scores. Fit quality is measured with adjusted R^2. A minimum threshold is used to focus on stocks for which the scores explain a meaningful share of short horizon variation. Selection scores are then computed from three elements. These are the estimated beta coefficients, a short recent mean of the component scores, and weights that reflect the share of variance explained by each component. The ten highest scores form the investable set for the next period. If a stock lacks a valid price on the trade date it is excluded.

Capital is distributed across the ten selected stocks by solving a minimum variance problem under simple constraints. Portfolio weights are constrained to be non-negative and to sum to one. The covariance used in the optimization is the same shrinkage estimate computed in the lookback window, restricted to the selected names. This keeps the estimation and allocation steps consistent.

Evaluation is based on out of sample period returns that follow from the rebalance schedule, that is, the cumulative return over the next 14 calendar days. Missing data are handled within each window through coverage requirements and by using only dates that are complete for the modelling steps. Missing values are not imputed. The next subsection presents the methods in detail.

4.3.2 Investment Strategy

First, the rolling window and symbols to be used later are defined. The portfolio is rebalanced every 14 calendar days, i.e., every two weeks. The rebalancing dates are therefore fixed. On a rebalance date t the method forms a lookback window of recent daily returns that ends at $t-1$. Component scores and portfolio weights are computed using only data available up to $t-1$. The new portfolio is opened on the first trading day after t and held until the last trading day on or before $t+13$. Trades occur on the next available trading session after the computation date.

Let N denote the number of stocks available on the rebalance date. Inside the lookback window let L denote the number of trading days. Then $\mathbf{R} \in \mathbb{R}^{\mathrm{L} \times \mathrm{N}}$ is the matrix of simple daily returns aligned on common trading days. Row u of $\mathbf{R}$ is $\boldsymbol{r}_{\mathrm{u}}^{\mathrm{T}}$. Then $\mathrm{r}_{u,i}$ is the return of stock i on day u. The portfolio is determined at the time of each rebalancing date. Only stocks for which sufficient data are available for the period in question are used in the backward-looking window. Calculations use only trading days that are complete for the retained assets inside each window. Missing values are not imputed. The fixed parameters for calculation are as follows: rebalancing interval 14 calendar days, portfolio size 10 stocks, lookback length 252 trading days (approx. one year), smallest K with cumulative explained variance at least 0.85, with $K \leq 4$, and regression gate adjusted R^2 threshold 0.10. Input data are daily simple returns from adjusted closes. Missing values are not imputed. Inside each lookback window only stocks with at least 80% data coverage are retained

The method for calculating a single holding period is described next. This method is repeated identically for each new period throughout the entire study period. The method uses the fixed parameters defined above.

Next, the calculation method for one holding period is described, which is repeated for each new period throughout the study. A shrinkage covariance estimator, as outlined in the Theoretical Framework, is applied to the lookback window. The sample covariance matrix $\boldsymbol{S}$ is computed directly from returns as:

$$\boldsymbol{S} = \frac{1}{L-1} \boldsymbol{X}^{\top} \boldsymbol{X},$$

where $\boldsymbol{X} = \boldsymbol{R} - \mathbf{1}\,\boldsymbol{\mu}^{\top}$, and $\boldsymbol{\mu}$ is the vector of column means of return matrix $\boldsymbol{R}$. Then the shrunk covariance matrix $\widetilde{\boldsymbol{\Sigma}}$ is obtained by shrinking the window covariance toward a structured target. The estimator is:

$$\widetilde{\boldsymbol{\Sigma}} = (1 - \delta)\,\widehat{\boldsymbol{\Sigma}} + \delta\,\boldsymbol{F},$$

where $\boldsymbol{F}$ is a target matrix with zero off-diagonal correlations and a pooled variance level on the diagonal. And $\delta \in [0,1]$ is the shrinkage intensity determined from the window. This $\widetilde{\boldsymbol{\Sigma}}$ is used to extract principal components next and later it is restricted to the selected stocks for the portfolio allocation.

The decomposition of the principal component analysis is:

$$\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\top},$$

where $\boldsymbol{\Lambda} = diag(\lambda_1, \dots, \lambda_N)$ with $\lambda_1 \geq \cdots \geq \lambda_N$, and $\boldsymbol{Q} = [\boldsymbol{q_1}, \dots, \boldsymbol{q_N}]$ is orthonormal. In this study, they are not rotated. The diagonal matrix $\boldsymbol{\Lambda}$ contains the component variances that order the importance of each direction over the current window, and the columns of $\boldsymbol{Q}$ give orthonormal directions in asset space that describe common co-movement patterns. In this study, they are not exchanged. They provide the coordinates used to form the time series of component scores that are fed into the regression phase.

The retained dimension $\boldsymbol{K}$ is the smallest value such that the cumulative explained variance reaches at least 0.85 with the cap $\boldsymbol{K} \leq 4$. Explained variance is calculated:

$$EV_K = \frac{\sum_{j=1}^{K} \lambda_j}{\sum_{j=1}^{N} \lambda_j} \geq 0.85.$$

Since the result of the solution is equally correct when $q = -q$, without an anchoring rule the component directions may vary between windows, which would reverse the scores and weaken comparability over time. To stabilize interpretation over time, eigenvector signs are oriented. The first direction is aligned to a market proxy,

$$s_1 = \text{sign}(\text{corr}(\mathbf{R}\boldsymbol{q_1}, \boldsymbol{r}^{mkt})),\ \ \boldsymbol{q_1} \leftarrow s_1\boldsymbol{q_1},$$

and $j \geq 2$ the sign is chosen for continuity with the previous window using cosine similarity on the common symbol set,

$$s_j = \text{sign}\big(\boldsymbol{q}_j^{\top}\,\boldsymbol{q}_j^{prev}\big),\ \ \boldsymbol{q}_j \leftarrow s_j\boldsymbol{q_j}.$$

Daily component scores are then calculated directly from the returns: $\boldsymbol{Z} = \boldsymbol{R}\boldsymbol{Q}_K$, where $\boldsymbol{Q}_K = [\boldsymbol{q}_1, \dots, \boldsymbol{q}_K]$, aligned on complete trading days inside the window. A ten-day mean of scores is applied later only for ranking. For later use in the selection score, explained-variance weights are the variance shares of the retained components,

$$w_j^{\mathrm{EV}} = \frac{\lambda_j}{\sum_{l=1}^{N} \lambda_l},\ j = 1, \dots, K.$$

Next, a regular OLS regression is estimated for each stock. The dependent variable is the 14-calendar day forward cumulative return. With t as the computation date, the window runs from the first trading day after t to the last trading day on or before $t + 13 = u$. The cumulative 14-calendar day return is:

$$R_{t,i}^{(14)} = \prod_{k=t+1}^{u} (1 + r_{k,i}) - 1.$$

Regression analysis is then applied. The regression of stock i is:

$$R_{t,i}^{(14)} = \alpha_i + \sum_{j=1}^{K} \beta_{i,j}\, z_{t,j} + \varepsilon_{t,i},$$

where $z_{t,j}$ is the day t score extracted from the matrix $\boldsymbol{Z}$ for component j. Stocks whose score-return fit shows little explanatory power based on adjusted R^2 are filtered out. The threshold for adjusted R^2 is at least 0.1. If fewer than ten pass, the remaining places are filled by ranking all stocks by adjusted R^2. The gate is only applied to the current window, meaning that no information outside the window is used.

Next, a scoring system is created to select the stocks for allocation. The score combines the fitted regression coefficients, β, with the current component state and applies an explained-variance weight for each component. Using explained-variance weights emphasises directions that account for most co-movement in the window and down-weights low-variance directions. This stabilizes the ranking, reduces sensitivity to estimation error in β coefficients in the scores, and focuses selection on dominant structures. The trade-off is that predictive information in lower-variance components receives less weight. However, the assumption is that factors with greater variance dominate the informational value of return dynamics. The current state is the ex-post mean of daily component points over the last ten trading days, approximating a 14-calendar-day cycle. The average is computed as $\bar{z}_{t,j} = \frac{1}{10}\sum_{k=0}^{9} z_{t-k,j}$. Each stock receives a score

$$S_i = \sum_{j=1}^{K} \beta_{i,j}\ \bar{z}_j\ w_j^{\mathrm{EV}}$$

and the ten highest-scoring names are retained for allocation.

Now the 10 highest-scoring stocks have been selected, and capital will be allocated to them using a minimum-variance portfolio. Optimization is performed on a shrunk covariance from the current lookback window restricted to the selected stocks. $\widetilde{\boldsymbol{\Sigma}}_S$ denotes this covariance matrix. The solution to the quadratic optimization problem is the weight vector $\mathbf{w}$ for the selected stocks. Weights are long-only and fully invested. The optimization problem is

$$\min_{\boldsymbol{w}\in R^{10}} w^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_S\boldsymbol{w} \quad \text{s.t. } 1^{\mathrm{T}}\boldsymbol{w} = 1,\ \ w \geq 0.$$

The optimization solution has been presented earlier in this work. If $\widetilde{\boldsymbol{\Sigma}}_S$ is not strictly positive definite, a small diagonal ridge is added to ensure feasibility: $\widetilde{\boldsymbol{\Sigma}}_S^{+} = \widetilde{\boldsymbol{\Sigma}}_S + \varepsilon\,\mathbf{1},\ \ \varepsilon > 0$. At the same time the ex-ante portfolio risk is computed as $\sigma = \sqrt{\boldsymbol{w}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_S\boldsymbol{w}}$. A simple historical average provides an annual return estimate that is only diagnostic and is not used for optimization. To test the performance of the strategy, the same process is repeated throughout the research period.

4.3.3 Performance and metrics

This subsection presents the portfolio performance evaluation methods and the meaning of each metric. Performance is measured on a grid of nonoverlapping 14-calendar-day periods aligned to the rebalancing dates. Period returns are compounded to form a cumulative equity curve. Annualization is calendar-based with approximately twenty-six periods per year. A constant risk-free rate of 4% per annum is used and converted to the period level. The benchmark is the S&P 500 aligned to the same period boundaries, so active returns are directly comparable.

The reported metrics summarize level, risk, and consistency. The cumulative return shows the evolution of capital implied by the period returns. The annualized return is the geometric growth rate mapped to one year and reflects long-run compounding. The annualized volatility scales the dispersion of period returns to the yearly horizon and serves as total risk. The Sharpe ratio uses the period risk-free rate and evaluates excess return per unit of total risk on an annualized basis. A higher value indicates better risk-adjusted performance. Zero implies performance equal to the risk-free rate, while a negative value indicates underperformance. (Sharpe, 1966.) A Sharpe of 1 means the expected annualized excess return equals annualized volatility and a Sharpe of 2 means twice that. Generally, a Sharpe ratio of > 1.0 is considered good, and between 0.5 and 1.0 is considered reasonable. Ratios below this are considered modest or weak. A rolling Sharpe over roughly one year provides a compact view of time variation in risk-adjusted performance. (Fernando, 2025.)

Benchmark and excess series are computed on the same grid. The information ratio reports annualized active return per annualized tracking error (Grinold & Kahn, 2000). CAPM alpha, α, and beta, β, are estimated with OLS regression on excess period returns. Alpha measures abnormal performance after controlling for market exposure. A positive alpha indicates value added beyond the benchmark. A negative alpha indicates a shortfall. A beta near one implies market like risk. A beta below one indicates a defensive profile. A beta above one indicates an aggressive profile. Rolling alpha is estimated on a trailing window of about twenty-six periods and reported on an annual basis. Persistent positive values with confidence intervals above zero indicate robust outperformance. Values varying around zero indicate that there is no stable edge. The section reports R^2, statistical significance, and a rolling alpha to assess stability through time. (Jensen, 1968.)

In addition to these, the maximum drawdown and the entire drawdown path are presented, illustrating the drawdown trend from peaks to troughs. The distribution of period returns is presented in the descriptive statistics table. The table shows the mean, median, minimum, maximum, skewness, kurtosis, and number of periods. The calculation method for these is not presented separately.

Period returns and portfolio weights follow the earlier definitions. The equity curve starts at one. The cumulative return CR at evaluation date T is the compounded product of period returns minus one:

$$CR_T = \prod_{k=1}^{T}(1 + R_{p,k}) - 1,$$

where k is period index and $R_{p,k}$ is the portfolio return in period k. The benchmark index return has been calculated using the same principle. In practice, the logic behind calculating returns is expressed simply as follows. With a 14-calendar day holding window and the backtest anchored to Monday 4 January 2010 (first trading day of 2010), weights are estimated on every second Monday using only information that ends on the previous trading day, typically the previous Friday adjusted close. Positions are opened at the adjusted close on t and closed at the adjusted close of the last trading day on or before t+13, which is normally the Friday of the following week. The next cycle begins at $t + 14$ under the same rules, and if an exchange holiday falls on t or on the terminal day the trade is entered or exited on the next open trading day.The calendar factor used in annualizing the figures presented is $A = 26$, if the calculation is based on daily returns $A = 252$.

Drawdowns are computed from the equity curve. The drawdown at date t is the fall from the running peak. The maximum drawdown is the worst value over the sample. Duration is the longest time spent below a prior peak until a full recovery. The maximum drawdown is reported in the summary table. The full drawdown path is shown in a separate chart.

The Sharpe ratio is reported on an annual basis. A rolling Sharpe over about one year is shown as a time series for stability. The Sharpe ratio for one period is calculated as follows:

$$Sharpe_{ann} = \frac{\bar{X}}{\sqrt{A}\sigma_p}, \qquad \bar{X} = \frac{1}{T}\sum_{k=1}^{T}(R_{p,k} - r_{f,per}).$$

Active returns, A_k, are the portfolio period return minus the benchmark period return. Tracking error is the dispersion of active returns. The information ratio reports annualized active return per annualized tracking error:

$$\text{IR} = \frac{A\overline{A}}{\text{TE}_{\text{ann}}}, \qquad \overline{A} = \frac{1}{T}\sum_{k=1}^{T} A_k.$$

Alpha and beta are estimated by OLS on excess period returns on the same 14-day grid. Excess returns subtract the period risk-free rate. The intercept is alpha. The slope is beta. Alpha is reported per period and annualized. A rolling alpha is plotted to assess stability. Confidence intervals use the OLS standard error. The regression is formed as follows:

$$X_{p,k} = \alpha + \beta, X_{m,k} + \varepsilon_k,$$

where $X_{p,k} = R_{p,k} - r_{f,\text{per}}$ and $X_{m,k} = R_{m,k} - r_{f,\text{per}}$. $X_{p,k}$ is excess return in period k and $X_{m,k}$ is excess return of the benchmark index of period k.

4.3.4 Limitations of the study

The model includes certain limitations stemming from the assumptions underlying the model and its practical implementation. The investable universe differs from the true set because of data gaps and missing delistings. This can bias results upward if weak delisted names are absent. Estimates are subject to sampling error and parameter choices. Results depend on the window length, the number of components, the shrinkage level, and the adjusted R^2 gate. Regression analysis also involves assumptions. It requires linearity, exogeneity of explanatory variables, homoscedastic and uncorrelated residuals, and normality for statistical inference. Stock returns typically do not fully satisfy these assumptions. In this study, the validity of the assumptions is not systematically tested. Robust standard errors and corrections for heteroscedasticity or autocorrelation are omitted. Regressions serve as a filter and a descriptive tool, and screening is performed only using the adjusted R^2 threshold. Outliers may affect the coefficients. The explanatory variables are PCA components and are orthogonal by construction, which mitigates multicollinearity but does not remove these limitations. Selections are designed to correspond as closely as possible to the chosen holding period. The limits are set to avoid unduly restricting eligible assets and to exclude names with incomplete or insufficient data. The risk-free rate is fixed at 4% per annum. Market rates vary through time, so excess returns are approximate. The impact of taxes is not considered. Taxation depends on the legislation of different jurisdictions and on the legal form of the trading entity. Net results therefore depend on the investor's circumstances. The strategy is long only. The absence of shorting limits hedging and may increase residual risk.

Transaction costs, bid and ask spreads, and slippage have not been modelled. They vary depending on the provider, trading venue and capital, so their impact is case specific. Slippage arises from execution timing. The model assumes execution at the closing price on the rebalance date, but in practice this price is often unattainable. In live trading these frictions reduce returns, especially when turnover is high. Liquidity constraints are not modelled. The default assumption is that all companies included in the index have sufficient liquidity. In practice, large assets under management can create capacity constraints and liquidity pressure. Capacity is here assumed to be unlimited. A concentrated top ten portfolio is hard to scale, and large assets under management raise market impact and constrain position sizes.

5 Results

This chapter presents the empirical results for the strategy from 4 January 2010 to 31 December 2024. Figures and tables summarize portfolio composition and ex-ante risk, cumulative and period returns, drawdowns, benchmark-relative performance, risk-adjusted metrics, and CAPM attribution. All quantities are computed on the non-overlapping 14 calendar day grid defined earlier and aligned with the S&P 500 on the same dates. Results are reported as levels and as differences relative to the benchmark. Rolling estimates are shown where useful to assess stability through time. The objective is to document realized behavior and evaluate economic and statistical significance as defined earlier.

5.1 Portfolio performance

Each period the portfolio selection comprised ten stocks, although some could receive a zero weight. Zero weights arise from the minimum variance optimization. The composition of the most recent portfolios is shown in Appendix 1. Over the full sample from 4.1.2010 to 16.12.2024 the portfolio held on average 8 names, and the average weight per stock was 10%. In the weight distribution the first quartile is 1.1%, the third quartile 14.9%, and the median weight 7.78%. Between consecutive periods the portfolio retained on average three names, so about seven names changed every two weeks. The average annual ex ante expected return of the portfolio was 21.23% with a risk of 20.27%. The maximum expected annual return was 317.9% and the minimum −32.6%. The largest ex ante volatility was 66.72% and the smallest 8.1%. On the two-week horizon the corresponding figures are 0.82% for expected return and 3.97% volatility.

The distribution of weighted beta values used in the stock scoring for the last rebalancing window is shown in the next figure. Based on the figure, the medians deviate from zero: PC1 is negative and PC2 positive, while PC3 and PC4 lie closer to zero. The medians of the first and second components are clearly further from zero than those of the third and fourth, which suggests a more consistent and informative EV weighted exposure to these directions in the selection.

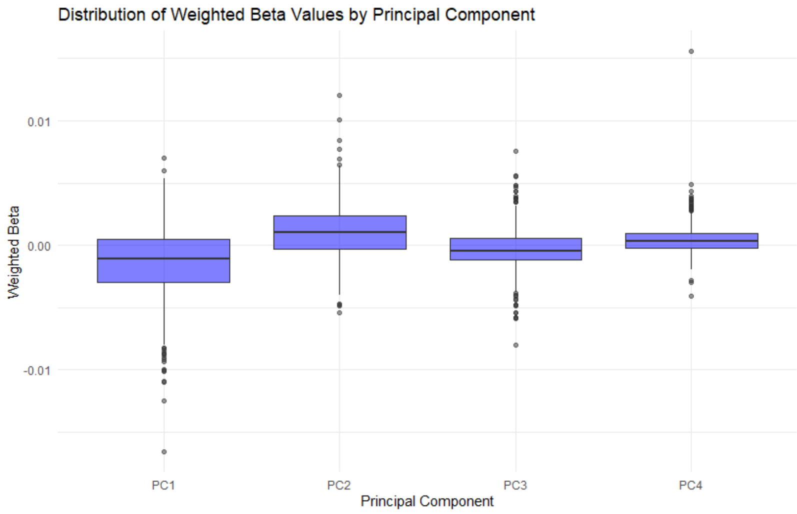

Figure 5 Distribution of Weighted Beta Values

The next figure shows the distribution of S_i- scores in the last rebalancing window. The distribution is centered close to zero with a slight right skew. Most observations lie around a small positive region, but there is a sparse right tail with clearly positive values. This is consistent with the selection logic, which picks the next period's portfolio from the upper tail of the distribution (the top ten by the score). Because the absolute magnitudes are small due to EV and scale weights, the natural interpretation is cross sectional: the score serves as an ordering within the period rather than as a large absolute forecast across time.

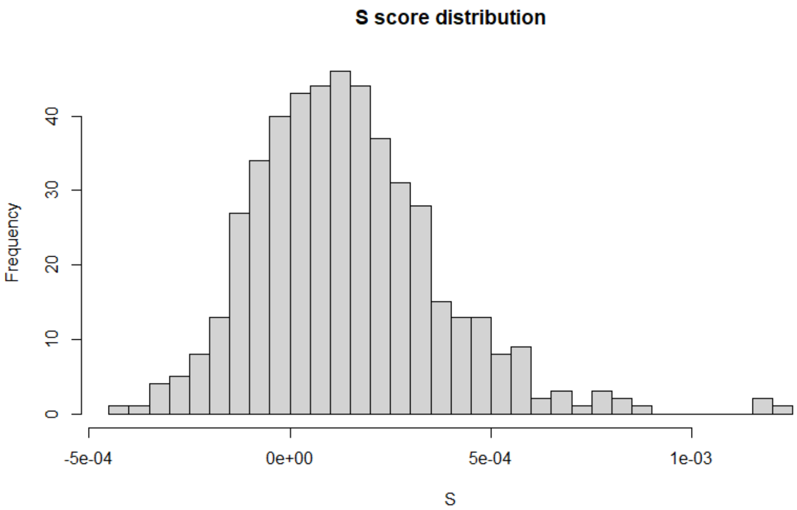

Figure 6 S Score Distribution

The next figure shows the portfolio's cumulative return over the full sample period.

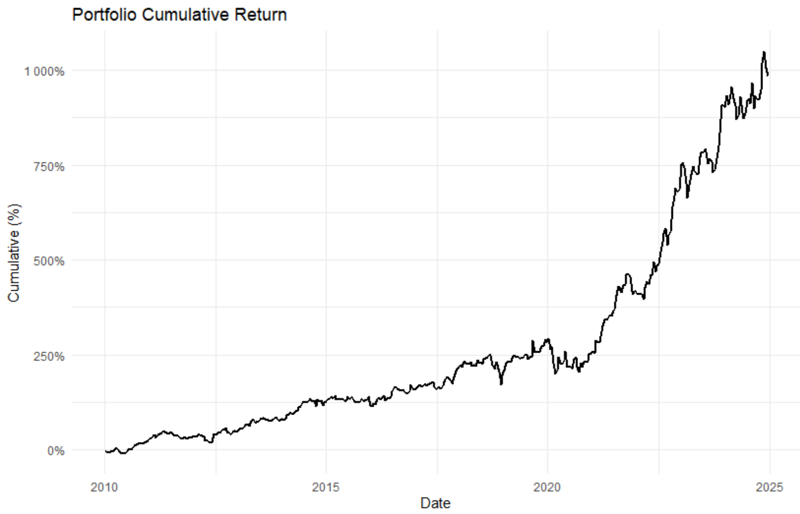

Figure 7 Portfolio Return

The cumulative path is upward sloping but includes several drawdown episodes. Progress was steady in the early years, with brief declines around 2011 and 2015. A deeper fall in early 2020 was followed by a rapid recovery and a strong advance during 2020–2021. Further pauses and a clear drawdown occurred in 2022. The curve then moved higher again and finished near its highs at the end of 2024.

A clearer view is obtained when the portfolio is shown together with the S&P 500.

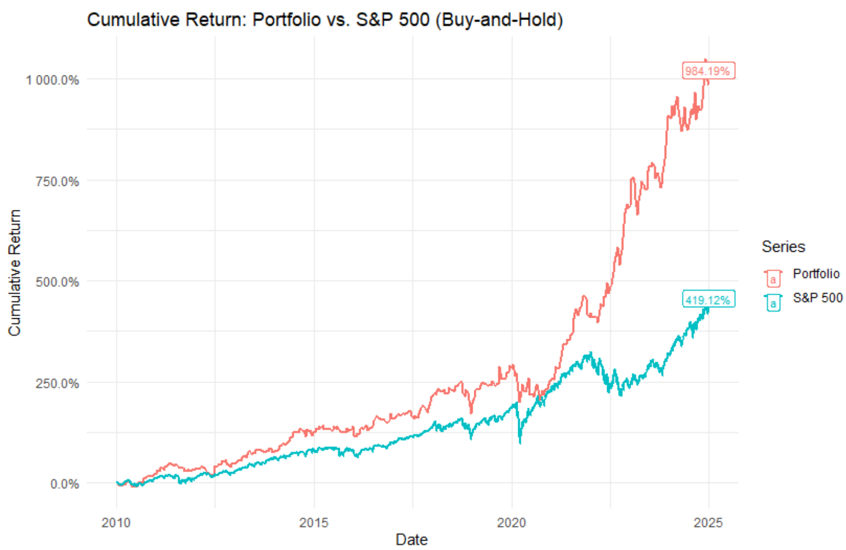

Figure 8 Portfolio and S&P 500 Return

The portfolio ended at a higher cumulative level than the index, close to 984% versus about 419% for the S&P 500. The gap widened most during 2020–2021 and remained wide thereafter. The figure also shows that the portfolio moved broadly with the index. During upswings the strategy tended to deliver higher gains, while drawdowns were also deeper.

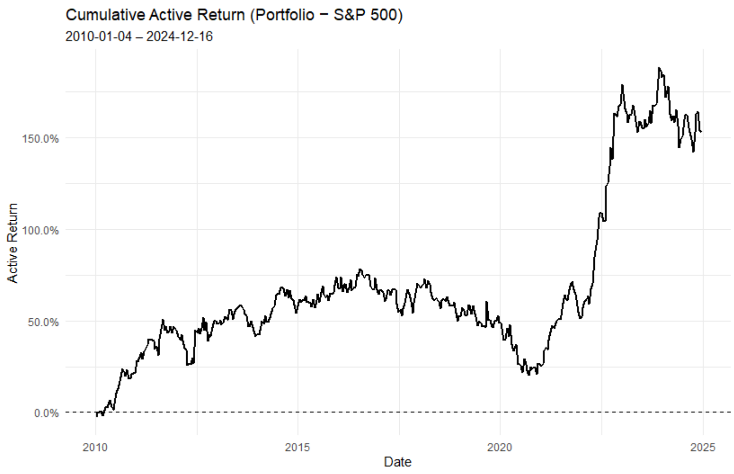

Figure 9 Cumulative Active Return

The cumulative active return confirms these observations. Cumulative active return is the compounded performance of the portfolio minus the S&P 500 on the same 14-day grid. The series starts at zero. Above zero means the portfolio has added value versus the index since inception. Below zero means a cumulative shortfall. When the line rises, the portfolio outperforms the index in that period, the steeper the rise, the stronger the outperformance. When the line falls, the index outperforms. Flat sections indicate similar returns. Cumulative active returns rose gradually in the first half of the sample, moved sideways and dipped before 2020, then increased sharply during 2021–2022. Thereafter it fluctuated around a high level and stayed well above zero.

The next table reports period-level descriptive statistics for the strategy and the S&P 500 on the same 14-day grid.

Table 1. Descriptive statistics

Period specific key figures for full sample period

	Portfolio	**S&P 500**
Mean	0.0068	0.0041
Median	0.0071	0.0063
Minimum	-0.1189	-0.1235
Maximum	0.1564	0.0887

	Portfolio	S&P 500
Std. Dev.	0.0368	0.0287
Skewness	-0.0166	-0.9116
Kurtosis	1.3779	2.8306
Observations	391	391

The portfolio's 14-day mean return is 0.68% and the median is 0.71% across 391 periods. The corresponding figures for the index are 0.41% and 0.63% over 391 observations. Dispersion is higher for the portfolio. The standard deviation is 3.68% versus 2.87% for the index. Extremes range from -11.89% to 15.64% for the portfolio and from -12.35% to 8.87% for the index.

The shapes of the distributions also differ. Portfolio skewness is close to zero (-0.02), which indicates an approximately symmetric period return distribution. Large gains and losses occur with similar frequency and magnitude. The index is clearly more negatively skewed (-0.91), which points to more frequent or larger negative outliers. Excess kurtosis is 1.38 for the portfolio and 2.83 for the index, indicating fatter tails for the benchmark. This means that although the portfolio varies more in general, extreme periods are relatively more common for the index, particularly in the left tail.

Table 2. Annualized descriptive statistics
Key figures for full sample period

	Portfolio	S&P 500
Cumulative Return	9.8419	4.1912
Annualized Mean Return	0.1717	0.1162
Annualized Volatility	0.1875	0.1723
Sharpe Ratio (ann.)	0.7318	0.4300

	Portfolio	S&P 500
Max. Drawdown	-0.2328	-0,3392

Table 2 summarizes full-period performance. The portfolio delivered an annualized geometric return of 17.17 % with an annualized volatility of 18.75. The annualized Sharpe ratio is 0.73 as defined earlier. The Sharpe ratio uses annualized period excess returns and the period standard deviation; it is not computed from the geometric rate. A 4 % annual risk-free rate is applied. The final cumulative return is 984.19 % and the maximum drawdown is -23.28 %. The index figures are computed from daily returns. The index Sharpe is calculated with the same 4% annual risk-free rate converted to a daily rate, which lowers the value relative to figures that assume ~0 % or realized short rates that were close to zero for much of the sample. For the S&P 500 the corresponding figures are 11.62 % for return, 17.23 for volatility, and 0.43 for the Sharpe ratio, with a cumulative return of 419.12 % and a maximum drawdown of -33.92 %. Based on these, the portfolio achieved a higher return with slightly higher risk and a modestly higher Sharpe. The drawdown comparison shows that the largest decline in the portfolio from peak value to bottom value is smaller in the sample.

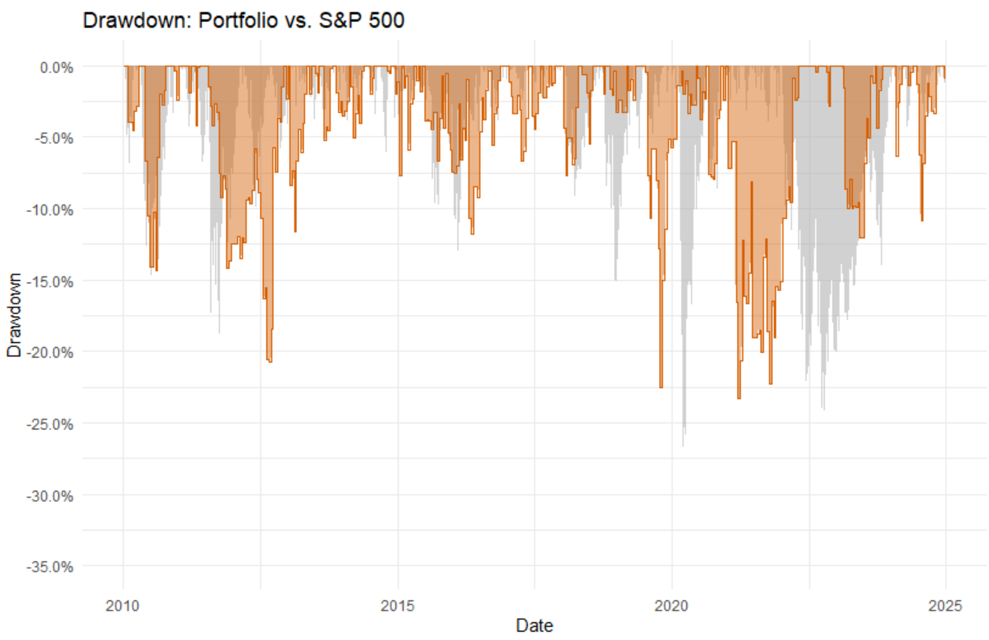

Figure 10 Drawdowns

Drawdowns are measured as the fall from the running peak. The portfolio's maximum drawdown was −23.28%. The S&P 500 reached −33.92%. Depth was therefore smaller for the strategy. Time under water differed as well. The longest spell for the portfolio was 35 periods (≈ 70 weeks). For the index the longest spell was 513 trading days (≈ 103 weeks).

Median time under water was 3 periods (≈ 6 weeks) for the portfolio and 3 trading days for the index. Durations are not directly comparable because the portfolio uses a non-overlapping 14-day grid whereas the index is measured daily. The five deepest portfolio episodes were:

Table 3. 5 deepest episodes for portfolio

From	To	Depth %	Length Weeks	To Trough Weeks	Recovery Weeks
2020-01-20	2021-03-15	-23.3 %	62	8	54
2018-09-17	2019-06-24	-22.5 %	42	14	28
2011-05-09	2012-08-27	-20.8 %	70	56	14
2010-04-26	2010-08-30	-14.4 %	20	10	10
2021-11-08	2022-05-23	-12.1 %	30	18	12

For comparison, the five deepest episodes for S&P 500:

Table 4. Deepest episodes for S&P 500

From	To	Depth %	Length Weeks	To Trough Weeks	Recovery Weeks
2020-02-20	2020-08-18	-33.9 %	25	5	21
2018-09-17	2019-06-24	-22.5 %	103	39	64
2011-05-09	2012-08-27	-20.8 %	29	13	16
2010-04-26	2010-08-30	-14.4 %	41	22	20
2021-11-08	2022-05-23	-12.1 %	27	10	17

The deepest episodes broadly coincide with market regimes. The same windows appear in both series: the 2010 correction, the 2011 euro-area stress, Q4-2018, the Covid shock in 2020, and the 2021–2022 bear phase.

The portfolio's maximum drawdown was −23.3 % versus −33.9 % for the S&P 500. Episode by episode, the portfolio was shallower in 2010 2020, and 2021–2022. It was deeper in 2011–2012 and in Q4-2018. Troughs formed quickly in the 2020 shock for both series (portfolio ~8 weeks, index ~5 trading weeks). In 2011–2012 the portfolio slid more slowly to the low (~56 weeks) compared with the index (~22 trading weeks), which indicates a prolonged decline rather than a crash-type episode for the strategy.

After the 2010 fall the portfolio recovered in ~10 weeks, faster than the index (~17 weeks). In Q4-2018 the portfolio recovery (~28 weeks) was slower than the index (~16 weeks). In 2020 the portfolio fell less but returned to the prior peak more slowly (~54 weeks) than the index (~21 trading weeks). In 2021–2022 the portfolio again showed a shallower trough and a faster recovery (~12 weeks) versus the index (~64 trading weeks). Across the full sample the longest portfolio spell under water was 35 periods (~70 weeks). The index recorded 513 trading days. Durations are not directly comparable because the portfolio is measured on a non-overlapping 14-day grid while the index is daily.

Downside capture is ~84%, which indicates that in down periods the strategy lost about four-fifths of the index loss on average. Upside capture is ~108%, so gains in up periods were slightly larger than the index. This asymmetry matches the drawdown overlays: the orange layer often remains inside the grey band during stress and lifts more in the recovery.

Taken together, the largest falls tend to occur at the same times as the market. The portfolio generally shows shallower troughs in crash-type episodes (2010, 2020, 2022) but can be deeper or more protracted in regime-driven slowdowns (2011–2012, Q4-2018). Recovery speed varies by episode, yet time under water is typically shorter on the portfolio grid, and the capture profile indicates a defensive down-market stance with strong participation in rebounds.

Benchmark-relative performance is evaluated on the same 14-day grid. The annualized active return is 7.10 %. The annualized tracking error is 13.64 %. The information ratio is 0.52. Tracking error is the dispersion of active returns on the 14-day grid, annualized as defined earlier. The information ratio is annualized active return divided by annualized tracking error.

It reports excess return per unit of active risk. These metrics are benchmark-relative and do not depend on the risk-free rate.

Beat-rate statistics give a complementary view. The share of periods with $R_p > R_{SP500}$ is 51.9 % overall. Conditional rates are 50.0 % when the S&P 500 rose on the period and 55.3% when it fell. The rolling one-year beat rate hovered around 50 % for most of the sample, dipped below 40 % around 2019, climbed to about 70-75 % during 2021-2022, and finished near 50 %. The conditional pattern is consistent with the capture profile reported earlier: performance was more resilient in down markets while still participating in up markets. The one-year rolling beat rate is shown in the next figure.

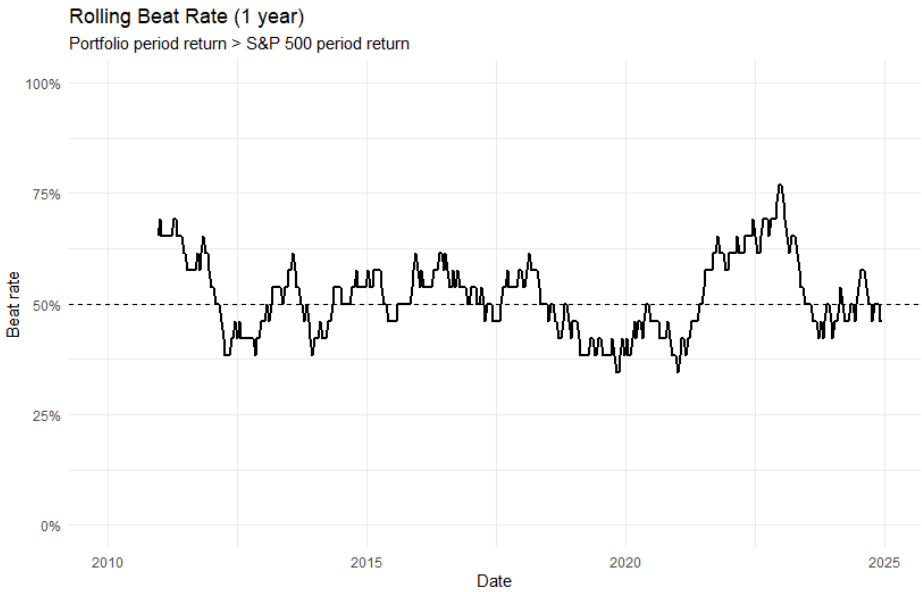

Figure 11 Rolling Beat Rate

The results indicate persistent outperformance delivered with a moderate active risk budget. An information ratio near 0.5 and an overall beat rate slightly above 50 % are consistent with a strategy that adds value relative to the index while avoiding large deviations from it.

CAPM attribution is presented next to assess alpha, beta, and stability through time. CAPM on excess period returns yields a period alpha of 0.0030. The annualized alpha is 7.85 % (arithmetic) and 8.16 % (geometric). Arithmetic scales the period alpha linearly to one year. Geometric compounds the period alpha to one year. For small positive α values the figures are close, with the compounded value slightly higher. For volatile multi-period returns the geometric rate is usually below the arithmetic mean.

The beta is 0.89, which indicates a defensive profile relative to the market. The R^2 is 0.48. Alpha is statistically significant with t = 2.24 and p = 0.026. Beta is highly significant with t = 18.88 and p < 0.001.

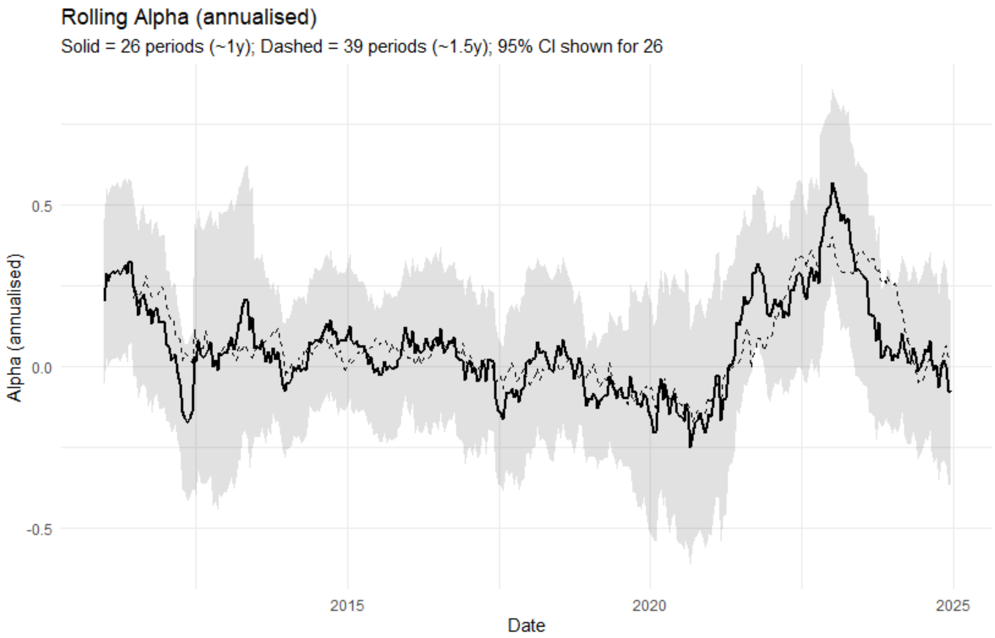

Figure 12 Rolling Alpha

The rolling alpha (annualize, arithmetic) indicates temporal variation. The values fluctuate around zero in the early years. There is a longer negative period before 2020. In 2021–2022, there is a continuous positive cluster, with a 95 % confidence interval above zero for a long period. The readings return to zero towards the end of 2024. This model suggests that the positive alpha throughout the period was due to separate phases rather than a steady change.

In summary, the CAPM attribution shows a positive and statistically reliable alpha, with beta below one and explanatory power being moderate. The rolling view supports this interpretation by showing periods where alpha is clearly positive and a return to neutrality at the end of the sample. However, it cannot be concluded that the strategy would work in all market environments, generating excess returns.

The full-period Sharpe ratio for the portfolio is 0.73, computed with a fixed 4 % annual risk-free rate as defined earlier.

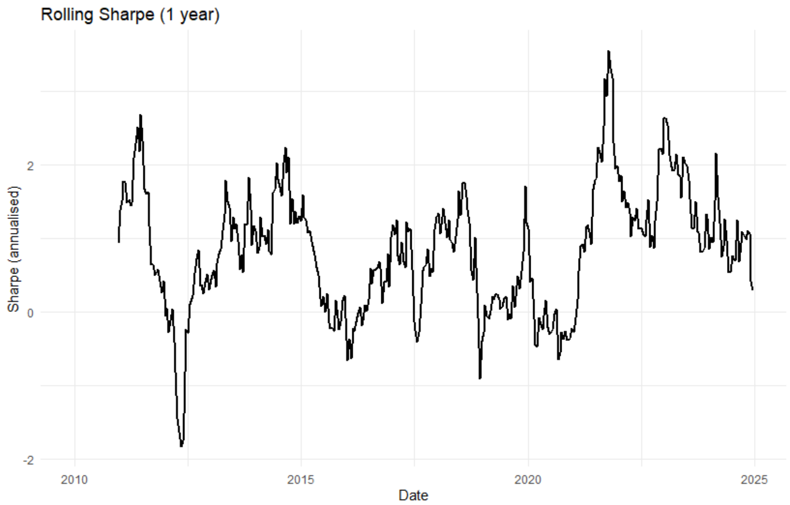

Figure 13 Rolling Sharpe

The rolling Sharpe ratio is presented in a 26-period window and annualized. The series is positive for most of the sample. Its behavior follows other metrics. The ratio turns negative around the time of the 2011-euro area stress and falls again during 2015. A larger decline can also be seen at the end of 2018. The shock in early 2020 causes a sharp decline, which quickly reverses and is followed by a strong positive cluster in 2020–2021. The decline in 2022 drives the ratio lower, after which the values return to a moderate positive range until the end of 2024. The pattern is consistent with the equity curve and drawdown overlay. When the portfolio declines, the rolling Sharpe ratio tends to decline or turn negative. During the recovery, it rises. This indicates that the risk-adjusted return was clustered rather than flat over time. In summary, the rolling profile shows that the risk-adjusted returns were concentrated. The strong periods were concentrated on the recovery after 2020. The weak periods coincide with the largest declines. Over the entire period, the process generated a positive long-term Sharpe ratio, but the result varied over time and was not stable.

6 Conclusions

This thesis evaluated a data-driven equity strategy that combines rolling principal component analysis for stock selection with minimum variance portfolio construction. The aim was to assess whether PCA-based selection together with Markowitz optimization delivers economically meaningful and robust performance relative to a broad market capitalization benchmark and to explain the economic forces behind the realized behavior. All conclusions refer to the studied sample and should not be generalized to other periods or markets without further evidence, since rolling alpha and Sharpe were negative in several subperiods.

The study was motivated by the observation that equity return co-movement and cross-sectional dependence vary over time, which can weaken static allocation rules and complicate risk control during market stress. PCA provides a parsimonious way to summarize common variation in returns, while minimum variance allocation translates the estimated dependence structure into an explicit risk focused weighting rule. A rolling implementation updates both the selection signal and the risk estimates through time, which is consistent with the premise that the underlying dependence structure is not stable. In addition, return patterns and factor exposures may change over time, which motivates protocols that update inputs and exposures rather than treating them as persistent (Lo, 2017). Despite extensive research on PCA in finance and on minimum variance portfolios, there is less evidence on a transparent rolling design that links PCA-based selection directly to minimum variance allocation in a broad large cap universe.

At the beginning of the thesis, three research questions were formulated and three hypotheses were derived from them. The conclusions are summarized by answering the three research questions through the corresponding hypotheses and by highlighting the practical relevance of the results. The first and principal research question was:

Q1: *Does combining PCA-based factor analysis with Markowitz's portfolio theory enable the construction of an efficient and flexible investment strategy?*

Based on this research question, the first hypothesis was:

H1: *Combining PCA-based factor selection with Markowitz's minimum variance optimization yields a more efficient and flexible investment strategy than a market-cap-weighted benchmark portfolio.*

On the sample studied this hypothesis is supported. The developed combination outperformed the benchmark index in terms of risk efficiency throughout the entire period and provided more stable downside protection during significant declines. However, returns varied between different subperiods and the advantage was not consistent. Changing the time period could have resulted in a different outcome.

Efficiency follows from two linked mechanisms. PCA organizes the co-movement structure into a small set of directions that are less noisy than individual names. This improves selection by concentrating exposure on the most informative components and by avoiding dispersion that does not carry persistent signal. Minimum variance then converts that selection into realized risk control by allocating capital to reduce total variance under the estimated covariance. In practice this meant lower vulnerability when correlations rose and a more even distribution of weights when single names became volatile. The joint effect is stronger than either step alone because selection shapes the opportunity set and allocation turns it into realized behavior.

Flexibility in this context means the ability of the strategy to adjust selection and allocation as the joint distribution of returns changes while preserving its stated objective. It arises because rolling PCA updates the feature space when co-movement shifts and minimum variance weights respond to the same changes in the estimated covariance. In the results this flexibility appeared as a positive average active return on the full sample and a period level beat rate that exceeded an even split. Outperformance clustered around regime shifts and recoveries, while downside capture remained lower than the index during major drawdowns with adequate participation in rising markets. In practical implementation, the current strategy requires more monitoring and rebalancing, but it concentrates capital in up to ten securities that the model ranks above the index at that time, which increases the potential for targeted exposure but also the risk of concentration. Therefore, it may be suitable for investors who accept the additional operational work and concentration in exchange for a potential improvement in risk-adjusted returns, while a market-cap-weighted index is a simpler option for those who prioritize low workload and very broad diversification.

The next research question was:

Q2: *How can the key factors identified through PCA be utilized in investment allocation?*

And the second hypothesis was:

H2: *The principal components extracted using PCA can be systematically translated into asset scores that guide stock selection and portfolio allocation.*

On the sample studied this hypothesis is supported. PCA-based scores provided a coherent basis for selection and, when combined with minimum variance weighting, produced positive risk-adjusted performance relative to the benchmark over the full period. Rolling alpha and Sharpe turned negative in several subperiods, so this conclusion concerns this sample.
In this context PCA captures the dominant directions of co-movement and reduces name specific noise, while minimum variance turns the selection into realized risk control. This explains why outperformance and underperformance appeared in clusters rather than uniformly through time.

The last and third research question was:

Q3: *How well does the proposed strategy perform across different economic conditions based on historical backtests?*

And hypothesis based on that:

H3: *A portfolio constructed using PCA-based factor selection and Markowitz optimization achieves a higher risk-adjusted return than a market-cap-weighted benchmark portfolio.*

On the sample studied this hypothesis is partially supported. The strategy delivered higher risk-adjusted performance on the full period, with a higher Sharpe, a positive alpha, and a beta below one. Measured by return alone, the strategy also outperformed the benchmark clearly over the whole sample. Results varied across subperiods, and the advantage was not persistent. Rolling alpha and Sharpe turned negative in several phases and moved close to zero toward the end of the sample. The conclusion therefore applies to this sample only and should not be generalized to other periods or markets without further evidence.

Across conditions the pattern was regime dependent. Performance improved when dispersion across stocks was elevated and when correlations rose from calm levels, since PCA concentrated exposure on the dominant directions of co-movement and minimum variance translated that structure into realized risk control. Downside episodes showed smaller losses and more stable behavior, while recoveries often produced periods of outperformance. Long and narrow rallies led by a small group of large names were more challenging, since concentration in the benchmark moved ahead faster than a diversified minimum variance allocation. Rotation phases required the selection to refresh, and the benefits did not materialize instantly. These features are consistent with a strategy that adapts to the evolving covariance structure and delivers its edge in clusters rather than uniformly through time.

The evidence in this thesis documents the behavior of the proposed strategy in this sample only. Within the studied period the combination of PCA-based selection and minimum variance allocation improved risk efficiency relative to the market capitalization benchmark, delivered more stable downside behavior during major declines, and exceeded the benchmark by return alone. Risk-adjusted metrics were consistent with this picture. At the same time performance was clearly regime dependent. Rolling alpha and Sharpe turned negative in several phases and moved close to zero toward the end of the period. These findings mean that the results do not support broader claims of continued efficiency or superior performance outside this period or in other markets. Any conclusions about long-term persistence would require separate evidence.

The findings contribute in two ways. For practice, the results indicate that the strategy primarily changes the risk profile, with more stable downside behavior and improved risk efficiency in major declines, while excess returns are not uniform across subperiods. Rolling alpha and Sharpe are negative in several windows, which shows that relative performance varies over time. Further evidence indicates that performance is regime dependent, with the strongest outcomes during elevated dispersion and drawdowns, while long benchmark led rallies can be more challenging. For previous literature, the thesis provides evidence from the studied S&P 500 sample that PCA-based equity selection does not translate into uniformly persistent outperformance but instead delivers clustered outcomes that are closely tied to time-varying dependence conditions.

Future work should focus on testing the robustness and generality of these results. Extending the analysis to longer horizons and additional markets would indicate whether the documented behavior repeats beyond this dataset. The findings also motivate explicit regime-based tests using measures of dispersion, correlations, and benchmark concentration, since performance differed across drawdowns, recoveries, and long benchmark led rallies. A systematic sensitivity analysis over key design choices, such as the rolling window length, rebalancing frequency, the number of principal components, and portfolio constraints, would clarify which elements drive the results. The volatility structure should be examined in greater depth and contrasted with the index to identify when risk control is most effective. Direct forecasting of future returns using principal component features could be compared with regularized linear models and nonlinear alternatives. These steps would help distinguish economic signal from sample specific noise and provide a firmer basis for implementation.

References

Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis (2nd ed.). New York: Wiley.

Agarwal, T., Quelle, H., & Ryan, C. (2021). Principal Component Analysis for Clustering Stock Portfolios. *Journal of Finance and Data Science*, 7(1), 1-12

Aït-Sahalia, Y., & Xiu, D. (2017). Using Principal Component Analysis to Estimate a High-Dimensional Covariance Matrix. *Journal of Econometrics*, 201(2), 384–399.

Ang, A., Hodrick, R. J., Xing, Y., & Zhang, X. (2006). The Cross-section of Volatility and Expected Returns. *Journal of Finance*, 61(1), 259–299.

Bai, J., & Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1), 191–221.

Bailey, D. H., Borwein, J., López de Prado, M., & Zhu, Q. (2017). The Probability of Backtest Overfitting. *The Journal of Computational Finance,* 20(4), 39-69.

Bali, T. G., Çakıcı, N., & Whitelaw, R. F. (2011). Maxing Out: Stocks as Lotteries and the Cross-Section of Expected Returns. *Journal of Financial Economics*, 99(2), 427–446.

Banz, R. W. (1981). The Relationship Between Return and Market Value of Common Stocks. *Journal of Financial Economics*, 9(1), 3–18.

Barber, B. M., & Odean, T. (2001). Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment. *Quarterly Journal of Economics*, 116(1), 261–292.

Barberis, N., & Thaler, R. (2003). A Survey of Behavioral Finance. In *Handbook of the Economics of Finance* (Vol. 1B, pp. 1053–1128). Elsevier.

Basu, S. (1977). Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios: A Test of The Efficient Market Hypothesis. *Journal of Finance*, 32(3), 663–682.

Black, F. (1972). Capital Market Equilibrium with Restricted Borrowing. *Journal of Business*, 45(3), 444–455.

Black, F., Jensen, M. C., & Scholes, M. (1972). The Capital Asset Pricing Model: Some Empirical Tests. In M. C. Jensen (Ed.), *Studies in the Theory of Capital Markets* (pp. 79–121). New York: Praeger.

Bodie, Z., Kane, A. & Marcus, A.J. (2024) Investments. 13th edition, McGraw-Hill, Boston.

Bogle, J. C. (1975). *Bogle on Mutual Funds: New Perspectives for the Intelligent Investor*. New York: McGraw-Hill

Boyd, S. & Vandenberghe, L. (2004). Convex Optimization. Cambridge University Press.

Breeden, D. T. (1979). An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities. *Journal of Financial Economics*, 7(3), 265–296.

Bufalo, D., Bufalo, G., Cesarone, F., & Orlando, G. (2025). Hybrid PCA Methods for Parsimonious S&P 500 Index Tracking. *Working paper*.

Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.

Cao, J., & Wang, J. (2020). Exploration of Stock Index Change Prediction Model Based on the Combination of Principal Component Analysis and Artificial Neural Networks. *Soft Computing*, 24, 7851-7860

Carhart, M. M. (1997). On Persistence in Mutual Fund Performance. *Journal of Finance*, 52(1), 57–82.

CFA Institute. (2024). *Trading Costs and Electronic Markets*. Charlottesville, VA: CFA Institute.

Chamberlain, G., & Rothschild, M. (1983). Arbitrage, Factor Structure, and Mean–Variance Analysis on Large Asset Markets. *Econometrica*, 51(5), 1281–1304.

Chen, N.-F., Roll, R., & Ross, S. A. (1986). Economic Forces and The Stock Market. *Journal of Business*, 59(3), 383–403.

Chincarini, L. B., & Kim, D. (2010). *Quantitative Equity Portfolio Management: An Active Approach to Portfolio Construction and Management*. New York, NY: McGraw-Hill.

Chitenderu, T., Maredza, A. & Sibanda, K. (2014). The Random Walk Theory and Stock Prices: Evidence from Johannesburg Stock Exchange. International Business & Economics Research Journal (IBER), 13(6), 1241-1250.

Cochrane, J. H. (2005). *Asset Pricing* (Rev. ed.). Princeton University Press.

Cochrane, J. H. (2011). Presidential Address: Discount Rates. *Journal of Finance*, 66(4), 1047–1108.

Connor, G., & Korajczyk, R. A. (1986). Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis. *Journal of Financial Economics*, 15(3), 373–394.

Connor, G., & Korajczyk, R. A. (1988). Risk and Return in an Equilibrium APT: Application of a New Test Methodology. *Journal of Financial Economics*, 21(2), 255–289.

Cont, R. (2001). Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues. *Quantitative Finance*, 1(2), 223–236.

Cornuéjols, G. & Tütüncü, R. (2007). Optimization Methods in Finance. Cambridge University Press.

Da, Z., Engelberg, J., & Gao, P. (2011). In Search of Attention. *Journal of Finance*, 66(5), 1461–1499.

Dal Pra, G., Guidolin, M., Pedio, M., & Vasile, F. (2018). Practical Applications of Regime Shifts in Excess Stock Return Predictability: An Out-of-Sample Portfolio Analysis. *Practical Application*, 6(1), 1-5

Daniel, K., & Titman, S. (1997). Evidence on the Characteristics of Cross-Sectional Variation in Stock Returns. *Journal of Finance*, 52(1), 1–33

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise Trader Risk in Financial Markets. *Journal of Political Economy*, 98(4), 703–738.

DellaVigna, S., & Pollet, J. M. (2009). Investor Inattention and Friday Earnings Announcements. *Journal of Finance*, 64(2), 709–749.

Disanaike, G. (1997) Do Stock Market Investors Overreact?, *Journal of Business Finance & Accounting*, 24(1), s.27-50

Elton, E. J., & Gruber, M. J. (1997). Modern Portfolio Theory, 1950 to Date. *Journal of Banking & Finance*, 21(11-12), 1743–1759

Fan, J., Liao, Y., & Mincheva, M. (2013). Large Covariance Estimation by Thresholding Principal Orthogonal Complements. *Journal of the Royal Statistical Society: Series B*, 75(4), 603–680.

Fan, Q., Hu, F., & Zhang, X.-P. (2019). Double-Selection Based High-Dimensional Factor Model with Application in Asset Pricing. *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 1-5

Fama, E. F. (1965). The Behavior of Stock-Market Prices. *Journal of Business*, 38(1), 34–105

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, 25(2), 383–417.

Fama, E. F., & French, K. R. (1992). The Cross-Section of Expected Stock Returns. *Journal of Finance*, 47(2), 427–465.

Fama, E. F., & French, K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33(1), 3–56.

Fama, E. F., & French, K. R. (2004). The Capital Asset Pricing Model: Theory and evidence. *Journal of Economic Perspectives*, 18(3), 25–46.

Fama, E. F., & MacBeth, J. D. (1973). Risk, Return, and Equilibrium: Empirical tests. *Journal of Political Economy*, 81(3), 607–636.

Fernando, J. (2025). Sharpe Ratio: Definition, Formula, and Examples, Website, Available: Sharpe Ratio: Definition, Formula, and Examples, Retrieved: 14.10.2025

Gewers, F. L., Ferreira, G. R., de Arruda, H. F., Silva, F. N., Comin, C. H., Amancio, D. R., and da F. Costa, L. (2021). Principal Component Analysis: A Natural Approach to Data Exploration. *ACM Computing Surveys*, 54(4), Article 70, 1–34.

Ghorbani, M., & Chong, E. K. P. (2020). Stock Price Prediction Using Principal Components. *PLoS ONE*, 15(3)

Glasserman, P. (2004). Monte Carlo Methods in Financial Engineering. New York, NY: Springer.

Graham, B., & Dodd, D. (1934). Security Analysis. New York: McGraw-Hill

Grinold, R. C., & Kahn, R. N. (2007). Active Portfolio Management (2nd ed.). McGraw-Hill.

Grossman, S. J., & Stiglitz, J. E. (1980). On the Impossibility of Informationally Efficient Markets. *American Economic Review*, 70(3), 393–408.

Hansen, L. P., & Singleton, K. J. (1982). Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models. *Econometrica*, 50(5), 1269–1286.

Hargreaves, C. A., & Mani, C. K. (2015). The Selection of Winning Stocks Using Principal Component Analysis. *American Journal of Marketing Research*, 1(3), 183–188.

Harris, L. (2003). Trading and Exchanges: Market Microstructure for Practitioners. New York, NY: Oxford University Press.

Haugen, R. A., & Baker, N. L. (1991). The Efficient Market Inefficiency of Capitalization-Weighted Stock Portfolios. *Journal of Portfolio Management*, 17(3), 35–40.

Herschberg, M. (2012). Trading Costs and Arbitrage Profitability. *Journal of Investment Strategies*, 1(3), 63–92.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24, 417–441; 498–520.

Jackson, J. E. (1991). A User’s Guide to Principal Components. New York: Wiley.

Jagannathan, R. & Ma, T. (2003). Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps. *The Journal of Finance*, 58(4), 1651–1683.

Jegadeesh, N. & Titman, S. (1993) Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency, *Journal of Finance*, 48(1), 65-91

Jensen, M. C. (1968). The Performance of Mutual Funds in the Period 1945–1964. *Journal of Finance*, 23(2), 389–416.

John, A., Oduro, F. T., & Ackora-Prah, J. (2016). Selection of Stocks on the Ghana Stock Exchange Using Principal Component Analysis. *International Journal of Theoretical and Applied Mathematics*, 2(2), 100–109.

Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). New York: Springer.

Jolliffe, I. T., & Cadima, J. (2016). Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.

Joubert, G. (2005). Excess Returns to Equity Market Neutral Hedge Funds. *Investment Analysts Journal*, 34, 31-41.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263–291.

Kendall, M. (1953) The Analysis of Economic Time-Series-Part 1: Prices. Journal of the Royal Statistical Society, Vol. 116 (1), 11–34.

Ledoit, O., & Wolf, M. (2004). A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *Journal of Multivariate Analysis*, 88(2), 365–411.

Ledoit, O., & Wolf, M. (2012). Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices. *The Annals of Statistics*, 40(2), 1024–1060.

Lehmann, B. N., & Modest, D. M. (1988). The Empirical Foundations of the Arbitrage Pricing Theory. *Journal of Financial Economics*, 21(2), 213–254.

Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *Review of Economics and Statistics*, 47(1), 13

Lo, A. W. (2017). Adaptive Markets: Financial Evolution at the Speed of Thought. Princeton, NJ: Princeton University Press.

Lo, A. W. & MacCinlay, A. C. (1988). Stock Market Prices do not Follow Random Walks: Evidence from a Simple Specification Test, *The Review of Financial Studies*, 1(1), 41-66

Longin, F., & Solnik, B. (2001). Extreme Correlation of International Equity Markets. *Journal of Finance*, 56(2), 649–676.

López de Prado, M. (2018). Advances in Financial Machine Learning. Hoboken, NJ: Wiley.

Lucas, R. E. (1978). Asset Prices in an Exchange Economy. *Econometrica*, 46(6), 1429–1445.

Norland, E., & Wilford, D. S. (2002). Global Portfolios Should Be Optimized in Excess, Not Total Returns. *Review of Financial Economics*, 11(3), 213-224

Malaj, E., & Malaj, V. (2016). Portfolio Allocation: An Empirical Analysis of Ten American Stocks for the Period 2010-2015. *Journal of Accounting and Finance*, 1(1), 11

Malkiel, B. G. (1973). *A Random Walk Down Wall Street*. W. W. Norton & Company.

Mandelbrot, B. (1963). The Variation of Certain Speculative Prices. *Journal of Business*, 36(4), 394–419

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*, 7(1), 77–91.

Menkveld, A. J. (2013). High Frequency Trading and the New Market Makers. *Journal of Financial Markets*, 16(4), 712–740.

Merton, R. C. (1973). An Intertemporal Capital Asset Pricing Model. *Econometrica*, 41(5), 867–887.

Merton, R. C. (1987). A Simple Model of Capital Market Equilibrium with Incomplete Information. *Journal of Finance*, 42(3), 483–510.

Mossin, J. (1966). Equilibrium in a Capital Asset Market. *Econometrica*, 34(4), 768–783.

Odean, T. (1998). Are Investors Reluctant to Realize Their Losses? *Journal of Finance*, 53(5), 1775–1798.

Park, E. W. (2023). Principal Component Analysis and Hidden Markov Model for Forecasting Stock Returns. *Journal of Financial Data Science,* Courant Institute of Mathematical Sciences, New York University

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Series 6, 2(11), 559–572.

Ritter, J. R. (2003). Behavioral Finance. *Pacific-Basin Finance Journal*, 11(4), 429–437.

Roll, R. (1977). A Critique of the Asset Pricing Theory's Tests. Part I: On Past and Potential Testability of the Theory. *Journal of Financial Economics*, 4(2), 129–176.

Roll, R., & Ross, S. A. (1980). An Empirical Investigation of the Arbitrage Pricing Theory. *Journal of Finance*, 35(5), 1073–1103.

Roncalli, T. (2013). Introduction to Risk Parity and Budgeting. CRC Press.

Rosenberg, B., Reid, K., & Lanstein, R. (1985). Persuasive Evidence of Market Inefficiency. *Journal of Portfolio Management*, 11(3), 9–16.

Ross, S. A. (1976). The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory*, 13(3), 341–360.

Rossi, M., & Gunardi, A. (2018). Efficient Market Hypothesis And Stock Market Anomalies: Empirical Evidence in Four European Countries, The Journal of Applied Business Research, Vol 34, (1), 183–192

Samuelson, P. A. (1965). Proof that Properly Anticipated Prices Fluctuate Randomly, *Industrial Management Review*, 6(2), 41-49.

Schneider, P., & Stübinger, J. (2020). PCA-Based Constituent Selection for Index Dispersion Trading. *Working paper*.

Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance*, 19(3), 425–442.

Sharpe, W. F. (1966) Mutual Fund Performance, *The Journal of Business*, 39(1), 119-138

Shiller, R. J. (2003). From Efficient Markets Theory to Behavioral Finance. *Journal of Economic Perspectives*, 17(1), 83–104.

Shleifer, A. (2000). Inefficient Markets: An Introduction to Behavioral Finance. Oxford University Press.

Shleifer, A., & Vishny, R. W. (1997). The Limits of Arbitrage. *Journal of Finance*, 52(1), 35–55.

Shukla, R. (2004). The Value of Active Portfolio Management. *Journal of Economics and Business*, 56(3), 331-346

Slezak, S. L. (2003). On the Impossibility of Weak-Form Efficient Markets. *Journal of Financial and Quantitative Analysis*, 38(3), 523-554.

Solnik, B. H. (1983). International Arbitrage Pricing Theory. *Journal of Finance*, 38(2), 449–457.

S&P Global, SP500, Website, Available: https://www.spglobal.com/spdji/en/indices/equity/sp-500/#overview, Retrieved: 07.10.2025

Stattman, D. (1980). Book Values and Stock Returns. *The Chicago MBA: A Journal of Selected Papers*, 4, 25–45.

Sun, H., Li, J., & Long, F. (2020). Empirical Research of Energy Sector Based on Principal Component Analysis. In *Proceedings of the 5th International Conference on Mechatronics, Control and Electronic Engineering (MCEE 2020)*.

Tsay, R. S. (2010). Analysis of Financial Time Series (3rd ed.). Hoboken, NJ: Wiley.

Tversky, A., & Kahneman, D. (1974). Judgment Under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–1131.

Vogl, M. (2022). Quantitative Modelling Frontiers: A Literature Review on the Evolution in Financial and Risk Modelling After the Financial Crisis (2008–2019). *SN Business & Economics*, 2, 183.

Waqar, M., Dawood, H., Guo, P., Shahnawaz, M. B., & Ghazanfar, M. A. (2017). Prediction of Stock Market by Principal Component Analysis. In *2017 13th International Conference on Computational Intelligence and Security (CIS)*, 599–604.

Wilmott, P. (2006). Paul Wilmott on Quantitative Finance (2nd ed.). Chichester, UK: Wiley.

Yang, L., Rea, W., & Rea, A. (2016). Stock Selection with Principal Component Analysis. *Journal of Interaction Science*, 5(1), 35-55

Zhou, G., & Wang, Y. (2015). Forecasting Stock Market Indexes Using Principal Component Analysis and Stochastic Time Effective Neural Networks. Neurocomputing, 167, 232–242.

Appendices

Appendix 1 Composition of most recent portfolios

Date	Symbol	Weight	Portfolio_Expected_Return	Portfolio_Risk
2024-10-21	ENPH	0	28,44	12,68
2024-10-21	EXR	7,99	28,44	12,68
2024-10-21	A	13,43	28,44	12,68
2024-10-21	MSCI	11,29	28,44	12,68
2024-10-21	INCY	10,01	28,44	12,68
2024-10-21	MOS	4,91	28,44	12,68
2024-10-21	DXCM	4,11	28,44	12,68
2024-10-21	LKQ	6,4	28,44	12,68
2024-10-21	CBOE	32,22	28,44	12,68
2024-10-21	IP	9,63	28,44	12,68
2024-11-04	PLTR	6,36	21,02	22,3
2024-11-04	UAL	12,67	21,02	22,3
2024-11-04	CDNS	7	21,02	22,3
2024-11-04	SBUX	20,63	21,02	22,3
2024-11-04	CHTR	20,72	21,02	22,3
2024-11-04	SNPS	14,23	21,02	22,3
2024-11-04	KLAC	0	21,02	22,3
2024-11-04	LULU	10,56	21,02	22,3
2024-11-04	CZR	4,56	21,02	22,3
2024-11-04	AMAT	3,27	21,02	22,3
2024-11-18	TSLA	3,06	38,3	14,13
2024-11-18	EPAM	5,86	38,3	14,13
2024-11-18	AXON	4,81	38,3	14,13
2024-11-18	DIS	13,73	38,3	14,13
2024-11-18	FTNT	12,34	38,3	14,13
2024-11-18	TPR	5,65	38,3	14,13
2024-11-18	DXCM	6,69	38,3	14,13
2024-11-18	TAP	12,27	38,3	14,13
2024-11-18	TSN	19,57	38,3	14,13
2024-11-18	TTWO	16,04	38,3	14,13
2024-12-02	CE	6,4	1,01	14,85
2024-12-02	APA	0,43	1,01	14,85
2024-12-02	DXCM	8	1,01	14,85
2024-12-02	LKQ	5,97	1,01	14,85
2024-12-02	MSCI	18,61	1,01	14,85
2024-12-02	SLB	18,76	1,01	14,85
2024-12-02	ENPH	4,42	1,01	14,85
2024-12-02	ACGL	30,47	1,01	14,85
2024-12-02	MOS	5,32	1,01	14,85
2024-12-02	MRNA	1,61	1,01	14,85
2024-12-16	AVGO	4,65	2,62	26,45
2024-12-16	WBA	23,21	2,62	26,45
2024-12-16	PLTR	8,41	2,62	26,45
2024-12-16	AMAT	2,3	2,62	26,45
2024-12-16	ENPH	7,14	2,62	26,45
2024-12-16	LULU	22,98	2,62	26,45
2024-12-16	KLAC	7,72	2,62	26,45
2024-12-16	LRCX	0	2,62	26,45
2024-12-16	CDNS	23,6	2,62	26,45
2024-12-16	MPWR	0	2,62	26,45