

CHARLES UNIVERSITY

Center for Economic Research and Graduate Education



Master's thesis

2026

Bc. Robert Oganessian

CHARLES UNIVERSITY

Center for Economic Research and Graduate Education



Comparison of Simple and Advanced ETF Portfolio Allocation Strategies across Major World Regions and Economic Sectors for a Retail Investor

Master's thesis

Author: Bc. Robert Oganessian

Study program: MA in Economic Research

Supervisor: Ctirad Slavík, Ph.D.

Year of defense: 2026

Declaration

1. I hereby declare that I have compiled this thesis using only the listed literature and resources.
2. I hereby declare that this thesis has not been used to obtain any other academic title.
3. I fully agree to this thesis being used for study and scientific purposes.
4. During the preparation of this thesis, the author used OpenAI ChatGPT, OpenAI Codex, and GitHub Copilot in Visual Studio Code to assist with brainstorming, language revision, LaTeX drafting, coding and debugging, and consistency and reproducibility checks. Python and Jupyter were used for data processing, statistical analysis, portfolio simulation, and visualization; LaTeX and MiKTeX for typesetting; and Git for version control. The author critically reviewed, edited, and, where applicable, tested all AI-assisted text and code. The author verified the data transformations, calculations, citations, interpretations, and conclusions and takes full responsibility for the final content of the thesis.

In Prague on
July 28, 2026

Robert Oganessian

Bibliographic Record

Oganesyan, Robert: *Comparison of Simple and Advanced ETF Portfolio Allocation Strategies across Major World Regions and Economic Sectors for a Retail Investor*. Master's thesis. Charles University, Center for Economic Research and Graduate Education, Prague. 2026, pages 183. Advisor: Ctirad Slavík, Ph.D.

Length of the Thesis: 229 334 characters including spaces, excluding the abstracts, bibliography, and appendices.

Abstract

This thesis examines whether advanced portfolio optimization methods applied to ETFs outperform simple investment rules for a retail investor. The analysis uses a growing universe of 71 ETFs covering international and sector equities, factor exposures, U.S. Treasuries, commodities, and listed real estate. Portfolios are formed monthly using only information available at the time.

The study first compares simple allocation rules with minimum-variance, equal risk contribution, maximum diversification, and mean-variance strategies. It then tests whether macroeconomic and financial information improves allocation by comparing macro-informed strategies with otherwise comparable models based only on historical returns or unconditional return moments. Finally, all strategies are evaluated under a retail constraint implementation with regular monthly contributions and transaction costs, and their terminal wealth is compared with 1,000 randomly constructed ETF investment paths.

The results show selective rather than general benefits from greater sophistication. Equal risk contribution provides the strongest overall balance of risk-adjusted performance, wealth accumulation, diversification, and implementation cost. Macroeconomic information adds value mainly in the regime-conditioned sleeve strategy, while the Black-Litterman results are weak and sensitive to the specification. None of the leading strategies reaches the upper tail of the Random DCA distribution, showing that simple and transparent allocation remains highly competitive.

JEL Classification G11, G17, C53

Keywords Portfolio optimization; ETF portfolio allocation; retail investor framework; macro-financial regime-switching; regional and sector diversification; equal-weight benchmark; dollar-cost averaging

Title Comparison of Simple and Advanced ETF Portfolio Allocation Strategies across Major World Regions and Economic Sectors for a Retail Investor

Author's e-mail robert.oganesyan@cerge-ei.cz

Supervisor's e-mail ctirad.slavik@cerge-ei.cz

Abstrakt

Tato práce zkoumá, zda pokročilé metody alokace ETF překonávají jednoduchá investiční pravidla pro drobného investora. Využívá rostoucí soubor 71 ETF zahrnujících regionální a sektorové akcie, faktorové strategie, americké státní dluhopisy, komodity a kótované nemovitosti. Portfolia jsou každý měsíc sestavována pouze z tehdy dostupných informací.

Nejprve jsou jednoduchá pravidla porovnána se strategiemi minimální variance, rovnoměrného příspěvku k riziku, maximální diverzifikace a optimalizace střední hodnoty a rozptylu. Následně práce zkoumá přínos makroekonomických a finančních informací porovnáním makroekonomicky podmíněných strategií s obdobnými modely založenými pouze na historických výnosech. Nakonec jsou všechny strategie hodnoceny při pravidelných měsíčních vkladech a transakčních nákladech a jejich konečná hodnota je porovnána s 1 000 náhodně sestavenými ETF portfolii.

Výsledky ukazují pouze dílčí přínos vyšší složitosti. Nejlepší celkovou rovnováhu mezi rizikově očištěnou výkonností, růstem majetku, diverzifikací a náklady nabízí strategie rovnoměrného příspěvku k riziku. Makroekonomické informace pomáhají především režimově podmíněné alokaci mezi skupinami aktiv, zatímco výsledky Blackova–Littermanova modelu jsou slabé a citlivé na specifikaci. Žádná z vedoucích strategií se nedostává do horní části rozdělení Random DCA, což ukazuje, že jednoduchá a transparentní alokace zůstává vysoce konkurenceschopná.

Klasifikace JEL

G11, G17, C53

Klíčová slova

optimalizace portfolia; alokace portfolia ETF; rámec drobného investora; přepínání makrofinančních režimů; regionální a sektorová diverzifikace; benchmark s rovnými vahami; pravidelné investování

Název práce

Srovnání jednoduchých a pokročilých strategií alokace portfolia ETF napříč hlavními světovými regiony a hospodářskými odvětvími z pohledu drobného investora

E-mail autora

robert.oganesyan@cerge-ei.cz

E-mail vedoucího práce

ctirad.slavik@cerge-ei.cz

Acknowledgments

I am especially grateful to Ctirad Slavík, Ph.D., for providing valuable insights and supervision. I would also like to thank my parents for their patience and support throughout my studies.

Contents

List of Tables	xiii
List of Figures	xv
Acronyms	xviii
Introduction	1
1 Literature review	6
1.1 Foundations of Portfolio Allocation	7
1.2 Simple Benchmarks	11
1.2.1 Equal-weight allocation	11
1.2.2 Market-capitalization-weighted benchmark	12
1.3 Constraints and Robust Input Estimation	13
1.3.1 Portfolio constraints	13
1.3.2 Linear covariance shrinkage	13
1.3.3 Nonlinear covariance shrinkage	14
1.3.4 Expected-return estimation	15
1.4 Advanced Portfolio Allocation Methods	16
1.4.1 Risk-based allocation	17
1.4.2 View-based and conditional allocation	18
1.5 ETF Portfolio Allocation and Diversification	20
1.5.1 International and factor-based ETF diversification	21
1.5.2 Direct ETF portfolio optimization	23
1.6 Macroeconomic Indicators and Conditional Allocation	24
1.6.1 Macroeconomic State Variables and Time-Varying Port- folio Inputs	25
1.6.2 Limits of direct return prediction	26
1.6.3 Regime-switching and conditional allocation	26

1.6.4	Regime-switching ETF allocation with macro-financial factors	27
1.6.5	Practical implementation risks	29
1.7	Summary	30
1.8	Research Gap, Research Question, and Hypotheses	31
2	Data and Methodology	34
2.0.1	Hypothesis Assessment Framework	36
2.1	Common Empirical Framework	37
2.1.1	ETF Prices, Currency Conversion, and Monthly Returns	38
2.1.2	Risk-Free Return	39
2.1.3	ETF Universe Selection	40
2.1.4	Economic Sleeve Classification	41
2.1.5	Growing Eligible Universe	43
2.1.6	Information Set and Portfolio Constraints	45
2.1.7	Common Return and Covariance Estimation	47
2.1.8	Common Benchmarks	49
2.2	Hypothesis 1: Unconditional Portfolio Allocation	51
2.2.1	Covariance-Based Allocation	52
2.2.2	Mean–Variance Allocation	54
2.3	Hypothesis 2: Added Value of Macroeconomic and Financial Information	54
2.3.1	Macroeconomic and Financial Information Set	55
2.3.2	Recursive Three-State Hidden Markov Model	57
2.3.3	Sleeve Returns and Regime-Conditioned Moments	59
2.3.4	Matched Black–Litterman Comparison	62
2.3.5	Matched Sleeve-Level Mean–Variance Comparison	64
2.4	Hypothesis 3: Retail-Investor Implementation	65
2.4.1	Monthly Contributions and Trading Costs	66
2.4.2	Random DCA Benchmark	67
2.5	Backtesting and Performance Evaluation	68
2.5.1	Common Starting Date and Monthly Sequence	68
2.5.2	Frictionless Evaluation for H1 and H2	70
2.5.3	Contribution-Based Evaluation for H3	72
2.5.4	Random DCA Distributional Comparison	73
2.5.5	Stationary-Bootstrap Inference for H1 and H2	74
2.5.6	Robustness Specifications and Model Diagnostics	75

2.5.7	Summary of Fixed Methodological Choices	76
3	Empirical Results	78
3.1	Evaluation Sample and Presentation of Results	78
3.2	Hypothesis 1: Unconditional Portfolio Allocation	79
3.2.1	Overall Performance	80
3.2.2	Risk, Drawdown, and Wealth Dynamics	80
3.2.3	Portfolio Composition, Concentration, and Turnover	82
3.2.4	Assessment of Hypothesis 1	83
3.3	Hypothesis 2: Added Value of Macroeconomic and Financial Information	84
3.3.1	Estimated Macro-Financial Regimes	84
3.3.2	Historical-View versus Macro-View Black–Litterman	87
3.3.3	Unconditional versus Regime-Conditioned Sleeve MVO	88
3.3.4	Broader Common-Period Context	89
3.3.5	Assessment of Hypothesis 2	91
3.4	Hypothesis 3: Retail-Investor Implementation	93
3.4.1	Contributions, Trading Costs, and Net Wealth	93
3.4.2	Gross-to-Net Performance Differences	95
3.4.3	Comparison with the Random DCA Distribution	96
3.4.4	Assessment of Hypothesis 3	97
3.5	External Market Benchmark: Vanguard VT	98
3.6	Robustness Analysis	102
3.7	Stationary-Bootstrap Inference	103
3.8	Summary of Empirical Findings	105
4	Discussion	107
4.1	Research Question and Empirical Contribution	107
4.2	Evaluation of the Hypotheses	108
4.2.1	Unconditional Allocation	108
4.2.2	Macroeconomic and Financial Information	109
4.2.3	Retail Implementation	110
4.3	Implications for a Retail Investor	111
4.4	Limitations	112
4.5	Future Research and Final Assessment	113
	Conclusion	115

Summary	117
Bibliography	119
A Supplementary Notes on Mean-Variance Optimization	I
A.1 Constrained and Utility-Based Formulations	I
A.2 Tobin Separation and the Tangency Portfolio	IV
A.2.1 Optimal Allocation Between the Risk-Free Asset and the Tangency Portfolio	VI
B Supplementary Notes on ETF Factor Portfolios	VIII
B.1 Description of Factor and Anomaly Portfolios	VIII
C Data Construction and Core ETF Universe	X
C.1 Synthetic U.S. Treasury-Bill Series	X
C.2 Real-Time Macroeconomic and Financial Series	XI
C.3 List of 71 ETFs	XIII
C.4 Permanent Sleeve Classification	XV
D Quadratic-Inverse Shrinkage Covariance Estimation	XVI
D.1 Data Matrix and Sample Covariance	XVI
D.2 Rotation-Equivariant Oracle	XVII
D.3 Smoothed Local Shrinkage Functions	XVIII
D.4 Nonsingular and Boundary Cases	XIX
D.5 Singular Case	XX
D.6 Trace Preservation and Matrix Reconstruction	XXI
D.7 Interpretation	XXI
D.8 Python Implementation	XXII
E Technical Portfolio Specifications	XXIV
E.1 Common Feasible Sets and Sleeve Mapping	XXIV
E.2 Numerical Solution of the Unconditional Models	XXV
E.2.1 GMV and Mean-Variance Optimization	XXV
E.2.2 Equal Risk Contribution	XXVI
E.2.3 Maximum Diversification	XXVI
E.3 Black-Litterman Implementation	XXVII
E.3.1 Reference-Portfolio-Implied Returns	XXVII
E.3.2 Sleeve View Matrix and View Vectors	XXVII
E.3.3 Forecast-Error Reliability	XXVIII

E.3.4	Posterior and Final Portfolio	XXVIII
E.4	Black–Litterman Performance	XXIX
E.4.1	Absolute Performance Across Specifications	XXIX
E.4.2	Comparison with Bessler, Opfer, and Wolff	XXX
E.4.3	Sources of Baseline Underperformance	XXX
E.5	Three-State Hidden Markov Model	XXXII
E.5.1	Scaled Forward Recursion and Likelihood	XXXII
E.5.2	Backward Recursion and Smoothed Probabilities	XXXIII
E.5.3	Expectation–Maximization Updates	XXXIII
E.6	Regime-Conditioned Sleeve Moments	XXXIV
E.7	Random DCA and Retail Wealth Accounting	XXXVI
E.7.1	External VT Reference	XXXVII
E.8	Stationary-Bootstrap Inference Results	XXXVIII
E.9	Transaction-Cost Sensitivity	XXXIX
E.10	Numerical Validation and Reproducibility	XXXIX

List of Tables

2.1	Empirical comparisons used to evaluate the hypotheses	36
2.2	Permanent economic sleeves used in portfolio construction	42
2.3	Growth of the ETF dataset and 60-month eligible universe	45
2.4	Policy allocation when all 15 sleeves are active	50
2.5	Baseline methodological specification	76
3.1	Evaluation periods and frozen empirical vintages	79
3.2	Standalone performance of unconditional ETF strategies	80
3.3	Economic characteristics and persistence of the final recursive HMM fit	84
3.4	Effective regime samples and pooling toward unconditional mo- ments	85
3.5	Matched macro-conditioning increments across baseline and pre- specified variants	88
3.6	Common-period performance of the ten core strategies	89
3.7	Common-period asset-class exposure and sleeve concentration . .	90
3.8	Retail implementation outcomes under monthly contributions and five-basis-point trading costs	94
3.9	External market context from Vanguard Total World Stock ETF	98
3.10	Stationary-bootstrap qualification of H1 and H2 Sharpe compar- isons	104
C.1	Core ETF universe, part 1	XIII
C.2	Core ETF universe, part 2	XIV
C.3	Permanent economic sleeves and member ETFs	XV
E.1	Absolute Black–Litterman performance across pre-specified vari- ants	XXX
E.2	Black–Litterman design comparison with Bessler, Opfer, and Wolff	XXXI

E.3	Pairwise stationary-bootstrap Sharpe-ratio differences	XXXVIII
E.4	Stationary-bootstrap sensitivity to expected block length	XXXVIII
E.5	Sensitivity of retail outcomes to proportional transaction costs	XXXIX
E.6	Numerical diagnostics for the baseline empirical results	XL

List of Figures

1.1	Illustration of the efficient frontier. The upper branch of the frontier represents efficient portfolios, while the lower branch is inefficient. The crosses represent feasible portfolios. Portfolios A, B, and C lie on the efficient frontier. Portfolio D is inefficient because it has the same level of risk as portfolio B but a lower expected return. The global minimum-variance portfolio (GMV) is the leftmost point on the frontier.	8
1.2	Illustration of the efficient frontier and the capital allocation line (CAL). The square point r_f denotes the risk-free asset, while point T denotes the tangency portfolio. The CAL is tangent to the efficient frontier at T and represents all portfolios formed as combinations of the risk-free asset and the tangency portfolio. Its slope equals the Sharpe ratio, that is, the excess return per unit of risk.	9
3.1	Cumulative frictionless wealth of the standalone H1 strategies, December 2009–June 2026. All paths begin at one. Flat $1/N$ is a robustness benchmark rather than a formal same-universe H1 comparator.	81
3.2	Turnover and concentration diagnostics for the standalone H1 strategies. The upper panel reports the 12-month moving average of one-way drift-adjusted turnover; the lower panel reports effective holdings.	82
3.3	Pooling weights applied to state-specific mean and covariance estimates, by HMM state and holding month. The remaining weight is assigned to the corresponding unconditional estimate. Panel bounds equal the theoretical maxima implied by the 24-month mean window, 60-month covariance window, and $\kappa = 36$	86

3.4	One-step predicted probabilities of the favourable, intermediate, and adverse macro-financial states from the recursive baseline HMM. The probabilities available at each formation date sum to one and include the pre-evaluation warm-up period.	87
3.5	Common-period target weights in Equity, Treasuries, Commodities, and Real estate for the ten core strategies, March 2012–June 2026. Each panel sums to one at every holding month.	91
3.6	Terminal net wealth percentiles of the ten target-weight strategies within the conditional 1,000-path Random DCA distribution, March 2012–June 2026. The Random DCA mean and median terminal wealth are 424.212 and 422.137 contribution units; its 5th and 95th percentiles are 378.906 and 473.780. Higher strategy percentiles indicate more favourable terminal wealth within the fixed simulation design.	96
3.7	VT and the formal strategy wealth paths over the standalone H1 and common H1/H2 periods. VT is distinguished as an external all-equity reference and does not enter the formal hypothesis comparisons.	99
3.8	Regional composition of Vanguard Total World Stock ETF. North America accounts for 64.9% of the fund’s equity exposure; the remaining exposure is distributed mainly across Europe, the Pacific region, and emerging markets.	100
3.9	Ten largest VT securities as of 30 June 2026. Together they represented 20.42% of net assets and were concentrated in large technology and technology-adjacent platform companies. Alphabet appears through two share classes.	100
3.10	VT and the ten formal strategy net contribution-based wealth paths, March 2012–June 2026. All paths use the common initial wealth, beginning-of-month contributions, fractional shares, and proportional trading-cost assumption.	101
3.11	VT within the conditional Random DCA terminal-net-wealth distribution. The dashed line identifies VT and the dotted line identifies the Random DCA median. The comparison is conditional on the realized market history and the fixed Random DCA design.	102

3.12	Observed paired Sharpe-ratio differences and 95% stationary-bootstrap percentile intervals under the primary 12-month expected block length. Blue rows belong to H1 and red rows to H2; intervals crossing zero indicate that the ranking reverses in some plausible pseudo-histories.	104
C.1	Synthetic monthly U.S. Treasury-bill return constructed from the FRED daily 3-month Treasury constant maturity yield . . .	XI

Acronyms

ALFRED	Archival Federal Reserve Economic Data
API	Application Programming Interface
BL	Black–Litterman
BL-H	Historical-view Black–Litterman
BL-M	Macro-view Black–Litterman
CAL	Capital Allocation Line
CAPM	Capital Asset Pricing Model
CFNAI	Chicago Fed National Activity Index
CFNAI-MA3	three-month moving average of the Chicago Fed National Activity Index
CVaR	Conditional Value-at-Risk
DCA	Dollar-Cost Averaging
ERC	Equal Risk Contribution
ETF	Exchange-Traded Fund
FRED	Federal Reserve Economic Data
FX	Foreign Exchange
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GMV	Global Minimum Variance
HEW	Hierarchical Equal Weight
HML	High Minus Low

HMM	Hidden Markov Model
MD	Maximum Diversification
MVO	Mean–Variance Optimization
MWR	Money-Weighted Return
QIS	Quadratic-Inverse Shrinkage
RDA	Random Dollar-Cost Averaging
SLSQP	Sequential Least Squares Programming
SMB	Small Minus Big
TWR	Time-Weighted Return
UCITS	Undertakings for Collective Investment in Transferable Securities
USD	United States Dollar
VIX	CBOE Volatility Index
VT	Vanguard Total World Stock ETF

Introduction

Retail investors can now obtain broad exposure to international equity markets, economic sectors, government bonds, commodities, and real estate through exchange-traded funds (ETFs). This access creates an opportunity, but also an allocation problem. A large investable menu does not by itself determine how capital should be distributed across exposures, how frequently portfolios should be adjusted, or whether statistical optimization justifies its informational and implementation demands. These questions are especially important for investors who contribute regularly, face trading costs, and have limited capacity to monitor complex models.

ETFs provide a suitable setting in which to study this problem. They combine the diversification of an investment fund with exchange trading and make regional, sectoral, factor, fixed-income, commodity, and real-estate exposures accessible through individual instruments (Lettau & Madhavan 2018). Unlike conventional mutual funds priced once per day, ETFs can be traded during exchange hours, while regular disclosure of holdings and benchmark objectives makes their economic exposure comparatively transparent. The creation and redemption mechanism permits authorized participants to exchange ETF shares for baskets of underlying securities, supporting the alignment of market prices with net asset values through arbitrage. For a retail investor, these features offer a practical way to construct a broad portfolio without selecting and monitoring a large number of individual securities.

The use of ETFs also supports a more controlled empirical design. It avoids the need to reconstruct the changing constituents of numerous international stock indices and reduces dependence on firm-specific events, while still allowing capital to be allocated across economically distinct markets and risk exposures. ETFs do not eliminate every data or implementation problem: funds have different inception dates, liquidity varies, several products represent overlapping exposures, and historical data on liquidated funds are incomplete. Moreover, a universe containing many specialized funds can give excessive importance to

categories represented by more tickers. A credible comparison therefore requires a common information set, a growing eligibility rule, economically meaningful benchmarks, and constraints applied consistently across strategies.

The portfolio-allocation literature offers no presumption that additional sophistication will improve realized outcomes. Mean–variance optimization can respond strongly to errors in expected returns, while covariance-based and risk-based methods may obtain better risk control without relying on return forecasts. Simple diversification rules remain demanding benchmarks because they avoid much of the estimation error and turnover associated with optimized portfolios (DeMiguel *et al.* 2009). Black–Litterman and regime-conditioned models provide structured ways to incorporate additional information, but their value depends on the quality of the views, the stability of estimated states, and the transmission of those inputs into portfolio weights. Macroeconomic and financial conditions may affect expected returns, volatility, and correlations, yet this does not guarantee that lagged state information produces superior out-of-sample ETF allocations.

Against this background, the thesis asks whether increasingly sophisticated portfolio-allocation methods, and macro-financial conditioning in particular, add practical value relative to simple ETF allocation rules for a constrained retail investor. The objective is not to identify one universally best portfolio. Instead, the analysis determines which methods improve particular dimensions of performance, whether macro-financial information adds value within matched model architectures, and whether any advantages remain economically meaningful after regular contributions and trading costs are introduced. This distinction is necessary because higher terminal wealth, a higher Sharpe ratio, a smaller drawdown, lower turnover, and simpler implementation need not point to the same strategy.

The empirical analysis has three levels. Hypothesis 1 evaluates whether unconditional classical and risk-based allocation strategies outperform Policy and hierarchical equal-weight benchmarks on the same growing ETF universe. The tested strategies are global minimum variance, equal risk contribution, maximum diversification, and ETF-level mean–variance optimization. Hypothesis 2 isolates the incremental contribution of macro-financial information through two matched comparisons. Macro-view Black–Litterman is compared with an otherwise identical historical-view version, while regime-conditioned sleeve-level mean–variance optimization is compared with an unconditional sleeve model. A recursively estimated three-state hidden Markov model uses real-

time economic activity together with lagged volatility, term-spread, and credit-spread information. Hypothesis 3 carries the complete strategy set into a retail implementation with beginning-of-month contributions and proportional transaction costs and locates each terminal outcome within a pre-specified distribution of 1,000 Random DCA paths.

The investable universe contains 71 U.S.- and European-listed ETFs covering broad and country equity markets, sectors, factors, U.S. Treasuries, commodities, and listed real estate. Every fund is assigned to one of 15 permanent economic sleeves. Eligibility begins only after 60 monthly returns are available, so the opportunity set grows without granting newly launched funds an informational advantage. ETF prices and foreign-exchange rates are converted to U.S. dollars, and all formation decisions use information available by the end of the preceding month. The baseline analysis applies long-only, fully invested portfolios, a 20% ETF cap, regularized covariance estimation, and explicit validation of every numerical solution. Standalone H1 covers December 2009 through June 2026, whereas H2, the common strategy comparison, and H3 cover March 2012 through June 2026.

Performance is assessed using annualized return, volatility, Sharpe ratio, maximum drawdown, terminal wealth, turnover, concentration, and numerical diagnostics. The retail analysis additionally distinguishes gross from net wealth, direct transaction cost from terminal cost drag, and time-weighted from money-weighted returns. Rather than comparing each model with only one deterministic benchmark, the thesis simulates 1,000 feasible Random DCA paths that repeatedly select ETFs from the eligible universe under a fixed contribution and review rule. This distribution locates each strategy's terminal net wealth among a range of possible outcomes generated under the same realized market history and investable opportunity set. Its percentiles and exceedance frequencies provide conditional simulation evidence, not classical population-level probabilities.

This simulation-based comparison is a central contribution of the thesis from the perspective of a retail investor. It asks not only whether an optimized strategy beats a selected benchmark, but whether its implemented outcome is unusual relative to many simple ETF allocation paths that the investor could feasibly have followed. Combined with monthly contributions, transaction costs, and strategy-specific trading intensity, the framework evaluates whether statistical sophistication produces a practically meaningful outcome after the conditions of gradual wealth accumulation are incorporated. Van-

guard Total World Stock ETF (VT) is reported separately as an external all-equity reference; because its opportunity set differs from the formal multi-asset strategy problem, it provides market context rather than evidence for or against the three hypotheses.

The thesis contributes a controlled comparison across model complexity, information requirements, and implementation intensity. All strategies share a common data vintage, eligibility calendar, information boundary, and constraint set. Macro-conditioned models are evaluated against matched controls rather than only against unrelated benchmarks, and the contribution-based experiment reuses frozen target paths instead of re-estimating models after observing earlier results. The combination of a growing ETF universe, retail cash-flow accounting, and a distribution of random but feasible allocation outcomes permits weak empirical performance to be distinguished from numerical failure and observed strategy success to be assessed against more than one arbitrarily chosen alternative.

The remainder of the thesis is structured as follows. Chapter 1 develops the theoretical and empirical foundation of the study, beginning with modern portfolio theory and simple benchmark allocation. It then reviews covariance-based, risk-based, Black–Litterman, and conditional allocation methods, together with evidence on ETFs and macro-financial information, before identifying the research gap and stating the three hypotheses.

Chapter 2 defines the common empirical framework, including data construction, currency conversion, the growing 71-ETF universe, economic sleeves, eligibility, portfolio constraints, and benchmark rules. It specifies the unconditional strategies, the recursive hidden Markov model, the matched macro-conditioned strategies, and the pre-specified robustness analysis. The chapter then formalizes the out-of-sample backtest, monthly contribution and transaction-cost accounting, Random DCA simulation, performance measures, and numerical validation rules.

Chapter 3 reports the empirical evidence in the same sequence as the hypotheses. It first evaluates the unconditional strategies, then presents the matched macro-conditioned comparisons, and finally examines contribution-based net outcomes relative to the Random DCA distribution. Robustness analysis, stationary-bootstrap inference, and the separate VT market comparison provide additional context, while detailed numerical validation is reported in the appendices.

Chapter 4 interprets the evidence, evaluates the three hypotheses, draws

practical implications for a constrained retail investor, and discusses the study's limitations and future research. An unnumbered conclusion then states the central answer to the research question, followed by concise English and Czech summaries.

Chapter 1

Literature review

This chapter reviews the theoretical and empirical literature that motivates the portfolio allocation problem studied in this thesis. The central question is not whether sophisticated allocation methods are theoretically more elegant than simple rules, but whether they add practical value for a constrained retail investor who invests in ETFs over time. The chapter therefore builds from classical portfolio theory to benchmark strategies, advanced allocation methods, ETF-specific evidence, and finally macro-conditioned allocation. The main argument is that macro-financial information may be useful because investment opportunities vary over time, but it should be used cautiously as regime or risk-state information rather than as a mechanical return-forecasting device.

The chapter is organized as follows. Section 1.1 reviews the foundations of portfolio allocation, including diversification, the risk–return trade-off, the role of the risk-free asset, and the relationship between systematic risk and broad market benchmarks. Section 1.2 discusses simple benchmark strategies and explains why they remain difficult to outperform out of sample. Section 1.3 examines portfolio constraints and robust estimation methods designed to reduce input uncertainty and unstable portfolio weights. Section 1.4 reviews risk-based, view-based, and conditional allocation methods. Section 1.5 summarizes the empirical evidence on ETF diversification, ETF universe selection, and direct ETF portfolio optimization. Section 1.6 reviews the literature on macroeconomic indicators, return predictability, conditional risk estimation, and regime-based allocation. Section 1.7 summarizes the main conclusions of the reviewed literature. Finally, Section 1.8 identifies the research gap, formulates the central research question and hypotheses, and connects the literature review with the empirical design developed in the next chapter.

1.1 Foundations of Portfolio Allocation

Modern portfolio theory begins with the contribution of Markowitz (1952), who formalized diversification as a risk–return problem. The key insight is that the risk of a portfolio depends not only on the risk of each asset separately, but also on how asset returns co-move. Let w denote the vector of portfolio weights, μ the vector of expected asset returns, and Σ the covariance matrix of asset returns. The expected return and variance of a portfolio are

$$\mathbb{E}(R_p) = w' \mu, \quad (1.1)$$

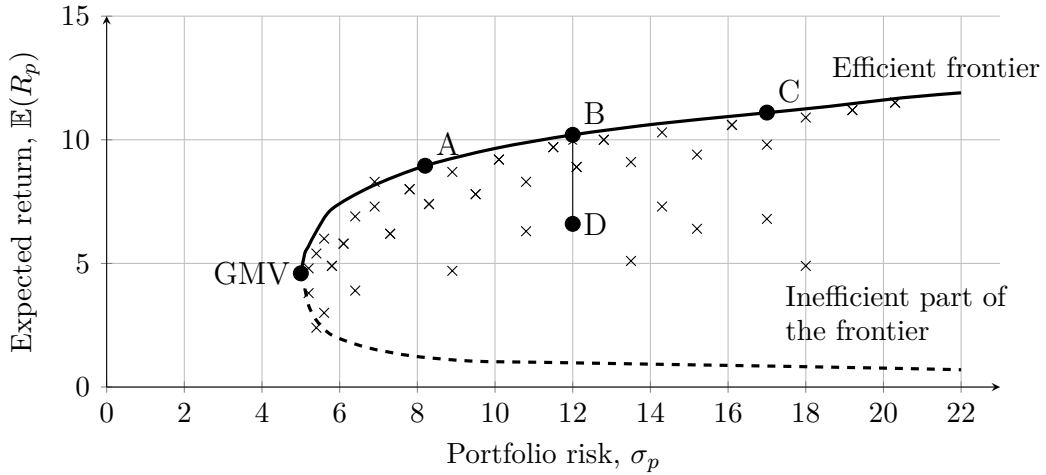
$$\sigma_p^2 = w' \Sigma w. \quad (1.2)$$

Equation (1.2) contains the central idea of diversification. A portfolio can reduce risk when assets are imperfectly correlated, even if the individual assets are risky. This logic is directly relevant for ETF allocation because ETFs already diversify within an index, while allocation across regions, sectors, factors, and asset classes can provide an additional layer of diversification.

The classical mean–variance problem can be stated as minimizing portfolio variance for a target expected return μ_p :

$$\begin{aligned} \min_w \quad & w' \Sigma w \\ \text{subject to} \quad & w' \mu = \mu_p, \\ & \mathbf{1}' w = 1. \end{aligned} \quad (1.3)$$

Solving this optimization problem for different target returns produces the efficient frontier (see Figure 1.1); full derivation is provided in Section A.1. Portfolios on the frontier offer the lowest variance for a given expected return, or equivalently the highest expected return for a given level of variance. The framework remains fundamental because it shows that portfolio selection is a joint problem: the attractiveness of an asset depends on its expected return, its volatility, and its covariance with the rest of the portfolio.



Source: Author for illustration purposes.

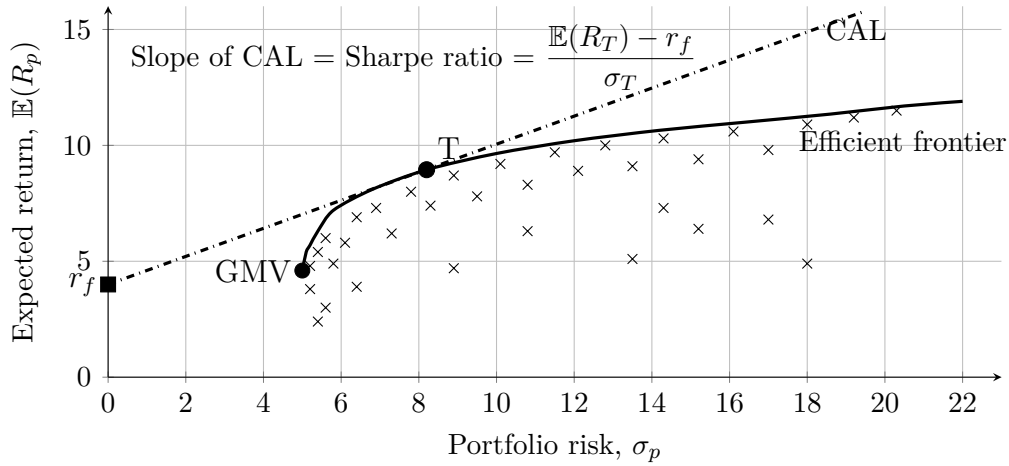
Figure 1.1: Illustration of the efficient frontier. The upper branch of the frontier represents efficient portfolios, while the lower branch is inefficient. The crosses represent feasible portfolios. Portfolios A, B, and C lie on the efficient frontier. Portfolio D is inefficient because it has the same level of risk as portfolio B but a lower expected return. The global minimum-variance portfolio (GMV) is the leftmost point on the frontier.

The same framework also exposes the main empirical weakness of portfolio optimization. Expected returns, variances, and covariances are not observed and must be estimated. Expected returns are especially noisy, and small errors in estimated inputs can lead to large changes in optimal weights. Without additional restrictions, mean–variance optimization may produce concentrated, unstable, or economically implausible portfolios. This practical weakness is central for the rest of the literature review. It explains why simple benchmarks are necessary, why constraints and shrinkage are widely used, and why risk-based methods often reduce reliance on expected-return forecasts.

The introduction of a risk-free asset is an important extension of the Markowitz framework, developed by Tobin (1958). Tobin’s original problem was formulated in the context of liquidity preference: investors choose between cash, which is safe in nominal terms, and interest-bearing assets, which offer yield but expose investors to capital gains and losses when interest rates change. The broader portfolio implication is that holding a safe asset is not irrational even when risky or interest-bearing assets offer higher expected returns. It is a way to control exposure to capital loss and uncertainty.

In the modern mean-variance setting, this idea leads to the separation theorem. Investors first identify the tangency portfolio of risky assets T , that is, the

risky portfolio with the highest Sharpe ratio¹. They then decide how much total wealth to allocate to this tangency portfolio and how much to allocate to the risk-free asset. All such combinations form the capital allocation line (CAL), which becomes the relevant efficient set once the risk-free asset is available. If borrowing at the risk-free rate is not permitted, the feasible part of the CAL is restricted to the segment between the risk-free asset and T . If borrowing is permitted, the CAL extends beyond T , representing leveraged positions in the tangency portfolio financed by borrowing at the risk-free rate. The technical derivation of Tobin's original cash–risky-asset problem, the tangency portfolio, and the optimal split between the risk-free asset and risky portfolio is provided in Section A.2 of the Appendix.



Source: Author for illustration purposes.

Figure 1.2: Illustration of the efficient frontier and the capital allocation line (CAL). The square point r_f denotes the risk-free asset, while point T denotes the tangency portfolio. The CAL is tangent to the efficient frontier at T and represents all portfolios formed as combinations of the risk-free asset and the tangency portfolio. Its slope equals the Sharpe ratio, that is, the excess return per unit of risk.

For the present thesis, the important implication is conceptual: portfolio composition and total risk exposure are separate decisions. This matters for retail investors because a practical ETF allocation framework must distinguish between the choice of ETFs and the investor's overall exposure to risky assets.

¹The Sharpe ratio is the excess expected return per unit of total risk. For a risky portfolio P , it is usually written as $(\mathbb{E}(R_P) - r_f)/\sigma_P$, where r_f is the risk-free rate and σ_P is the standard deviation of portfolio returns. In Tobin's analysis, the same object appears geometrically as the slope of the line connecting the safe asset to a risky portfolio. The term is used here as modern shorthand; Sharpe later formalized this performance measure as the reward-to-variability ratio (Sharpe 1966).

The Capital Asset Pricing Model (CAPM) extends this logic to market equilibrium. This extension is important for portfolio construction because it answers a question that remained open after the separation theorem: if investors combine a risk-free asset with an optimal risky portfolio, which risky portfolio should this be in equilibrium? Developed by Sharpe (1964), Lintner (1965), and Mossin (1966), the CAPM states that the expected return of asset i depends on its exposure to systematic market risk:

$$\mathbb{E}(R_i) = r_f + \beta_i [\mathbb{E}(R_M) - r_f], \quad (1.4)$$

where r_f is the risk-free rate, R_M is the return on the market portfolio, and β_i measures the asset's covariance with the market relative to the variance of the market:

$$\beta_i = \frac{\text{Cov}(R_i, R_M)}{\text{Var}(R_M)}. \quad (1.5)$$

The CAPM links portfolio theory with asset pricing by showing that, in equilibrium, only systematic risk is rewarded². Put simply, in the CAPM framework, investors earn expected returns above the risk-free rate only as compensation for systematic risk. Asset-specific risk can be diversified away and therefore does not command a risk premium.

For this thesis, the CAPM is relevant for two reasons. First, it provides the theoretical basis for broad market portfolios and market-capitalization-weighted benchmarks. If the market portfolio is the equilibrium risky portfolio, then a broad index ETF is a natural passive alternative to optimized allocation. Second, the model clarifies why diversification matters. The objective is not simply to hold many assets, but to manage exposure to systematic sources of risk. Later factor and ETF studies build on this idea by treating ETFs as investable vehicles for broad market, sector, regional, and factor exposures.

Classical portfolio theory therefore provides the foundation for the empirical analysis but not a complete solution. Markowitz establishes the diversification problem, Tobin separates risky portfolio choice from risk exposure, and the CAPM connects diversified portfolios to systematic risk and broad market benchmarks. At the same time, the direct empirical use of these models is limited by noisy inputs, restrictive assumptions, and implementation fric-

²Systematic risk is the part of asset risk that is related to broad market movements and cannot be eliminated through diversification, such as risk from recessions, inflation shocks, or changes in interest rates. Unsystematic risk is asset-specific risk, such as firm-level management problems, lawsuits, or product failures. In a diversified portfolio, these asset-specific shocks tend to offset each other, so the CAPM does not assign them a separate risk premium.

tions. This motivates the use of simple benchmarks and more robust allocation methods.

1.2 Simple Benchmarks

Empirical portfolio allocation requires credible benchmark strategies. A portfolio that looks optimal in theory may perform poorly out of sample when expected returns and covariances are estimated with error. Optimized portfolios may also require frequent rebalancing, generate high turnover, or produce concentrated weights that are difficult to justify for a retail investor. Simple benchmarks therefore serve two roles. They represent realistic alternatives that investors could actually implement, and they provide a strict test of whether additional model complexity adds value.

1.2.1 Equal-weight allocation

The equal-weighted portfolio, or $1/N$ portfolio, assigns the same weight to each asset:

$$w_i^{EW} = \frac{1}{N}, \quad i = 1, \dots, N, \quad (1.6)$$

where N is the number of assets in the investment universe. This rule does not require expected-return forecasts or a covariance matrix. Its strength is precisely its simplicity: it avoids the input-estimation problem that often harms optimized portfolios.

The empirical importance of the $1/N$ rule is shown by DeMiguel *et al.* (2009). They compare the naive diversification rule with 14 optimized portfolio models across seven empirical datasets and show that $1/N$ often performs better in out-of-sample Sharpe ratio terms. Their result does not imply that optimization is useless, but it sets a demanding standard. A sophisticated allocation method must improve enough out of sample to compensate for estimation risk, turnover, and implementation complexity.

Tu & Zhou (2011) provide a more constructive interpretation of the strong performance of $1/N$. Instead of treating naive and optimized allocation as mutually exclusive, they combine the equal-weighted portfolio with optimized portfolios. Their combined portfolio can be written as

$$\hat{w}_c = (1 - \delta)w_e + \delta\hat{w}, \quad (1.7)$$

where $w_e = \mathbf{1}_N/N$ is the equal-weight portfolio, $\hat{w}$ is an estimated optimized portfolio, and $\delta \in [0, 1]$ controls the weight placed on the optimized rule. Their evidence shows that such combinations can outperform $1/N$ alone. This suggests that the poor performance of standalone optimized portfolios does not necessarily mean that estimated return and covariance information has no value. Rather, optimization may contribute useful information when its sensitivity to estimation error is controlled by anchoring the resulting allocation to a robust benchmark. This finding motivates the use of constraints, shrinkage, and other stabilizing methods in the empirical analysis of this thesis.

1.2.2 Market-capitalization-weighted benchmark

A commonly used benchmark is the market-capitalization-weighted portfolio, which assigns larger weights to companies with higher market values. The weight of security i is defined as

$$w_i^{MC} = \frac{MC_i}{\sum_{j=1}^N MC_j}, \quad (1.8)$$

where MC_i denotes the market capitalization of security i , and N is the number of securities included in the portfolio. This weighting approach is transparent, inexpensive to implement, and closely related to the market portfolio in asset-pricing theory.

For a retail investor, a broad market ETF represents a practical passive alternative to active allocation across regions or sectors. In the empirical analysis, the Vanguard Total World Stock ETF (VT) is used as an external global-equity market reference.³ VT provides context for whether the multi-asset strategies justify their additional complexity relative to a single-fund equity solution. It uses a different opportunity set and therefore does not serve as a formal benchmark for Hypotheses introduced later in Section 1.8, but it is a useful reference for the practical question of whether a multi-asset allocation adds value for a retail investor.

³VT seeks to track the FTSE Global All Cap Index, a float-adjusted market-capitalization-weighted index covering large-, mid-, and small-cap companies across developed and emerging markets; see The Vanguard Group (2026).

1.3 Constraints and Robust Input Estimation

The instability of empirical portfolio optimization originates from two closely related sources: an excessively broad feasible set and noisy estimates of the required inputs. Portfolio constraints address the first problem by restricting extreme allocations, while shrinkage methods address the second by replacing unstable sample estimates with more regularized inputs. Because covariance matrices and expected returns present different estimation problems, they are discussed separately below.

1.3.1 Portfolio constraints

A direct way to improve empirical portfolio optimization is to restrict the feasible set of portfolios. Long-only constraints, the exclusion of leverage, and upper bounds on individual weights prevent the optimizer from taking extreme positions in response to small differences in estimated returns or covariances. Although these restrictions reduce the theoretical opportunity set, they may improve out-of-sample performance by limiting the effect of estimation error.

Jagannathan & Ma (2003) show that no-short-selling constraints can improve the risk properties of estimated optimal portfolios and may produce an effect similar to covariance shrinkage. By preventing extreme positive and negative weights, the constraints implicitly reduce the influence of unstable covariance estimates. This result is directly relevant to the present thesis because the retail-investor framework excludes short selling and leverage and imposes upper bounds on ETF weights.

1.3.2 Linear covariance shrinkage

Covariance shrinkage addresses estimation error more explicitly by accepting a controlled amount of bias in exchange for lower sampling variance. A linear shrinkage estimator combines the sample covariance matrix S with a more stable target matrix F :

$$\hat{\Sigma}^{LS} = \delta F + (1 - \delta)S, \quad \delta \in [0, 1], \quad (1.9)$$

where δ is the shrinkage intensity. The sample covariance matrix uses all pairwise information contained in the data but may have substantial estimation variance, particularly when the number of assets is large relative to the available

return history. The target matrix imposes a simpler covariance structure and can therefore be estimated more precisely. Combining the two estimators may produce a lower overall estimation error than relying on either one alone.

The choice of F determines the structure imposed on the covariance matrix. Common targets assume shared variances or covariances, constant correlations, or a factor-based covariance structure. A suitable target should contain substantially fewer free parameters than the sample covariance matrix, remain well-conditioned, and provide a reasonable approximation to the underlying covariance structure. It should be selected on economic and statistical grounds rather than according to its realized backtest performance.

For a given target, the Ledoit–Wolf procedure estimates the shrinkage intensity by minimizing the expected squared Frobenius distance between $\hat{\Sigma}^{LS}$ and the unknown population covariance matrix. The required population quantities are replaced with consistent sample estimates. Intuitively, more weight is assigned to the target when the sample covariance matrix is estimated imprecisely and when the target provides a relatively accurate approximation of the population covariance structure (Ledoit & Wolf 2004; 2022a).

1.3.3 Nonlinear covariance shrinkage

Linear shrinkage applies one common shrinkage intensity to the entire covariance matrix. Nonlinear shrinkage instead allows different parts of the estimated risk structure to receive different adjustments. Let the spectral decomposition of the sample covariance matrix be

$$S = U\Lambda U', \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_N), \quad (1.10)$$

where U contains the sample eigenvectors and λ_i are the corresponding sample eigenvalues. A rotation-equivariant nonlinear shrinkage estimator takes the form

$$\hat{\Sigma}^{NL} = U\tilde{\Lambda}U', \quad \tilde{\Lambda} = \text{diag}(\tilde{\lambda}_1, \dots, \tilde{\lambda}_N), \quad (1.11)$$

where each sample eigenvalue is replaced by an individually regularized estimate.

The eigenvalues represent estimated directions of portfolio risk, and their sampling errors are not uniform. In high-dimensional samples, the smallest eigenvalues may be severely underestimated and can therefore appear to the

optimizer as artificially low-risk directions. Linear shrinkage moves all eigenvalues toward a common target using the same intensity, whereas nonlinear shrinkage adjusts them according to their position in the estimated covariance spectrum (Ledoit & Wolf 2012; 2017; 2022a).

Quadratic Inverse Shrinkage (QIS) implements this principle by applying a data-dependent quadratic shrinkage rule in inverse-eigenvalue space and subsequently transforming the regularized values back into covariance eigenvalues (Ledoit & Wolf 2022b). Unlike linear shrinkage, it does not require the researcher to select a constant-correlation, factor-model, or other economic target matrix. QIS is also defined when the sample covariance matrix is singular, including cases in which the number of assets exceeds the effective number of observations. This property makes it relevant for a growing ETF universe in which the cross-sectional dimension may be similar to or larger than the rolling covariance-estimation window.

Empirically, Ledoit & Wolf (2017) find that nonlinear shrinkage dominates linear shrinkage for unrestricted global minimum-variance portfolios. However, the performance differences between the sample covariance matrix, linear shrinkage, and nonlinear shrinkage become considerably smaller when short selling is prohibited. A recent comparative preprint by Yadav *et al.* (2026) similarly finds that no covariance estimator dominates across all markets and investor objectives. Two-parameter linear shrinkage performs best most frequently overall, while QIS is selected in several high-dimensional applications.

Nonlinear shrinkage should therefore be interpreted as a robust covariance-estimation method rather than as an approach guaranteed to improve portfolio performance in every sample. Its usefulness depends on the size of the ETF universe, the length of the estimation window, the portfolio constraints, and the allocation objective.

1.3.4 Expected-return estimation

Expected returns are even more difficult to estimate reliably than covariances because the mean-return signal is weak relative to return volatility. Merton (1980) emphasizes the substantial amount of data required to estimate expected returns with reasonable precision. This estimation problem is one of the main reasons why unconstrained mean-variance portfolios often produce unstable and concentrated weights.

Shrinkage and Bayesian methods can reduce dependence on raw historical

averages by combining sample estimates with more stable prior information. The Bayes–Stein estimator of Jorion (1986), for example, shrinks individual sample means toward a common portfolio return. The Black–Litterman framework instead combines benchmark-implied equilibrium returns with structured views and their associated uncertainty (Black & Litterman 1992). Related Bayesian approaches use asset-pricing models to construct economically motivated priors (Pastor 2000; Pastor & Stambaugh 2000).

Factor models and macroeconomic variables can also provide expected-return inputs. The CAPM and the Fama–French models relate differences in expected returns to systematic exposures such as market, size, value, profitability, and investment factors (Fama & French 1993; 2015). Conditional asset-pricing studies further suggest that expected returns may vary with state variables such as dividend yields, interest-rate spreads, default spreads, inflation, and business-cycle indicators (Ferson & Harvey 1991; Cochrane 2008).

A comprehensive comparison of expected-return estimators is outside the scope of this thesis. The empirical analysis focuses on ETF allocation rather than firm-level return prediction and therefore does not employ large sets of company characteristics or high-dimensional forecasting models. Expected-return inputs are kept parsimonious, while the use of structured views and macro-financial conditioning is examined separately in the subsequent sections.

1.4 Advanced Portfolio Allocation Methods

Building on the preceding discussion of portfolio constraints and robust input estimation, this section reviews allocation methods that extend the portfolio-construction framework. Risk-based methods reduce reliance on uncertain expected-return estimates by allocating capital according to portfolio risk and diversification. Black–Litterman combines equilibrium-implied returns with structured views, while conditional allocation allows expected returns, risk estimates, or portfolio weights to vary with observable characteristics or market states. These approaches differ in design, but all seek to improve the stability and practical usefulness of portfolio allocation relative to unrestricted classical mean–variance optimization.

1.4.1 Risk-based allocation

Risk-based methods reduce the role of expected-return forecasts. They are motivated by the view that risk and covariance estimates are more stable than expected returns. The most direct example is the global minimum-variance portfolio, which solves

$$\begin{aligned} \min_w \quad & w' \Sigma w \\ \text{subject to} \quad & \sum_{i=1}^N w_i = 1, \end{aligned} \tag{1.12}$$

usually with additional long-only or maximum-weight constraints in empirical applications. Since this method does not require expected-return estimates, it can be more robust than full mean–variance optimization when return forecasts are noisy.

Risk parity and equal risk contribution portfolios follow a related logic. Instead of equalizing capital weights, they aim to equalize the contribution of each asset to total portfolio risk. Maillard *et al.* (2010) formalize equal risk contribution portfolios and show that they are connected to the broader objective of risk diversification. This matters because a portfolio may look diversified in capital weights while most of its volatility comes from one asset class or one common factor.

Maximum-diversification portfolios also focus on the risk structure of the asset universe. Choueifaty & Coignard (2008) propose maximizing the diversification ratio,

$$DR(w) = \frac{w' \sigma}{\sqrt{w' \Sigma w}}, \tag{1.13}$$

where σ is the vector of individual asset volatilities. The method seeks portfolios that benefit strongly from imperfect correlations across assets. In an ETF universe with regional, sector, and factor exposures, this approach may be useful because correlations can differ substantially across market conditions.

Downside-risk methods modify the definition of risk. Classical variance treats positive and negative deviations symmetrically, while investors are often more concerned with losses. Conditional Value-at-Risk (CVaR) optimization focuses on expected losses in the tail of the return distribution. Rockafellar & Uryasev (2000) show that minimizing CVaR can be formulated as a tractable convex optimization problem and, under a finite set of return scenarios, solved using linear programming. CVaR is not included in the empirical analysis because, with monthly rolling windows, a 95% CVaR estimate would be based

on only a few worst observations, making portfolio weights highly sensitive to individual crisis months.

1.4.2 View-based and conditional allocation

View-based and conditional methods introduce forward-looking information into portfolio construction. These approaches are relevant because purely historical estimates may not capture changes in economic conditions, monetary policy, inflation, or market regimes. They also provide a bridge between standard optimization and the macro-conditioned allocation framework considered later in this thesis.

The Black–Litterman model of Black & Litterman (1992) addresses the instability of expected-return estimates by combining market-implied equilibrium returns with structured investor views. This makes the framework relevant for macro-informed allocation, since macro-financial indicators can be translated into views on broad asset classes, regions, sectors, or risk exposures. For example, elevated financial stress may justify a negative view on cyclical sectors, while declining interest rates may support a positive view on duration-sensitive assets. The credibility of the resulting allocation, however, depends on whether these views are formulated *ex ante*, based exclusively on lagged information, and assigned realistic confidence levels. Without such discipline, Black–Litterman may merely provide a formal structure for overfitted return expectations.

Empirical evidence from Bessler *et al.* (2017) supports this cautious interpretation. They implement a sample-based Black–Litterman model in a global multi-asset setting consisting of stock, bond, and commodity indices over January 1993–December 2011. Their views are not discretionary analyst forecasts or macro narratives. Instead, they use 12-month rolling historical mean return estimates as the subjective views and combine them with implied returns derived from strategic benchmark weights. The reliability of each view is measured through the variance of past forecast errors, while the covariance matrix is estimated with a 36-month rolling window. The model is rebalanced monthly and is implemented with realistic restrictions, including a budget constraint, no short selling, transaction costs, and volatility limits for conservative, moderate, and offensive investor types.

The reported results provide more concrete support for Black–Litterman than the original methodological paper. In the base case, Bessler *et al.* (2017)

report net Sharpe ratios of 0.91, 0.75, and 0.74 for conservative, moderate, and offensive Black–Litterman portfolios, respectively. The corresponding mean–variance Sharpe ratios are 0.68, 0.49, and 0.53, while the equally weighted $1/N$ benchmark has a Sharpe ratio of 0.27. The same pattern appears in maximum drawdown: Black–Litterman drawdowns are 5.99%, 9.46%, and 14.39%, compared with 9.41%, 16.18%, and 24.59% for mean–variance optimization and 40.59% for the $1/N$ benchmark. The authors also find that Black–Litterman portfolios are generally less concentrated and have lower turnover than mean–variance portfolios.

The subperiod evidence is also relevant. For the moderate investor, the Black–Litterman strategy performs especially well in recessionary periods: the net Sharpe ratio is 0.92 during February 2001–June 2004 compared with 0.39 for mean–variance optimization, and 0.63 during March 2008–December 2011 compared with 0.27 for mean–variance optimization. In expansionary periods, however, the advantage is smaller and not always statistically significant; in one bullish subperiod, naive benchmark portfolios perform competitively. This result is useful for the present thesis because it suggests that structured views may be most valuable when market conditions are stressed or changing, but it also warns that view-based allocation does not dominate uniformly.

For this thesis, the main lesson from Bessler *et al.* (2017) is that forward-looking information may improve allocation when views are constructed systematically, their uncertainty is measured, and implementation constraints are imposed. A macro-conditioned ETF version would therefore need to define macro views or risk-state tilts in a transparent, lagged, and testable way.

Conditional allocation rules provide a more direct way to link information variables to allocation. Brandt *et al.* (2009) propose modelling portfolio weights as functions of observable asset characteristics.⁴ Their empirical application uses U.S. stock returns from CRSP and firm-level characteristics from the CRSP–Compustat database over 1964–2002. The baseline allocation rule overweights small firms, value firms, and past winners, and underweights large firms, growth firms, and past losers. For an investor with relative risk aversion

⁴The investor chooses portfolio weights $w_{i,t}$ to maximize conditional expected utility, $\mathbb{E}_t[u(r_{p,t+1})] = \mathbb{E}_t[u(\sum_{i=1}^{N_t} w_{i,t} r_{i,t+1})]$. The characteristic-based allocation rule writes each weight as a function of characteristics, $w_{i,t} = f(x_{i,t}; \theta)$, and in the simple linear case uses $w_{i,t} = \bar{w}_{i,t} + N_t^{-1} \theta' \hat{x}_{i,t}$. Here $\bar{w}_{i,t}$ is the benchmark weight, $\hat{x}_{i,t}$ is the standardized characteristic vector, and θ is the vector of coefficients that determines how characteristics tilt the portfolio away from the benchmark.

equal to five, they report a certainty-equivalent gain relative to the market portfolio of 11.1% in sample and 5.4% out of sample.

The paper is relevant for this thesis because the authors also allow weight coefficients to depend on state variables, such as the slope of the yield curve. This shows how economic conditions can affect portfolio weights directly. At the same time, the approach is complex and creates a high risk of overfitting if many characteristics, macro variables, lags, and transformations are tested. For a retail ETF setting, the broader lesson is that conditioning information can be useful, but it must be implemented parsimoniously and tested against simple benchmarks.

Overall, advanced allocation methods do not provide a single universally superior solution. Their value depends on which weakness they address and how they perform out of sample. Constraints and shrinkage reduce instability; risk-based portfolios reduce reliance on expected-return forecasts; Black–Litterman and conditional allocation rules introduce forward-looking information. These methods motivate the empirical comparison in this thesis, but they also reinforce the need for simple benchmarks and careful robustness checks.

1.5 ETF Portfolio Allocation and Diversification

ETFs are especially suitable for studying practical portfolio allocation. They provide liquid and transparent exposure to broad markets, sectors, regions, factors, bonds, commodities, and other asset classes. As Lettau & Madhavan (2018) explain, ETFs combine features of index mutual funds and exchange-traded securities, which makes them accessible portfolio building blocks for individual investors. For this thesis, the ETF literature is important because it connects the theoretical allocation methods reviewed above with an implementable investment universe.

The main question in the ETF literature is not simply whether ETFs are diversified instruments. It is whether allocation across ETFs can improve performance relative to simple passive alternatives. The evidence suggests three broad conclusions. First, ETFs can provide efficient access to international and systematic exposures. Second, the choice of ETF universe may be as important as the allocation model. Third, sophisticated optimization often improves performance only modestly and depends strongly on estimation windows, constraints, and benchmark choice.

1.5.1 International and factor-based ETF diversification

Huang & Lin (2011) examine whether ETFs provide effective international diversification. They compare direct investment in foreign market index portfolios with indirect investment through 19 iShares ETFs traded on NYSE Arca, covering European, American, Asian, and African markets. The authors construct regional optimal portfolios using mean–variance optimization, Sharpe-ratio analysis, and a Value-at-Risk framework under both normal and non-normal distribution assumptions. For the sample from June 2003 to March 2009, they report that international ETF portfolios achieve higher Sharpe ratios than the corresponding target market indices, especially in emerging markets. They also find that portfolios combining the S&P 500 with foreign market exposure perform better than portfolios invested only in the S&P 500, including during the subprime crisis period.

This paper supports the idea that ETFs can be practical vehicles for international diversification. At the same time, its results should not be interpreted as general proof that optimized ETF portfolios always dominate broad benchmarks. The sample period is relatively short and includes a specific crisis episode. For this thesis, the main value of the paper is that it motivates the use of ETFs for regional diversification and provides empirical results that can later be compared with the performance of broader ETF allocation strategies.

A different perspective is provided by Puelz *et al.* (2015), who examine whether a small set of ETFs can replicate major dimensions of return variation in U.S. equity markets. They treat ETFs as investable proxies for eight well-known academic factor and anomaly portfolios: market, size, value, profitability, investment, momentum, short-term reversal, and long-term reversal. The first five are related to the Fama–French asset-pricing framework (Fama & French 1992; 2015), while momentum and reversal come from the anomaly literature (Jegadeesh & Titman 1993; Carhart 1997; Lehmann 1990; De Bondt & Thaler 1985). A brief description of these factor and anomaly portfolios is provided in Section B.1.

Using Bayesian variable selection, Puelz *et al.* (2015) show that a small number of broad-market and style ETFs can capture much of the covariance structure of factor portfolios that are difficult for retail investors to implement directly. This supports the use of ETFs as practical building blocks for systematic exposures. The result is also relevant for the ETF-universe decision in this thesis: if a limited number of liquid ETFs can approximate broad factor

dimensions, the empirical design can focus on implementable ETFs rather than theoretical long–short factor portfolios.

Bessler *et al.* (2021) provide particularly relevant evidence because they compare factor-based and sector-based allocation strategies. The study uses investable U.S. factor and sector indices that can be implemented through low-cost ETFs. The factor universe consists of six MSCI USA factor indices: momentum, minimum volatility, value, quality, size, and carry/high dividend yield. The sector universe consists of ten S&P 500 sector indices. Over the out-of-sample period from May 2007 to November 2020, the authors compare equal weighting, risk parity, minimum variance, mean–variance optimization, Bayes–Stein, and Black–Litterman.

Their central result is that factor portfolios generally outperform sector portfolios over the full sample. The equal-weighted factor portfolio achieves an annualized Sharpe ratio of 1.17, compared with 0.90 for the equal-weighted sector portfolio. The factor $1/N$ portfolio produces approximately 15% annualized return with 13% volatility, while the sector $1/N$ portfolio produces approximately 13% annualized return with 14% volatility. This suggests that the investment universe itself can have a large effect on performance.

The evidence for sophisticated allocation is more modest. The best long-only mean–variance and Bayes–Stein factor portfolios achieve Sharpe ratios around 1.23–1.26, only moderately above the factor $1/N$ Sharpe ratio of 1.17. Black–Litterman performs close to $1/N$, with long-only Sharpe ratios around 1.15–1.19. Thus, the paper does not show a large and stable dominance of advanced allocation methods over simple diversification. Instead, it suggests that factor ETFs may be more attractive building blocks than sector ETFs in some samples, while the incremental benefit of optimization is limited.

The paper also shows that relative performance varies across market regimes. Factor portfolios dominate over the longer full sample and during normal periods, while sector portfolios may perform better during crisis periods because of stronger diversification benefits. This point is directly connected to the macro-conditioned part of the thesis. If the relative attractiveness of sectors, factors, or regions changes with market conditions, then allocation rules that condition on macro-financial risk states may be worth testing.

1.5.2 Direct ETF portfolio optimization

While the previous studies motivate ETFs as vehicles for international and factor exposure, Emekter *et al.* (2024) study the direct construction of efficient portfolios using ETFs exclusively. They use monthly data for 34 U.S.-listed ETFs over 2007–2017 and apply classical Markowitz mean–variance optimization to construct portfolios for different risk-aversion parameters. The ETF universe includes equity, fixed income, real estate, commodity, and international exposure.

Their findings show both the promise and the fragility of direct ETF optimization. In an ex-post setting, where realized returns are known, optimized portfolios achieve competitive risk-adjusted performance. In the more realistic ex-ante setting, where weights are estimated using historical rolling windows and then tested out of sample, the performance is weaker. Standard mean–variance portfolios do not clearly outperform the best individual ETFs, and the results are sensitive to the estimation window and rebalancing frequency. The authors also test a Ledoit–Wolf covariance estimator, but the improvement is only marginal.

The most practically useful element of Emekter *et al.* (2024) is the average-weight approach. Instead of implementing the unstable weights generated in each rolling window, the authors average the optimized ETF weights over time and use these average weights to construct portfolios. This smoothing procedure improves stability and produces stronger out-of-sample results than the standard ex-ante Markowitz portfolios. They report that the average-weight approach improves the return of the best-performing portfolio by approximately 370 basis points and increases the Sharpe ratio.

This finding is relevant for the thesis because it shows that ETF optimization may require weight stabilization. However, the study also has limits. The sample contains 34 U.S.-listed ETFs, the strongest portfolios are influenced by U.S. large-cap growth exposure such as QQQ, and the evidence does not establish that optimization consistently dominates simple ETF choices or broad benchmarks. The main lesson is therefore not that Markowitz optimization should be applied mechanically, but that ETF allocation methods must be evaluated out of sample and under realistic implementation rules.

Petronio *et al.* (2014) compare portfolio-selection criteria in the U.S. and European ETF markets using daily returns for 702 U.S. and 587 European ETFs from January 2006 to April 2012. Portfolios are estimated from rolling

125-day windows and recalibrated every five trading days, producing 288 portfolio decisions. At each recalibration, the authors exclude ETFs with average daily trading volume below 10,000 units, preselect the 100 ETFs with the highest Sharpe ratios, reduce the return dimension through principal component analysis, and construct long-only portfolios after proportional transaction costs of five basis points. They compare the Sharpe ratio with the Rachev ratio of Biglova *et al.* (2004), which relates expected upper-tail gains to expected lower-tail losses, and a new performance measure based on conditional expected value (PCEV).

In the U.S. sample, the Sharpe-based portfolio achieves the highest mean daily return, 0.0225%, and the lowest volatility, 0.6215%. In Europe, PCEV slightly outperforms Sharpe in mean return, 0.0183% versus 0.0178%, while Sharpe again delivers the lowest volatility. Overall, Sharpe-based optimization performs best, although PCEV remains competitive in some periods. The authors relate this result to the comparatively regular distribution of ETF returns, as normality is not rejected in approximately 57% of the recalibration samples. The findings should nevertheless be interpreted cautiously because ETF selection is based on prior Sharpe ratios and the PCEV benchmark is itself Sharpe-optimal, which may favour the Sharpe strategy.

Overall, the ETF literature supports the use of ETFs as practical instruments for diversified allocation. ETFs make it possible to build portfolios across regions, sectors, factors, and asset classes without selecting individual securities. However, the literature also shows that sophisticated optimization does not automatically dominate simple rules. The ETF universe, estimation window, constraints, weight stability, and benchmark choice often matter as much as the allocation model. This motivates the empirical design of the thesis, which compares simple, optimized, and macro-conditioned ETF strategies under long-only retail-investor constraints.

1.6 Macroeconomic Indicators and Conditional Allocation

The macro-financial literature provides the main motivation for the conditional part of the thesis. Macroeconomic and financial conditions may affect expected returns, volatility, correlations, factor exposures, and regime probabilities. Relevant variables include inflation, interest rates, yield spreads, credit spreads,

financial stress, volatility indices, industrial production, unemployment, and real rates. The key issue is how to use such information. The literature does not support the simple claim that macro variables reliably forecast asset returns out of sample. A more defensible interpretation is that macro-financial indicators may help identify risk states or regimes in which portfolio risks and diversification benefits differ.

1.6.1 Macroeconomic State Variables and Time-Varying Portfolio Inputs

Campbell *et al.* (2003) provide a theoretical foundation for portfolio inputs that vary with economic conditions. In their multivariate strategic asset allocation model, asset returns and state variables such as dividend yields, interest rates, yield spreads, and inflation evolve jointly. This implies that long-term optimal portfolio weights may depend on macro-financial conditions. The paper is mainly a theoretical and intertemporal portfolio-choice contribution rather than an ETF allocation study. Therefore, it motivates the use of state variables but does not provide direct evidence that macro-conditioned ETF strategies outperform simple benchmarks.

Lettau & Ludvigson (2001) provide empirical evidence that aggregate state variables may contain information about future stock returns. Their consumption–wealth ratio, denoted cay_t , measures deviations from the long-run relationship among aggregate consumption, asset wealth, and labor income. They show that cay_t predicts real and excess stock returns, particularly at short and intermediate horizons. For this thesis, the paper supports the idea that aggregate macro-financial conditions can contain information relevant for allocation. At the same time, it studies return predictability rather than ETF portfolio performance and does not compare a macro-based allocation strategy with retail-investor benchmarks.

Nezafat & Slavík (2021) offer a structural explanation for why liquidity and business-cycle conditions may matter for asset prices. Their production-based model with financially constrained firms shows how liquidity shocks can affect equity demand, investment, equity prices, the equity premium, and the risk-free rate. The paper supports the broader economic mechanism behind conditional allocation: expected returns and risk premia may vary with liquidity, investment, output, and interest-rate conditions.

1.6.2 Limits of direct return prediction

The strongest caution against naive macro-based return forecasting is provided by Welch & Goyal (2008). They re-examine many variables proposed as predictors of the U.S. equity premium, including dividend-price ratios, dividend yields, earnings-price ratios, payout ratios, interest rates, term and default spreads, inflation, book-to-market ratios, stock variance, investment-capital ratios, issuance variables, and the consumption–wealth–income ratio. Their main finding is that most predictors perform poorly out of sample and often fail to beat the historical mean equity premium benchmark. Apparent predictability is unstable and often concentrated in particular historical episodes.

This evidence has direct implications for the empirical design of the thesis. A fragile approach would be to regress next-period ETF returns on macro variables and use the fitted values as expected returns in a mean–variance optimizer. Such a procedure would combine two sources of instability: weak return predictability and optimizer sensitivity to expected-return inputs. It would also increase the risk of overfitting, look-ahead bias, and excessive turnover. Therefore, the macro component of this thesis is framed more cautiously. Lagged macro-financial indicators are used as conditioning information for regimes or risk states, not as a guarantee of precise return forecasts.

1.6.3 Regime-switching and conditional allocation

Regime-switching models provide another way to incorporate changing market conditions. Instead of asking whether one variable predicts next-period returns, these models allow returns to be generated by different states with different expected returns, volatilities, and correlations. This is relevant for ETF allocation because diversification benefits often change during stress periods, when correlations across risky assets may rise.

Ang & Bekaert (2003) examine regime changes in international asset allocation. They argue that equity returns are characterized by persistent normal and bear-market regimes, where bear markets have lower expected returns, higher volatility, and higher cross-market correlations. In their first application, they use monthly MSCI returns for six developed equity regions in U.S. dollars from February 1975 to December 2000. The estimated model identifies a normal regime with monthly world excess return of 0.90% and volatility of 2.81%, and a high-volatility regime with monthly excess return of 0.13% and volatility of 5.04%. In an out-of-sample test from January 1985 to December 2000, the

regime-switching all-equity strategy achieves a Sharpe ratio of 1.07, compared with 0.52 for the world market portfolio and 0.90 for a non-regime-dependent mean–variance strategy. The result is promising, but the authors allow unconstrained positions and do not include transaction costs, which limits direct applicability to retail ETF allocation.

Guidolin & Timmermann (2007) provide further evidence that regime-dependent returns, volatilities, and correlations can matter for allocation. They estimate a multivariate Markov-switching model using monthly U.S. data on large-cap stocks, small-cap stocks, and 10-year Treasury bonds, with excess returns measured relative to 30-day Treasury bills. The regimes are latent states, interpreted as crash, slow growth, bull, and recovery regimes. Portfolio weights differ substantially across regimes as investors update filtered state probabilities. In pseudo-real-time out-of-sample tests over 1980–1999, the four-state model performs favourably relative to IID, VAR, minimum-variance, and tangency portfolio benchmarks when evaluated using realized utility. The study is technically demanding and uses a small U.S. asset universe, but it supports the idea that regime-dependent investment opportunities may improve allocation.

1.6.4 Regime-switching ETF allocation with macro-financial factors

The closest ETF-specific contribution is Ma *et al.* (2011), who develop a regime-dependent portfolio optimization model for U.S. sector ETFs. Their model combines ETF returns, macro-financial factors, style factors, and hidden market regimes. ETF returns are specified through a regime-switching linear factor structure:

$$R_t = A_{s_t} + B_{s_t}F_t + \Gamma_{s_t}e_t, \quad (1.14)$$

where R_t is the vector of ETF returns, F_t is a vector of risk factors, and s_t is the unobserved market regime. The parameters A_{s_t} , B_{s_t} , and Γ_{s_t} are regime-specific, allowing ETF alphas, factor sensitivities, and residual risks to change across states. The empirical application uses nine U.S. sector SPDR ETFs and six risk factors: the market factor, size factor, value factor, VIX, yield spread, and credit spread.

The regime process is modelled as a hidden Markov chain. Transition probabilities are given by

$$p_{mn} = \Pr(s_{t+1} = n \mid s_t = m), \quad (1.15)$$

which allows expected returns and risks to be calculated as probability-weighted outcomes across possible future states. Using the Bayesian information criterion, the authors identify three regimes, interpreted as bear, transition, and bull regimes. This framework is important because it allows the same macro or style factor to have different implications depending on the market state.

The portfolio step is a regime-dependent stochastic linear programming problem. The authors maximize expected portfolio alpha while constraining exposure to risk factors. The expected alpha conditional on the previous regime is written as

$$\Psi_m(w_t) = \sum_{n=1}^N \sum_{i=1}^I w_{ti} \alpha_{in} p_{mn}, \quad (1.16)$$

while portfolio beta exposure to each factor is restricted within a tolerance band:

$$-\delta_n \leq \sum_{i=1}^I w_{ti} \beta_{ijn} \leq \delta_n. \quad (1.17)$$

The model is therefore best understood as a regime-switching factor model combined with a market-neutral alpha-maximization strategy. It separates expected excess return from systematic factor exposure and limits unintended exposure to market, style, volatility, yield-spread, and credit-spread risks.

Empirically, Ma *et al.* (2011) report that the regime-dependent strategy generally outperforms an equally weighted sector ETF benchmark. The out-of-sample test covers 150 trading days from February 27 to September 30, 2009. Across different levels of factor-risk tolerance, the regime-dependent portfolio achieves higher mean excess returns and higher cumulative returns than the benchmark, although with higher volatility. The results are strongest in the bull regime, where both return and risk-adjusted performance improve relative to equal weighting. In the transition regime, the advantage is weaker and often close to the benchmark.

The paper is central for this thesis because it shows that macro-financial and style factors can be incorporated into ETF allocation through regimes. At the same time, its empirical design has important limits. The ETF universe is narrow and U.S.-specific, consisting of nine sector ETFs. The out-of-sample evaluation is short and does not include a bear regime. The strategy allows short selling and is designed as a long-short market-neutral allocation, which is not fully realistic for many retail investors. Finally, the benchmark comparison is limited mainly to equal weighting. Therefore, the paper motivates regime-conditioned ETF allocation but does not settle whether such information adds

practical value in a broader, long-only, retail-oriented ETF framework.

This thesis builds on the motivation of Ma *et al.* (2011) without attempting to replicate their model exactly. The empirical analysis tests a macro-conditioned ETF allocation framework in which lagged macro-financial indicators classify or proxy market regimes and risk states. These states may affect allocation through regime-conditioned inputs, constraints, or allocation tilts. The exact procedure is specified in the methodology chapter. The key difference from Ma *et al.* (2011) is the practical focus: broader ETF exposure, long-only and no-leverage constraints, longer out-of-sample evaluation, and comparison with simple benchmarks such as $1/N$ and broad market allocation.

1.6.5 Practical implementation risks

The macro-financial literature supports the relevance of time-varying conditions, but it also highlights implementation risks. Kollár & Schmieder (2019) provide a practical macro-financial-cycle perspective on allocation. Their study suggests that macro-financial cycle indicators may be useful when combined rather than interpreted in isolation. At the same time, they emphasize difficulties that are directly relevant for implementable strategies: model complexity, turning-point identification, data availability, and the robustness of real-time signals.

These issues are particularly important for retail ETF allocation. Macro variables are often released with delays and revised after publication. A back-test that uses final revised data may overstate the information actually available to investors at the time. Testing many indicators, lags, transformations, and regime definitions also creates overfitting risk. A strategy may appear successful because it was selected after observing the sample rather than because the relationship is stable. Regime turning points are difficult to identify in real time, and a model that reacts too slowly may miss the most important periods, while a model that reacts too quickly may trade excessively.

Transaction costs, turnover, and simplicity are also central. Even liquid ETFs are not costless to trade, and frequent reallocations may reduce net performance. More importantly, a strategy designed for a retail investor must remain transparent and maintainable. A macro-conditioned strategy therefore needs to be judged not only by return and Sharpe ratio, but also by drawdown, turnover, robustness across samples, and performance relative to simple passive alternatives.

The macro literature therefore motivates a cautious empirical question. Macro-financial variables may contain useful information about risk states, but they should not be assumed to improve allocation automatically. The thesis tests whether such information adds practical value in a constrained ETF setting. This framing is consistent with the evidence: macro indicators may matter because investment opportunities vary over time, but direct return predictability is unstable and must be evaluated carefully out of sample.

1.7 Summary

The reviewed literature establishes the theoretical foundation for diversified portfolio allocation while also revealing the difficulties of its empirical implementation. Markowitz (1952) formalizes diversification through expected returns, variances, and covariances, Tobin (1958) separates risky portfolio selection from the investor's overall exposure to risk, and the CAPM (Sharpe 1964; Lintner 1965; Mossin 1966) links diversified portfolios to systematic risk and broad market benchmarks. However, the practical application of these models is limited by noisy inputs, restrictive assumptions, and unstable portfolio weights.

Simple allocation rules therefore remain essential benchmarks. The equal-weighted portfolio is difficult to outperform consistently out of sample because it avoids the estimation errors that affect optimized portfolios (DeMiguel *et al.* 2009). Nevertheless, the evidence of Tu & Zhou (2011) suggests that estimated return and covariance information may still contain useful allocation signals when dependence on unstable optimized weights is reduced. Optimization should therefore not be rejected entirely, but its value must be evaluated within a framework that controls estimation risk and compares performance with transparent and implementable alternatives.

Advanced allocation methods respond to specific weaknesses of classical optimization. Constraints and shrinkage reduce estimation error and extreme weights. Risk-based methods such as minimum variance, risk parity, maximum diversification, and CVaR reduce reliance on expected-return forecasts or focus on alternative risk concepts. Black–Litterman and conditional allocation policies introduce structured forward-looking views or state-dependent information. However, the literature does not imply that advanced methods always dominate simple rules. Performance depends on the quality of inputs, the sta-

bility of weights, the constraints imposed, and the out-of-sample benchmark comparison.

The ETF literature supports ETFs as practical instruments for diversified portfolio construction. Huang & Lin (2011) show that international ETF portfolios can improve diversification relative to target market indices in their sample. Puelz *et al.* (2015) show that a small number of broad-market and style ETFs can approximate important factor dimensions. Bessler *et al.* (2021) demonstrate that factor ETF universes may offer stronger risk–return properties than sector universes, while also showing that sophisticated allocation adds only modest value beyond simple allocation in some cases. Emekter *et al.* (2024) show that direct ETF optimization is feasible but fragile, and that average optimized weights can improve stability. Overall, ETF studies suggest that ETF allocation is empirically meaningful, but performance depends strongly on ETF-universe selection, constraints, estimation windows, and benchmark design.

The macro-financial literature provides the motivation for conditional allocation but also the main caution. Campbell *et al.* (2003) and Lettau & Ludvigson (2001) support the view that portfolio inputs and optimal allocations may change with macro-financial conditions. Welch & Goyal (2008) show that direct equity premium prediction is unstable out of sample, which warns against naive macro-based return forecasting. Flavin & Wickens (2003) suggest that macro variables may be more useful through conditional covariance and risk estimation. Ang & Bekaert (2003) and Guidolin & Timmermann (2007) show that regimes can affect expected returns, volatilities, correlations, and optimal allocations. Ma *et al.* (2011) provide the closest ETF-specific regime-switching application, but their study is limited by a narrow U.S. sector ETF universe, a short out-of-sample period, short selling, and limited benchmark comparison. Taken together, the literature supports testing macro-conditioned ETF allocation, but it does not establish that such strategies will outperform simpler alternatives under realistic investment constraints.

1.8 Research Gap, Research Question, and Hypotheses

The reviewed literature leaves a specific empirical gap. Existing studies separately demonstrate that ETFs can be used for diversified portfolio construc-

tion, that robust allocation methods may reduce the instability of classical optimization, and that macro-financial conditions may affect expected returns, volatilities, correlations, and optimal portfolio weights. However, the reviewed evidence does not establish whether observable and lagged macro-financial information used to define regime or risk states provides incremental allocation value in a broader international and cross-sector ETF universe. ETF optimization studies generally rely on unconditional return information, while regime-switching studies often use narrow asset universes, latent return-based states, short out-of-sample periods, unconstrained positions, or limited benchmark comparisons. The practical performance of these approaches under long-only retail-investor constraints, regular contributions, and transaction costs also remains insufficiently examined.

This thesis addresses the gap through a three-level empirical design. The first level tests whether unconditional classical and risk-based allocation methods improve performance relative to simple benchmark allocations when all strategies use the same growing ETF universe and information set. The second level isolates the incremental contribution of macro-financial information by comparing macro-conditioned strategies with otherwise comparable strategies based on historical returns or unconditional return moments. The third level evaluates whether any performance advantage remains economically meaningful when the strategies are implemented through regular contributions and proportional transaction costs and are compared with a distribution of feasible retail-investor outcomes.

The central research question is therefore whether increasingly sophisticated portfolio-allocation methods—and macro-financial conditioning in particular—add practical value relative to simple ETF allocation rules for a constrained retail investor.

This question is examined through the following hypotheses:

Hypothesis 1. Classical and risk-based portfolio allocation strategies achieve higher out-of-sample risk-adjusted performance than simple benchmark allocations when evaluated on the same growing international and cross-sector ETF universe.

Hypothesis 2. Macro-conditioned allocation strategies achieve higher out-of-sample risk-adjusted performance than otherwise comparable strategies based only on historical returns or unconditional return moments.

Hypothesis 3. Strategy-specific performance advantages observed in the frictionless comparisons remain positive in net time-weighted return and terminal wealth after monthly contributions and proportional transaction costs are incorporated.

The hypotheses correspond to three increasingly demanding tests of allocation value. Hypothesis 1 examines whether unconditional estimation and optimization add value beyond simple policy-weighted and equal-weight benchmark allocations. Hypothesis 2 examines the incremental contribution of macro-financial information through matched comparisons, including historical-view versus macro-view Black–Litterman portfolios and unconditional versus regime-conditioned sleeve allocation. Hypothesis 3 evaluates whether previously favourable frictionless comparisons remain positive in net time-weighted return and terminal wealth after monthly contributions and trading costs are incorporated. A separate conditional distributional comparison then locates each implemented terminal outcome within the pre-specified Random DCA design.

The next chapter translates these hypotheses into a complete empirical design. It defines the ETF universe, information set, estimation windows, portfolio constraints, contribution and transaction-cost assumptions, strategy specifications, benchmark construction, and performance-evaluation procedure. It also presents the strategy comparisons used to test each hypothesis.

Chapter 2

Data and Methodology

This chapter formalizes the three hypotheses stated in Section 1.8. It first introduces the common investment setting and then specifies the hypothesis-specific models and implementation features. This structure makes the strategy results comparable and translates each hypothesis into a corresponding empirical comparison.

Section 2.1 fixes the elements that should not influence the comparison between strategies. Hypotheses 1 and 2 therefore share the same growing ETF universe, monthly information boundary, portfolio constraints, estimation methodology, and frictionless wealth path beginning with one unit of initial wealth. Holding these features constant is necessary to ensure that differences in performance can be attributed to the allocation rules themselves rather than to unequal investment opportunities, timing conventions, or modelling inputs. Hypothesis 3 subsequently retains the previously specified target-weight rules but changes the wealth-accounting environment by introducing monthly contributions and transaction costs. The same strategies can therefore be evaluated first as abstract portfolio-allocation rules and then under conditions that more closely resemble their practical use by a retail investor.

The design of Hypothesis 1 reflects the need to distinguish the value added by portfolio optimization from the consequences of benchmark construction. The model selection rationale for Hypothesis 1 is straightforward: this thesis employs classical mean–variance optimization and the three covariance-based strategies—GMV, equal risk contribution, and maximum diversification—and tests whether they achieve higher performance than the benchmarks. A flat $1/N$ rule would mechanically allocate more capital to exposures represented by a larger number of ETFs, which is problematic because the ETF universe re-

mains highly unbalanced (see Subsection 2.1.3). The policy-weighted and hierarchical equal-weight benchmarks are therefore constructed in Subsection 2.1.8 to reduce this ticker-count bias and to test whether statistical optimization improves on transparent allocation rules defined across economic exposures rather than individual fund symbols.

Hypothesis 2 is structured around paired comparisons in order to isolate the incremental contribution of macroeconomic and financial information. The historical-view and macro-view Black–Litterman strategies share the same prior, covariance estimator, constraints, and portfolio-construction procedure, while differing only in the information used to formulate views. Similarly, unconditional and regime-conditioned sleeve-level MVO use the same sleeve structure and optimization framework, with the latter allowing expected returns and covariances to vary with the estimated macro-financial regime. These paired designs provide the most controlled tests of whether the additional information improves portfolio decisions. The macro-informed strategies are also compared with the H1 models and common benchmarks, but this broader comparison serves a different purpose: it assesses whether the added modelling complexity produces strategies that are competitive within the complete empirical framework.

Hypothesis 3 provides a practical stress test of the complete strategy set developed under Hypotheses 1 and 2. It introduces monthly contributions and proportional trading costs and compares each realized strategy outcome with a simulated distribution of Random DCA paths. This distribution represents probable investment paths generated under the realized market history, growing eligible universe, and fixed random-selection and review rules. It provides a conditional distributional comparison of implemented outcomes within that design and is not a classical sampling distribution over alternative economic histories.

The following Table 2.1 concisely summarizes the models and benchmarks against which performance is evaluated.

Table 2.1: Empirical comparisons used to evaluate the hypotheses

<i>Hypothesis Core empirical comparison</i>	
H1	Policy-weighted and hierarchical equal-weight benchmarks compared with GMV, equal risk contribution, maximum diversification, and ETF-level MVO
H2	Historical-view Black–Litterman compared with macro-view Black–Litterman; unconditional sleeve-level MVO compared with regime-conditioned sleeve-level MVO
H3	Each pre-specified strategy evaluated with monthly contributions and transaction costs; outcomes additionally located within the conditional Random DCA distribution

Source: Author’s empirical comparison design.

2.0.1 Hypothesis Assessment Framework

Each hypothesis is evaluated using an explicit assessment rule. The primary metric determines whether a comparison supports a hypothesis, while the remaining reported measures are used to interpret the associated economic trade-offs. Return, volatility, maximum drawdown, terminal wealth, turnover, and concentration therefore remain substantively important, but they do not override the primary assessment metric.

For H1, the primary metric is the out-of-sample Sharpe ratio. The hypothesis is decomposed into four sub-comparisons: GMV versus the two principal benchmarks, ERC versus the two principal benchmarks, MD versus the two principal benchmarks, and ETF-level MVO versus the two principal benchmarks. A sub-comparison is counted as supported only if the strategy’s Sharpe ratio exceeds those of both Policy and HEW on the same evaluation sample. H1 receives full support if all four sub-comparisons are supported, partial support if at least one but not all are supported, and no support if none are supported.

For H2, the primary metric is again the out-of-sample Sharpe ratio. The hypothesis is decomposed into two matched sub-comparisons: BL-M versus BL-H and Sleeve-MVO-RC versus Sleeve-MVO-U. A sub-comparison is counted as supported only if the macro-conditioned strategy records the higher baseline Sharpe ratio on the same evaluation sample. H2 receives full descriptive support if both sub-comparisons are supported, partial support if exactly one is supported, and no support if neither is supported. Robustness results qualify the economic interpretation but do not alter this baseline classification.

For H3, only strategy comparisons that are favourable in the common-period

frictionless evaluation are assessed. A favourable comparison survives the retail implementation if the previously leading strategy retains both a higher net time-weighted return and higher terminal net wealth after contributions and transaction costs. For the H1 comparisons, the optimized strategy must remain above both Policy and HEW on both measures; for the H2 comparisons, the macro-conditioned strategy must remain above its matched control. H3 receives full support if all favourable frictionless sub-comparisons survive, partial support if at least one but not all survive, and no support if none survive. The Random DCA distribution is supplementary and does not enter this support classification.

The chapter proceeds accordingly. Section 2.1 defines the data, investable universe, information boundary, constraints, common estimators, and benchmarks. Section 2.2 introduces the unconditional allocation models. Section 2.3 develops the shared macro-regime engine and the two macro-informed comparisons. Section 2.4 specifies the contribution-based implementation and its distributional benchmark. The common backtesting sequence, evaluation measures, robustness specifications, and numerical diagnostics are subsequently consolidated in Section 2.5. Technical derivations and implementation details that are not required to follow the main empirical argument are reported in the appendices.

2.1 Common Empirical Framework

The common empirical framework fixes the features that should not vary across strategies: data construction, ETF eligibility, the monthly information boundary, portfolio constraints, and the baseline estimation methodology. This is necessary to ensure that a strategy does not appear superior merely because it receives earlier access to ETFs, uses later information, or faces less restrictive constraints. All strategies therefore use the same growing ETF universe, long-only and fully invested feasible set, 20% cap per ETF, and 60-month Quadratic-Inverse Shrinkage covariance estimator. Expected returns are estimated over 24 months in the baseline specification and over 12 and 48 months in robustness tests, while the covariance window remains fixed at 60 months so that the responsiveness of expected returns can be varied without simultaneously changing the risk model. Hypotheses 1 and 2 are evaluated in a constrained but frictionless setting in which every strategy begins with one unit of wealth, receives no subsequent contributions, and incurs no transaction costs. Hypoth-

esis 3 retains the same target-weight rules and portfolio constraints but changes the wealth-accounting environment by introducing monthly contributions and proportional trading costs. Long-only positions, full investment, no leverage, and the ETF weight cap therefore remain characteristics of the underlying investment problem rather than implementation frictions.

2.1.1 ETF Prices, Currency Conversion, and Monthly Returns

The common data block contains ETF prices, foreign-exchange rates, and the risk-free proxy required throughout the empirical analysis. The macroeconomic and financial series used only by H2 are introduced later, after the unconditional strategies have been defined.

Daily ETF prices are obtained through the Yahoo Finance API. The empirical universe combines U.S.-listed ETFs with European-listed UCITS ETFs. The analysis therefore adopts the perspective of a hypothetical international retail investor who can access the selected listings, rather than the perspective of an investor subject to one jurisdiction-specific product-distribution regime. Investor-specific taxes, brokerage permissions, and account restrictions remain outside the scope of the study.

Adjusted close prices are used as total-return proxies because they incorporate cash distributions and capital actions such as stock splits. Fund operating expenses are not deducted again because they are already reflected in fund net asset values and market performance. The resulting returns are pre-tax and exclude bid-ask spreads, market impact, and investor-specific foreign-exchange charges. The retail experiment later adds only the common proportional trading cost specified in Section 2.4.

All prices are converted to U.S. dollars before monthly returns are calculated. Of the 71 ETFs, 48 are reported directly in U.S. dollars, 18 German-listed ETFs are quoted in euros, and five London-listed ETFs are quoted in pounds sterling or pence. The foreign-exchange series are the daily Yahoo Finance close rates $\text{EURUSD}=\mathbf{X}$ and $\text{GBPUSD}=\mathbf{X}$. Let $P_{i,d}^{LC}$ denote the adjusted close of ETF i in its listing currency on day d , and let $FX_{LC/USD,d}$ denote U.S. dollars per unit of that currency. The converted price is

$$P_{i,d}^{USD} = P_{i,d}^{LC} FX_{LC/USD,d}. \quad (2.1)$$

For London listings quoted in pence rather than pounds, the local-currency price is divided by 100 before applying (2.1).¹ Exchange and foreign-exchange markets do not always share the same holiday calendar. When an exchange-rate observation is unavailable on an ETF trading day, the most recent earlier rate is carried forward. A later exchange rate is never used.

The monthly USD price $P_{i,t}^{USD}$ is the last available adjusted USD close in calendar month t . The corresponding simple monthly return is

$$R_{i,t} = \frac{P_{i,t}^{USD}}{P_{i,t-1}^{USD}} - 1. \quad (2.2)$$

A return is retained only when both consecutive month-end prices are available. The processed monthly ETF histories contain no internal missing observations after the first valid return. Consequently, once an ETF has accumulated 60 monthly returns, it also has a complete trailing 60-month matrix for covariance estimation. All portfolio returns, wealth values, and contributions are expressed in U.S. dollars. Currency movements therefore remain part of the return on unhedged non-U.S.-dollar listings.

2.1.2 Risk-Free Return

The monthly risk-free proxy is constructed from the FRED daily 3-month Treasury constant maturity yield, series DGS3M0. A synthetic return is preferred to a cash ETF because it provides a longer history and avoids introducing another fund-specific inception date. Let y_{t-1} denote the annualized percentage yield last observed by the end of month $t - 1$. The approximate one-month return used during month t is

$$r_{f,t} = \left(1 + \frac{y_{t-1}}{100}\right)^{1/12} - 1. \quad (2.3)$$

This transformation is not the exact holding-period return on a continuously rolled three-month Treasury bill. It is a consistent monthly approximation used to calculate excess returns, Sharpe ratios, and Black–Litterman views. The synthetic series is an evaluation and modelling input; it is not an investable asset in the optimization problems. Its construction is summarized in Section C.1.

¹The conversion is therefore $P_{i,d}^{USD} = (P_{i,d}^{GBP}/100)FX_{GBP/USD,d}$ for a pence-denominated quotation.

2.1.3 ETF Universe Selection

The empirical universe contains 71 ETFs representing broad and country-level equity markets, sector and factor exposures, short- and long-duration U.S. Treasuries, commodities, and listed real estate. ETF selection follows four principles. First, each fund must represent an identifiable economic exposure. Second, adjusted historical prices must be available under a sufficiently stable listing symbol. Third, the overall set must contain both growth-oriented and defensive or real-asset exposures. Fourth, the number of funds must remain small enough for the resulting allocations to be interpreted and replicated. The complete list, economic classification, and first valid monthly return are reported in Section C.3.

The final selection is informed by the reviewed literature and the research gap identified earlier. Bessler *et al.* (2021) use U.S. sector and factor ETFs, Petronio *et al.* (2014) examine ETFs covering U.S. and European markets, and Huang & Lin (2011) construct optimized portfolios on a global ETF universe. The equity opportunity set is complemented by short- and long-duration U.S. Treasuries, commodities and precious metals, and listed real estate. Treasury maturities provide different exposures to interest-rate risk (Campbell & Viceira 2001), commodities can provide diversification relative to stocks and bonds (Gorton & Rouwenhorst 2006), and real estate represents an additional diversifying asset class in mixed-asset portfolios (Hoesli *et al.* 2004). The selected universe is intended to provide a broad and interpretable set of commonly available exchange-traded investment opportunities.

The universe is constructed top-down. The required economic exposures are defined first from the research question and the reviewed literature, after which an ETF is selected to represent each exposure. Where several funds offer substantially similar exposure, preference is given to the fund with the clearer exposure definition, longer and more stable adjusted-price history, and greater consistency with the ETF family used within the corresponding sleeve. Realised returns, Sharpe ratios, and subsequent backtest performance are not used to select among alternative funds. The resulting universe is therefore purposively constructed rather than generated by a screening algorithm.

The use of ETFs reduces the security-level constituent-reconstruction problem but does not eliminate selection bias. The final list is constructed from identifiable funds available through the selected data source and does not fully reconstruct ETFs liquidated before the universe-selection date. The results may

therefore retain fund-level survivorship bias. This limitation affects the interpretation of long-run performance but does not justify return-based screening within the surviving sample.

No return- or correlation-based screening rule is applied either before or after the initial economic selection. High correlations between ETFs are expected because regional, sector, and factor ETFs share broad equity exposure, but the funds represent different economic allocation decisions. Removing ETFs according to trailing correlations or Sharpe ratios would introduce an additional data-dependent selection rule before the portfolio models are evaluated. The consequences of overlap among the retained funds are instead managed through the economic sleeve structure, the individual ETF weight cap, and the regularized covariance estimator.

2.1.4 Economic Sleeve Classification

Every ETF is assigned to one top-level asset class and one permanent economic sleeve. The hierarchy is

$$\text{asset class} \longrightarrow \text{economic sleeve} \longrightarrow \text{individual ETF}.$$

The four top-level classes are equities, Treasuries, commodities, and listed real estate. Within these classes, the 71 ETFs are organized into 15 sleeves. The regional equity block is divided into four sleeves rather than one because U.S. broad equity, broad developed ex-U.S. exposure, developed-country funds, and emerging-market or regional funds represent distinct geographic allocation decisions. This separation also prevents a single 16-ETF regional sleeve from receiving the same benchmark weight as one narrower sector or factor sleeve.

The classification follows four empirical roles. It removes ticker-count bias from the hierarchical benchmarks, defines the portfolios on which Black–Litterman views are expressed, reduces the dimension of the regime-conditioned allocation problem, and preserves an economic interpretation as new ETFs become eligible. The classification is based on investment exposure rather than historical performance. EXH8.DE is assigned to listed real estate rather than to European sectors; SHY and TLT form separate duration sleeves; and GDX is grouped with precious-metal exposure because of its economic sensitivity to gold, although it holds mining equities rather than physical metal.

The following Table 2.2 summarizes the permanent economic sleeves. The table reports the number of ETFs in each sleeve, the corresponding economic

exposure, and the total number of ETFs in each top-level asset class. The 71-ETF universe contains 60 equity, two Treasury, five commodity, and four listed-real-estate funds. The sleeve classification is used throughout the empirical analysis to define benchmarks, Black–Litterman views, and regime-conditioned moments. The complete ticker-level mapping is reported in Section C.4, while ETF eligibility and the resulting active-sleeve set are defined in the following subsection.

Table 2.2: Permanent economic sleeves used in portfolio construction

<i>No.</i>	<i>Class</i>	<i>Sleeve</i>	<i>N</i>	<i>Content</i>
1	Equity	U.S. broad equity	1	Broad U.S. market exposure through SPY
2	Equity	Broad developed ex-U.S. equity	2	Broad developed-market and European regional exposure through VEA and VGK
3	Equity	Developed-country equity	7	Australia, Germany, Italy, Japan, Spain, France, and the United Kingdom
4	Equity	Emerging-market and regional equity	6	Broad emerging markets, Taiwan, South Korea, Latin America, India, and China
5	Equity	U.S. sectors	10	Main U.S. sector ETFs, with VOX as the longer-history communication-services proxy
6	Equity	European sectors	17	STOXX Europe 600 sector ETFs, excluding EXH8.DE
7	Equity	Asia/China sectors	3	China consumer discretionary, technology, and internet exposures
8	Equity	U.S. factors	6	Momentum, quality, value, minimum volatility, size, and dividend exposures
9	Equity	European factors	4	Momentum, quality, value, and minimum-volatility exposures
10	Equity	International factors	4	Developed ex-U.S. momentum, quality, value, and minimum-volatility exposures

Continued on next page

Table 2.2 – continued

<i>No.</i>	<i>Class</i>	<i>Sleeve</i>	<i>N</i>	<i>Content</i>
11	Treasuries	Short-duration Treasuries	1	SHY
12	Treasuries	Long-duration Treasuries	1	TLT
13	Commodities	Precious metals and gold-linked exposure	3	Gold, silver, and gold-mining equity exposure through GLD, SLV, and GDX
14	Commodities	Broad and industrial commodities	2	Broad commodity and base-metal exposure through DBC and DBB
15	Real estate	Listed real estate	4	U.S., international, and European listed-real-estate exposures
Total			71	60 equity, 2 Treasury, 5 commodity, and 4 real-estate ETFs

Source: Author's classification based on the processed ETF universe.

2.1.5 Growing Eligible Universe

The 71 selected ETFs record their first valid monthly returns at different dates, as summarized in Table 2.3. A fixed complete panel would discard the earlier history of established funds and would implicitly give past investors access only to products that survived long enough to enter the final sample. The empirical design therefore uses a growing universe: a fund may enter only after it has accumulated the required history, while earlier portfolio decisions remain unchanged. This construction better represents the expansion of the ETF opportunity set and prevents later fund introductions from affecting earlier eligibility, estimation, or portfolio weights.

ETF i becomes eligible for the portfolio held during month t when it has a complete set of 60 monthly USD returns available before formation:

$$\mathcal{U}_t = \{i : R_{i,t-\ell} \text{ is observed for every } \ell = 1, \dots, 60\}. \quad (2.4)$$

Because the processed ETF histories are continuous after their first valid return, this rule is equivalent in the present data to requiring at least 60 pre-

viously observed monthly returns. The consecutive-window formulation makes the connection to the covariance estimator explicit: every ETF in $\mathcal{U}_t$ can be placed in the same balanced $60 \times N_t$ return matrix.

Given the eligible universe, let $\mathcal{U}_{g,t} \subseteq \mathcal{U}_t$ denote the ETFs in sleeve g that are eligible for the portfolio held during month t . The number of eligible constituents is

$$N_{g,t} = |\mathcal{U}_{g,t}|. \quad (2.5)$$

A sleeve is active when it contains at least one eligible ETF. The active-sleeve set is therefore

$$\mathcal{G}_t = \{g : N_{g,t} > 0\}. \quad (2.6)$$

The economic definition of each sleeve remains fixed. Only the number of eligible constituents and, during the early sample, the set of active sleeves change. A newly eligible ETF therefore expands an existing economic sleeve without rewriting its historical definition.

The 60-month eligibility rule is applied to optimized portfolios and benchmarks in baseline scenario alike. A simple benchmark does not require five years of estimation data, but allowing it to invest in a new ETF immediately while requiring an optimized strategy to wait would create different opportunity sets. The common rule ensures that performance differences arise from allocation rather than availability. A fund may therefore appear in the descriptive dataset before it can influence weights, sleeve returns, expected-return estimates, or covariance matrices.

Table 2.3: Growth of the ETF dataset and 60-month eligible universe

<i>End of year</i>	<i>Observed</i>	<i>Eligible</i>	<i>Main development</i>
1999	17	1	SPY, early country funds, and the first U.S. sector ETFs are observed, but most have not accumulated five years of returns.
2004	26	17	Regional and sector exposures form the first broad eligible equity set.
2009	51	26	Treasury, real-estate, precious-metal, and additional regional exposures begin to satisfy the history requirement.
2014	65	52	European sectors and several commodity and factor sleeves enter the eligible universe.
2015	71	54	All final funds are observed, although recently launched factor ETFs remain ineligible.
2019	71	65	Most factor and China-sector funds have accumulated the required history.
2020	71	71	All core ETFs satisfy the 60-month eligibility requirement.

Source: Author's calculations from the first valid monthly USD return of each ETF.

The universe grows but does not contract in the processed sample because the included return series contain no internal gaps after inception. When a new ETF becomes eligible, the covariance matrix, sleeve composition, and all target weights are recalculated at the next formation date. Earlier portfolio decisions remain unchanged.

2.1.6 Information Set and Portfolio Constraints

The information boundary is common to every strategy. Let s index a strategy and let t denote the month in which its portfolio is held. Target weights are selected at the end of month $t - 1$ according to

$$w_t^s = f_s(\mathcal{I}_{t-1}, \mathcal{U}_t; \theta_s), \quad (2.7)$$

where $\mathcal{I}_{t-1}$ contains only information available by the formation date, $\mathcal{U}_t$ is the eligible universe, and θ_s contains parameters fixed before performance is evaluated. The information set includes ETF returns, exchange rates, risk-free

rates, and, for H2 strategies, macroeconomic and financial observations released by the end of month $t - 1$. It excludes the return realized during month t and all later data.

Portfolio decisions are therefore formed recursively using only information available at each formation date, yielding a strict recursive historical out-of-sample evaluation. Throughout the thesis, “out-of-sample” refers to recursively generated one-month-ahead historical returns and should not be interpreted as performance from a prospectively untouched holdout.

This thesis is out-of-sample in the recursive time-series sense because each month- t portfolio is formed with information through $t - 1$, and month- t returns are observed afterward. Selected strategies, constraints, estimation choices, and macroeconomic variables are motivated by the literature rather than selected according to their realised performance. However, the ETF universe and its classification into sleeves is constructed retrospectively while the historical evaluation period was observable. The results are therefore interpreted as recursive historical out-of-sample evidence rather than evidence from a prospectively pre-registered holdout.

The realized gross portfolio return is

$$R_{p,t}^s = (w_t^s)^\top R_{\mathcal{U},t}, \quad (2.8)$$

where $R_{\mathcal{U},t}$ is the vector of month- t returns. Formation and realization are therefore separated by construction. The same timing rule governs eligibility, parameter estimation, Black–Litterman views, regime probabilities, and transaction-cost calculations.

All ETF-level models use the feasible set

$$\mathcal{W}_t = \left\{ w_t : \mathbf{1}^\top w_t = 1, \quad 0 \leq w_{i,t} \leq 0.20 \text{ for every } i \in \mathcal{U}_t \right\}. \quad (2.9)$$

The long-only and no-leverage restrictions match the retail focus of the thesis. The 20% upper bound is an economically motivated concentration constraint rather than an estimated optimum. It limits dominance by a single ETF, requires at least five holdings in a fully invested portfolio, and regularizes unstable optimized solutions (Jagannathan & Ma 2003). The same cap is applied to every core strategy and benchmark without imposing sleeve-specific limits.

Cash is not included as a strategic asset. An unrestricted near-zero-volatility asset would dominate a pure minimum-variance objective and would change the

economic meaning of the risk-contribution and diversification-ratio strategies. The synthetic Treasury-bill return is therefore used as an excess-return reference rather than as another security.

2.1.7 Common Return and Covariance Estimation

The strategy set does not use one identical collection of inputs. The policy and hierarchical benchmarks require only eligibility and sleeve membership. GMV, equal risk contribution, and maximum diversification use the covariance matrix but not expected returns. MVO uses both. Black–Litterman adds posterior expected returns, and the regime-conditioned strategy adds conditional sleeve moments. To keep the comparisons interpretable, common inputs are estimated in the same way whenever a model requires them.

Expected returns are estimated from a 24-month rolling arithmetic mean in the baseline specification. This baseline design balances responsiveness to changing return conditions against the instability of very short sample means; the 12- and 48-month robustness specifications bracket it with materially shorter and longer alternatives. Covariance estimation instead uses 60 months to provide a more stable five-year risk history while preserving a meaningful evaluation period for the growing ETF universe. A substantially longer requirement would delay the eligibility of newer ETFs, while QIS regularizes the covariance estimate as the cross-section grows. None of these windows is selected according to realized backtest performance.

Let $W_\mu \in \{12, 24, 48\}$ denote the expected-return window, with $W_\mu = 24$ in the baseline. For ETF $i \in \mathcal{U}_t$,

$$\hat{\mu}_{i,t}^{(W_\mu)} = \frac{1}{W_\mu} \sum_{\ell=1}^{W_\mu} R_{i,t-\ell}. \quad (2.10)$$

The estimates remain in monthly units. They are not annualized before optimization because the covariance matrix and risk-aversion parameters are also specified at monthly frequency.

Let $R_{t-\ell}$ be the $N_t \times 1$ vector of eligible ETF returns in month $t - \ell$. The 60-month sample mean used only to centre the covariance matrix is

$$\bar{R}_t^{(60)} = \frac{1}{60} \sum_{\ell=1}^{60} R_{t-\ell}, \quad (2.11)$$

and the sample covariance is

$$S_t = \frac{1}{59} \sum_{\ell=1}^{60} \left(R_{t-\ell} - \bar{R}_t^{(60)} \right) \left(R_{t-\ell} - \bar{R}_t^{(60)} \right)^\top. \quad (2.12)$$

A covariance matrix for N_t assets contains $N_t(N_t + 1)/2$ distinct elements. With all 71 ETFs eligible, 2,556 quantities are estimated from only 60 monthly return vectors. Demeaning also limits the rank of S_t to 59, so the matrix is singular whenever $N_t > 59$ and may be poorly conditioned before that point. Optimization can then mistake an underestimated eigenvalue for a low-risk allocation direction.

The baseline risk input is the Quadratic-Inverse Shrinkage estimator of Ledoit & Wolf (2022b). If $S_t = U_t \Lambda_t U_t^\top$, QIS retains the sample eigenvectors and replaces the sample eigenvalues with nonlinear, spectrum-dependent estimates:

$$\hat{\Sigma}_t^{\text{QIS}} = U_t \tilde{\Lambda}_t U_t^\top. \quad (2.13)$$

The shrinkage adjustment depends on the full sample spectrum and on the ratio of the number of assets to the effective sample size. In the singular case, QIS assigns positive variance to directions associated with zero sample eigenvalues and therefore produces an invertible matrix without requiring a researcher-chosen constant-correlation or factor target.

QIS is used because the ETF dimension eventually exceeds the effective number of observations, the universe combines heterogeneous exposures for which one external target would impose strong economic assumptions, and all optimized strategies require a stable risk matrix. The estimator is not assumed to guarantee superior performance. Its narrower purpose is to prevent the empirical comparison from being dominated by a singular or severely ill-conditioned sample covariance matrix. The full nonsingular and singular formulas, trace rescaling, and implementation checks are provided in Appendix D.

For H1 and ETF-level Black–Litterman, QIS is applied to the balanced $60 \times N_t$ ETF return matrix. For sleeve-level H2 models, the same procedure is applied to the balanced $60 \times |\mathcal{G}_t|$ sleeve return matrix. Keeping one covariance methodology at both levels reduces the number of dimensions along which the models differ.

2.1.8 Common Benchmarks

The literature reviewed in Chapter 1 shows that an optimized portfolio is recommended to be tested against some simple allocation rule, to evaluate the added value of optimizing portfolio allocation. A useful benchmark should be transparent, investable on the same universe, and sufficiently different in construction to reveal how much optimization contributes. Two hierarchical benchmarks satisfy these requirements. The policy-weighted portfolio represents a strategic multi-asset allocation with an explicit equity preference. The hierarchical equal-weight portfolio removes that preference and allocates equally across economic decisions. Both avoid giving an exposure more capital simply because it is represented by more ETF tickers.

The policy benchmark allocates capital in three stages: predetermined weights are assigned to the four asset classes, each class allocation is divided equally among its active sleeves, and each sleeve allocation is divided equally among its eligible ETFs. The hierarchical equal-weight benchmark omits the first stage and begins by assigning equal weights to all active sleeves.

Let the top-level classes be equities (E), Treasuries (B), commodities (C), and listed real estate (RE). The class weights define a stylized strategic benchmark rather than an estimated optimal allocation. They are fixed before the out-of-sample comparison and are not selected according to realized performance. The 60% equity allocation gives the benchmark a conventional growth-oriented profile, while the remaining 40% provides material exposure to Treasuries, commodities, and listed real estate. The policy portfolio is not treated as a universally optimal allocation; it provides a stable economically structured reference point. The hierarchical equal-weight benchmark is reported alongside it so that conclusions do not depend solely on these predetermined class weights.

Let $\mathcal{G}_{a,t} \subseteq \mathcal{G}_t$ denote the active sleeves in class a . The policy class weights are fixed before the backtest:

$$q_E = 0.60, \quad q_B = 0.20, \quad q_C = 0.15, \quad q_{RE} = 0.05. \quad (2.14)$$

Within a class, each active sleeve receives an equal share:

$$x_{g,t}^{\text{Policy}} = \frac{q_a}{|\mathcal{G}_{a,t}|}, \quad g \in \mathcal{G}_{a,t}. \quad (2.15)$$

The sleeve allocation is then divided equally among its eligible ETFs:

$$w_{i,t}^{\text{Policy}} = \frac{x_{g,t}^{\text{Policy}}}{N_{g,t}}, \quad i \in \mathcal{U}_{g,t}. \quad (2.16)$$

The class weights and allocation rule are fixed, but the resulting ETF weights are not numerically constant while the universe grows. New active sleeves change the equal division within their class, and new ETFs change the equal division within their sleeve. Once all 15 sleeves are active, the construction is especially transparent, as shown in Table 2.4.

Table 2.4: Policy allocation when all 15 sleeves are active

<i>Top-level class</i>	<i>Class weight</i>	<i>Active sleeves</i>	<i>Weight per sleeve</i>
Equities	60%	10	6.0%
Treasuries	20%	2	10.0%
Commodities	15%	2	7.5%
Listed real estate	5%	1	5.0%

Source: Author's illustration from (2.14)–(2.16).

For example, the U.S. broad-equity sleeve receives 6%, all of which is invested in SPY. The developed-country sleeve also receives 6%, but that amount is divided equally among its seven eligible country ETFs when all are active. Each Treasury sleeve receives 10%; the precious-metals sleeve and broad-commodity sleeve each receive 7.5%; and the real-estate sleeve receives 5%. The regional equity block comprises four sleeves and therefore receives 24% in total, while sector and factor decisions each receive 18%. This is the intended geographic balance within the 60% equity allocation.

The policy portfolio also serves as the reference allocation in Black–Litterman reverse optimization. This role does not imply that it is a market-capitalization equilibrium portfolio. Rather, it provides a stable neutral allocation from which the Black–Litterman models depart only when their sleeve views are sufficiently strong relative to estimated uncertainty.

The hierarchical equal-weight benchmark removes the policy class preferences. Every active sleeve receives the same weight:

$$x_{g,t}^{\text{HEW}} = \frac{1}{|\mathcal{G}_t|}, \quad g \in \mathcal{G}_t, \quad (2.17)$$

and every eligible ETF within sleeve g receives

$$w_{i,t}^{\text{HEW}} = \frac{x_{g,t}^{\text{HEW}}}{N_{g,t}}, \quad i \in \mathcal{U}_{g,t}. \quad (2.18)$$

When all 15 sleeves are active, each receives $1/15 = 6.67\%$. SPY therefore receives 6.67% through its single-ETF sleeve, whereas each of the 17 European sector ETFs receives approximately $6.67\%/17 = 0.39\%$. The benchmark is equal across economic decisions, not across ticker symbols. It is therefore the principal same-universe comparator for whether statistical estimation improves on a transparent neutral allocation.

A flat ETF-level equal-weight portfolio is reported as a robustness benchmark:

$$w_{i,t}^{1/N} = \frac{1}{N_t}, \quad i \in \mathcal{U}_t. \quad (2.19)$$

The flat rule is easy to reproduce and remains a demanding empirical comparator (DeMiguel *et al.* 2009). However, it is equal only across ticker symbols, not across economic exposures. When all 71 ETFs are eligible, each fund receives approximately $1/71 = 1.41\%$. The 17 ETFs in the European sector sleeve therefore receive approximately $17/71 = 23.94\%$ in total, whereas the single-ETF U.S. broad-equity sleeve receives only $1/71 = 1.41\%$. The difference reflects the number of selected funds rather than an economic allocation decision. For this reason, flat equal weighting is reported as a robustness benchmark, while hierarchical equal weighting remains the principal neutral same-universe benchmark.

The Vanguard Total World Stock ETF (VT) is reported separately as an external global-equity market reference. It is a single all-equity fund rather than a strategy formed from the 71-ETF multi-asset opportunity set. Consequently, VT is not subject to the 20% ETF cap or the 60-month growing-universe rule and does not enter the formal assessment of Hypotheses 1–3. Chapter 3 reports VT over the standalone H1 and common evaluation periods as contextual evidence only.

2.2 Hypothesis 1: Unconditional Portfolio Allocation

Hypothesis 1 tests whether unconditional portfolio optimization improves out-of-sample performance relative to the policy-weighted and hierarchical equal-

weight benchmarks defined in Subsection 2.1.8. Flat ETF-level equal weighting is retained as a secondary robustness benchmark, not as a replacement for the two economically structured comparisons. No macro-financial conditioning, investor contribution, or transaction cost enters H1. The comparison therefore isolates the allocation rules before the retail implementation considered under Hypothesis 3.

For each holding month t , all four optimized strategies allocate directly across the common growing universe $\mathcal{U}_t$, and w_t denotes the corresponding vector of ETF weights. The weights are formed using information available through month $t - 1$ and must belong to the common feasible set $\mathcal{W}_t$ in (2.9). Thus, every strategy is long-only and fully invested, with each ETF weight bounded above by 20%. Each strategy also uses the QIS covariance estimate $\hat{\Sigma}_t^{\text{QIS}}$ in (2.13), calculated from the same balanced 60-month return window. Holding the universe, timing convention, covariance estimator, and constraints fixed makes the objective function the defining difference among the four H1 strategies.

GMV, equal risk contribution, and maximum diversification are covariance-only strategies. MVO uses the same covariance input but additionally uses a trailing arithmetic-mean return estimate. This distinction permits H1 to separate alternative uses of estimated risk from the additional estimation uncertainty introduced by expected returns. The selected models correspond to established covariance-based and mean-variance approaches in the portfolio-allocation literature (Markowitz 1952; Maillard *et al.* 2010; Choueifat & Coignard 2008).

2.2.1 Covariance-Based Allocation

The global minimum-variance strategy selects the feasible ETF allocation with the lowest QIS-estimated portfolio variance:

$$w_t^{\text{GMV}} = \arg \min_{w_t \in \mathcal{W}_t} w_t^\top \hat{\Sigma}_t^{\text{QIS}} w_t. \quad (2.20)$$

GMV depends only on the estimated covariance matrix and the common constraints; it contains no expected-return input.

Equal risk contribution uses the same covariance matrix but allocates according to normalized contributions to portfolio risk. Let portfolio volatility be

$$\sigma_p(w_t) = \sqrt{w_t^\top \hat{\Sigma}_t^{\text{QIS}} w_t}, \quad (2.21)$$

and define ETF i 's contribution to portfolio volatility as

$$RC_{i,t}(w_t) = \frac{w_{i,t} \left(\hat{\Sigma}_t^{\text{QIS}} w_t \right)_i}{\sigma_p(w_t)}. \quad (2.22)$$

The normalized contribution share is

$$NRC_{i,t}(w_t) = \frac{RC_{i,t}(w_t)}{\sigma_p(w_t)} = \frac{w_{i,t} \left(\hat{\Sigma}_t^{\text{QIS}} w_t \right)_i}{w_t^\top \hat{\Sigma}_t^{\text{QIS}} w_t}. \quad (2.23)$$

The implemented ERC strategy minimizes the squared dispersion of these shares around the equal target $1/N_t$:

$$w_t^{\text{ERC}} = \arg \min_{w_t \in \mathcal{W}_t} \sum_{i=1}^{N_t} \left[NRC_{i,t}(w_t) - \frac{1}{N_t} \right]^2. \quad (2.24)$$

The 20% cap may prevent exact equality, in which case (2.24) selects the feasible allocation with the smallest normalized contribution dispersion. This specification follows the equal-risk-contribution principle formalized by Mailard *et al.* (2010).

For maximum diversification, let $\hat{\sigma}_t$ contain the individual ETF volatilities implied by the diagonal of $\hat{\Sigma}_t^{\text{QIS}}$. The diversification ratio is

$$DR(w_t) = \frac{w_t^\top \hat{\sigma}_t}{\sqrt{w_t^\top \hat{\Sigma}_t^{\text{QIS}} w_t}}. \quad (2.25)$$

The maximum-diversification portfolio solves

$$w_t^{\text{MD}} = \arg \max_{w_t \in \mathcal{W}_t} DR(w_t). \quad (2.26)$$

The strategy uses no expected-return estimate and implements the criterion proposed by Choueifaty & Coignard (2008).

The three covariance-only strategies therefore impose distinct empirical objectives on the same risk estimate: GMV minimizes portfolio variance, ERC minimizes unequal normalized risk contributions, and maximum diversification maximizes the diversification ratio.

2.2.2 Mean–Variance Allocation

Mean–variance optimization adds expected returns to the common QIS risk model. For an expected-return window W_μ , the estimate $\hat{\mu}_t^{(W_\mu)}$ is the trailing arithmetic mean defined in (2.10). The portfolio solves

$$w_t^{\text{MVO}} = \arg \max_{w_t \in \mathcal{W}_t} \left\{ w_t^\top \hat{\mu}_t^{(24)} - \frac{\gamma}{2} w_t^\top \hat{\Sigma}_t^{\text{QIS}} w_t \right\}, \quad \gamma = 5. \quad (2.27)$$

The baseline fixes $W_\mu = 24$ months and $\gamma = 5$ before performance is evaluated. The coefficient $\gamma = 5$ represents an economic calibration for a moderately risk-averse investor under monthly return and covariance inputs; robustness for γ is out of scope of current research. Robustness specifications replace the baseline mean with $\hat{\mu}_t^{(12)}$ or $\hat{\mu}_t^{(48)}$ while retaining the same 60-month QIS covariance, eligible universe, feasible set, risk-aversion coefficient, and monthly timing convention. MVO is therefore the only H1 strategy whose weights respond directly to estimated mean returns; the other three respond only to the covariance structure.

The numerical implementation and code-level validation of all four H1 models are documented in Section E.2.

2.3 Hypothesis 2: Added Value of Macroeconomic and Financial Information

Hypothesis 2 asks whether additional information improves allocation after the investment problem itself has been held constant. An unrestricted comparison between unrelated macro strategies would not answer that question because differences could arise from the optimizer, benchmark, universe, or constraints. The H2 design therefore uses two matched pairs. Historical-view and macro-view Black–Litterman share the same reference portfolio, prior, covariance matrix, view portfolios, confidence methodology, feasible set, and final optimizer. Unconditional and regime-conditioned sleeve MVO share the same 15 sleeves, risk-aversion coefficient, ETF cap, and within-sleeve equal weighting. Within each pair, the macro-regime information is the only substantive addition.

Both pairs are supplied by one common regime engine. The engine first converts one direct macroeconomic activity measure and three financial-condition measures into probabilities of favourable, intermediate, and adverse states. Those probabilities are then combined with historical sleeve returns. The re-

sulting regime-conditioned means support both the macro Black–Litterman views and the regime-conditioned MVO, while the regime-conditioned covariances enter only the latter. Using one information architecture avoids selecting a different macro signal for each portfolio model.

2.3.1 Macroeconomic and Financial Information Set

The baseline regime vector contains four variables:

1. A_t^{RT} : the most recent real-time value of the Chicago Fed National Activity Index three-month moving average, FRED series `CFNAIMA3`, available by the final trading date of month t ;
2. $\overline{\log(VIX)}_t$: the monthly mean of the logarithm of daily CBOE Volatility Index observations, FRED series `VIXCLS`;
3. $\overline{TS}_t$: the monthly mean of the daily 10-year minus 3-month U.S. Treasury term spread, FRED series `T10Y3M`;
4. $\overline{\log(CS)}_t$: the monthly mean of the logarithm of the daily Moody’s Baa corporate yield spread over the 10-year Treasury yield, FRED series `BAA10Y`.

The variables have distinct economic roles. CFNAI-MA3 is a direct measure of broad U.S. economic activity. It summarizes indicators related to production and income, employment and hours, consumption and housing, and sales, orders, and inventories. A value of zero corresponds to growth near its historical trend, a negative value indicates below-trend activity, and a positive value indicates above-trend activity. Its three-month moving average is used rather than the more volatile monthly index because the objective is to identify persistent conditions rather than one-month macroeconomic noise.

VIX reflects option-implied expectations of near-term U.S. equity-market volatility and therefore captures uncertainty and risk aversion that may not yet appear in realized monthly returns. The term spread summarizes the slope of the Treasury yield curve and contains information about monetary conditions and the business cycle. The Baa–Treasury spread measures the additional yield required on lower-investment-grade corporate debt relative to the 10-year Treasury benchmark. It is therefore used as a long-history proxy for corporate credit stress. It is not described as a high-yield option-adjusted spread.

The daily financial variables are converted to monthly observations using arithmetic averages of all available daily values in each calendar month. Logs are taken before averaging VIX and the credit spread because both series are positive and strongly right-skewed. Let $\mathcal{D}_{j,t}$ denote the available daily observations for variable j in month t . The aggregation is

$$\overline{\log(VIX)}_t = \frac{1}{|\mathcal{D}_{V,t}|} \sum_{d \in \mathcal{D}_{V,t}} \log(VIX_d), \quad (2.28)$$

$$\overline{TS}_t = \frac{1}{|\mathcal{D}_{TS,t}|} \sum_{d \in \mathcal{D}_{TS,t}} TS_d, \quad (2.29)$$

$$\overline{\log(CS)}_t = \frac{1}{|\mathcal{D}_{CS,t}|} \sum_{d \in \mathcal{D}_{CS,t}} \log(CS_d). \quad (2.30)$$

Monthly averaging uses the full information accumulated during the month and reduces sensitivity to one noisy month-end observation. An exponentially weighted average is not used because it would add an extra decay parameter without a hypothesis-specific reason.

CFNAI-MA3 requires different timing treatment because it is released after the reference month and its historical values may be revised. For each month-end, A_t^{RT} is defined as the latest CFNAI-MA3 observation that had actually been published by that date. The reconstruction uses the Chicago Fed's archive of past releases and, for dates covered by it, the vintage-date functionality of ALFRED or the FRED API. Each stored observation records its source release and economic reference month. Today's revised historical series is therefore never shifted backward and treated as information that was already available. The implementation also verifies that an archived real-time observation exists for every HMM month; incomplete archival coverage, if encountered, is reported and addressed only through a pre-specified robustness specification rather than silently filled with revised data. The complete construction is reported in Section C.2.

The resulting raw information vector is

$$F_t = \left[A_t^{RT} \quad \overline{\log(VIX)}_t \quad \overline{TS}_t \quad \overline{\log(CS)}_t \right]^\top. \quad (2.31)$$

The different original data frequencies do not create a mixed-frequency model because every input is converted into one observation per information month before HMM estimation. For the month- t portfolio, the latest vector

included in the estimation sample is F_{t-1} .

2.3.2 Recursive Three-State Hidden Markov Model

The Hidden Markov Model (HMM) adapts the latent-state logic used by Ma *et al.* (2011) to the present long-only, multi-asset setting. It treats the economic state as unobserved. Rather than assigning each month to a regime through a fixed threshold, it estimates the probability that the observed macro-financial vector was generated by each of three latent states. Three states are the smallest specification that directly distinguishes favourable, intermediate, and adverse conditions while remaining parsimonious enough for monthly recursive estimation. This literature-informed choice is consistent with the three economically interpretable regimes used by Ma *et al.* (2011).

At every formation date t , each variable is standardized using only the expanding macro sample available through month $t-1$. For variable j , let $\bar{F}_{j,t-1}$ and $s_{j,t-1}$ be the mean and standard deviation calculated from the common macro starting month through $t-1$. Every observation in the estimation sample is transformed as

$$Z_{j,\tau}^{(t)} = \frac{F_{j,\tau} - \bar{F}_{j,t-1}}{s_{j,t-1}}, \quad \tau \leq t-1. \quad (2.32)$$

The superscript (t) indicates that the standardization is recalculated when the expanding sample grows. This does not introduce future information: the mean and standard deviation used at date t contain no observation after $t-1$. Standardization is required because the four variables are measured in different units and should not influence the likelihood merely because one has a larger numerical scale.

Let $Z_\tau^{(t)}$ collect the four standardized variables and let $r_\tau \in \{1, 2, 3\}$ denote the latent state. Conditional on state k , the observation vector is assumed to follow a Gaussian distribution with a state-specific mean and diagonal covariance matrix:

$$Z_\tau^{(t)} \mid r_\tau = k \sim \mathcal{N}_4(m_{k,t}, V_{k,t}), \quad V_{k,t} = \text{diag}(v_{1k,t}, \dots, v_{4k,t}). \quad (2.33)$$

The diagonal conditional covariance is a parsimonious modelling assumption. A full four-by-four matrix would require ten covariance parameters in each state, whereas the diagonal form requires four. The reduction is important be-

cause each state is supported by only part of a monthly sample. Cross-variable co-movement can still arise because the variables respond jointly to the common latent state, but residual correlations within a state are set to zero. The assumption is therefore a bias–variance trade-off rather than a claim that the four variables are economically independent.

Regime persistence follows a first-order Markov chain with transition matrix

$$A_t = [a_{jk,t}], \quad a_{jk,t} = \Pr(r_\tau = k \mid r_{\tau-1} = j). \quad (2.34)$$

Within one estimation at date t , the transition probabilities are treated as constant across the included historical months. At the next formation date, the sample expands by one observation and the complete matrix is re-estimated. Thus, the parameters are constant within an estimation sample but may evolve recursively across portfolio dates.

The first HMM estimate requires at least 120 common monthly macro observations. This ten-year minimum is a numerical and sample-length compromise: it provides a broad initial history for estimating three state distributions and their transitions, while a longer initialization would materially shorten the recursive evaluation period. Thereafter, the model is re-estimated each month on the expanding sample ending at $t - 1$. Parameters are obtained by the expectation–maximization algorithm. The expectation step computes the probabilities of states and state transitions under the current parameters; the maximization step updates the state means, variances, initial probabilities, and transition matrix to increase the likelihood. Multiple initializations are used because HMM likelihoods can contain local optima, and the converged solution with the highest likelihood is retained. The forward and backward recursions are provided in Section E.5.

The numerical state indices have no inherent economic meaning and may change order between recursive estimates. Economic labels are therefore assigned from state characteristics. Define the stress score

$$L_{k,t} = -m_{k,A,t} + m_{k,VIX,t} + m_{k,CS,t}, \quad (2.35)$$

where $m_{k,A,t}$, $m_{k,VIX,t}$, and $m_{k,CS,t}$ are the state means of standardized activity, log VIX, and log credit spread. High activity lowers the score, while high volatility and credit stress raise it. The state with the lowest score is labelled *favourable*, the state with the highest score is labelled *adverse*, and the remaining state is labelled *intermediate*. The term spread remains part of the

HMM but is excluded from the label score because a high or low slope can have different meanings across monetary-policy and business-cycle phases.

At the end of month $t - 1$, the filtered probability that the current state is j is

$$\alpha_{j,t-1|t-1} = \Pr(r_{t-1} = j \mid Z_1^{(t)}, \dots, Z_{t-1}^{(t)}). \quad (2.36)$$

The probability used for the portfolio held during month t is the one-step-ahead forecast

$$\pi_{k,t|t-1} = \sum_{j=1}^3 \alpha_{j,t-1|t-1} a_{jk,t}. \quad (2.37)$$

The allocation models use the complete probability vector rather than switching to the single most likely state. This allows weights to adjust gradually when state evidence is uncertain.

For historical conditional-moment estimation, the model also computes smoothed probabilities within the information boundary available at date t :

$$\eta_{\tau k}^{(t)} = \Pr(r_{\tau} = k \mid Z_1^{(t)}, \dots, Z_{t-1}^{(t)}), \quad \tau \leq t - 1. \quad (2.38)$$

A smoothed probability for an earlier month uses observations after that historical month, but never uses an observation later than the current formation date. It is therefore an estimate that could have been formed at date t , not a full-sample classification imported into the past.

2.3.3 Sleeve Returns and Regime-Conditioned Moments

The HMM identifies states independently of the ETF cross-section. Its probabilities must therefore be linked to investable returns in a second step. The link is constructed at the sleeve level because state-specific estimation is too data-intensive at the 71-ETF level and because the 15 sleeves correspond to economically interpretable allocation decisions.

At formation date t , the current eligible constituents of sleeve g are held fixed over the historical estimation windows. The return assigned to that sleeve in month $\tau < t$ is

$$R_{g,\tau}^{(t)} = \frac{1}{N_{g,t}} \sum_{i \in \mathcal{U}_{g,t}} R_{i,\tau}. \quad (2.39)$$

In words, the formula reconstructs how an equal-weight portfolio of the

sleeve's current eligible constituents would have performed in each historical month. The superscript (t) records that the composition is determined at the current formation date. Because the eligible constituent set is redefined at each formation date, the reconstructed historical sleeve-return series may change when a new ETF enters the sleeve; this affects current estimates but never revises an earlier portfolio decision. Since every current constituent has 60 continuous prior returns, the resulting sleeve panel is balanced over both the 24-month mean window and the 60-month covariance window. Holding composition fixed prevents changes in historical sleeve returns from being confused with month-to-month changes in membership.

Expected returns and covariances use different windows. Let $W_\mu = 24$ in the baseline and $W_\Sigma = 60$ throughout. For a generic window W , define the total regime weight and effective sample size as

$$G_{k,t}(W) = \sum_{\tau=t-W}^{t-1} \eta_{\tau k}^{(t)}, \quad (2.40)$$

$$M_{k,t}^{\text{eff}}(W) = \frac{G_{k,t}(W)^2}{\sum_{\tau=t-W}^{t-1} \left(\eta_{\tau k}^{(t)}\right)^2}. \quad (2.41)$$

The effective sample size equals the number of state observations when assignments are zero or one. It becomes smaller when probabilities are diffuse, reflecting that a collection of weak state memberships contains less state-specific information than the same number of clear assignments.

The probability-weighted regime mean is estimated over W_μ months:

$$\hat{\mu}_{k,t}^{(W_\mu)} = \frac{\sum_{\tau=t-W_\mu}^{t-1} \eta_{\tau k}^{(t)} R_{\mathcal{G}_{t,\tau}}^{(t)}}{G_{k,t}(W_\mu)}, \quad (2.42)$$

where $R_{\mathcal{G}_{t,\tau}}^{(t)}$ is the vector of active sleeve returns. For covariance centring, let $\hat{\mu}_{k,t}^{(60,c)}$ denote the same probability-weighted state mean calculated over the 60-month covariance window. The regime covariance is then estimated using the unequal-weight degrees-of-freedom correction

$$D_{k,t}(60) = G_{k,t}(60) - \frac{\sum_{\tau=t-60}^{t-1} \left(\eta_{\tau k}^{(t)}\right)^2}{G_{k,t}(60)}, \quad (2.43)$$

and

$$S_{k,t}^r = \frac{1}{D_{k,t}(60)} \sum_{\tau=t-60}^{t-1} \eta_{\tau k}^{(t)} \left(R_{\mathcal{G}_{t,\tau}}^{(t)} - \hat{\mu}_{k,t}^{(60,c)} \right) \left(R_{\mathcal{G}_{t,\tau}}^{(t)} - \hat{\mu}_{k,t}^{(60,c)} \right)^\top. \quad (2.44)$$

The mean inside the covariance expression is the probability-weighted 60-month state mean because covariance centring must use the same observations as the covariance window. It is an intermediate risk-estimation quantity and does not replace the 24-month expected return used in the portfolio objective.

Even at the sleeve level, each regime contains substantially less information than the unconditional windows. The conditional estimates are therefore pooled toward unconditional sleeve inputs. Let $\hat{\mu}_t^{U,(W_\mu)}$ be the W_μ -month unconditional sleeve mean and let $\hat{\Sigma}_t^{U,\text{QIS}}$ be the 60-month sleeve QIS covariance. One prior-equivalent constant, $\kappa = 36$, governs both pooling rules, while separate effective sample sizes acknowledge the different mean and covariance windows:

$$\lambda_{k,t}^\mu = \frac{M_{k,t}^{\text{eff}}(W_\mu)}{M_{k,t}^{\text{eff}}(W_\mu) + \kappa}, \quad (2.45)$$

$$\lambda_{k,t}^\Sigma = \frac{M_{k,t}^{\text{eff}}(60)}{M_{k,t}^{\text{eff}}(60) + \kappa}, \quad \kappa = 36. \quad (2.46)$$

The pooled moments are

$$\tilde{\mu}_{k,t}^{(W_\mu)} = \lambda_{k,t}^\mu \hat{\mu}_{k,t}^{(W_\mu)} + (1 - \lambda_{k,t}^\mu) \hat{\mu}_t^{U,(W_\mu)}, \quad (2.47)$$

$$\tilde{\Sigma}_{k,t} = \lambda_{k,t}^\Sigma S_{k,t}^r + (1 - \lambda_{k,t}^\Sigma) \hat{\Sigma}_t^{U,\text{QIS}}. \quad (2.48)$$

The common pooling constant is set to $\kappa = 36$, so the unconditional estimate contributes the equivalent of three years of monthly observations. Under the implemented rule, an effective regime sample of 36 gives equal weight to the conditional and unconditional estimates. One common κ is used across regimes, means, and covariances to avoid separately tuning multiple hyperparameters; the resulting reliability weights still differ with their effective samples. Since the positive-definite unconditional QIS matrix always receives positive weight, the pooled covariance remains suitable for optimization even when the regime-specific covariance is singular.

The state for month t is unknown at formation. Predictive inputs therefore

integrate across the three one-step-ahead probabilities. The regime-conditioned expected-return vector is

$$\hat{\mu}_t^{RS, (W_\mu)} = \sum_{k=1}^3 \pi_{k,t|t-1} \tilde{\mu}_{k,t}^{(W_\mu)}. \quad (2.49)$$

The predictive covariance follows the law of total covariance:

$$\begin{aligned} \hat{\Sigma}_t^{RS} = \sum_{k=1}^3 \pi_{k,t|t-1} & \left[\tilde{\Sigma}_{k,t} + \left(\tilde{\mu}_{k,t}^{(W_\mu)} - \hat{\mu}_t^{RS, (W_\mu)} \right) \right. \\ & \left. \times \left(\tilde{\mu}_{k,t}^{(W_\mu)} - \hat{\mu}_t^{RS, (W_\mu)} \right)^\top \right]. \end{aligned} \quad (2.50)$$

The first component is expected covariance within states. The second captures uncertainty arising from differences between state means. Omitting the second component would understate risk precisely when the economic states imply materially different expected returns. An analogous set of pooled means is calculated from sleeve excess returns for use in the macro Black–Litterman views.

2.3.4 Matched Black–Litterman Comparison

Following Black & Litterman (1992), Black–Litterman is used as a disciplined expected-return framework rather than as an additional covariance model. The empirical implementation also adapts the view-based structure of Bessler *et al.* (2017) to the present sleeve classification. Both H2 versions start from the policy benchmark and the same ETF-level QIS covariance. They differ only in whether sleeve views are generated from recent returns or from predicted macro-regime conditions.

Let w_t^{Policy} be the policy portfolio in (2.16). Its reference-portfolio-implied monthly excess returns are

$$\Pi_t = \delta \hat{\Sigma}_t^{\text{QIS}} w_t^{\text{Policy}}, \quad \delta = 2. \quad (2.51)$$

The term *reference-portfolio-implied* is used instead of *market-equilibrium* because the universe contains overlapping regional, sector, factor, and real-asset exposures for which a unique market-capitalization portfolio is not defined. The prior is

$$\mu_t^e \sim \mathcal{N}(\Pi_t, \tau \hat{\Sigma}_t^{\text{QIS}}), \quad \tau = 0.1. \quad (2.52)$$

Following Bessler *et al.* (2017), the reverse-optimization risk-aversion coefficient is fixed at $\delta = 2$ and the prior-uncertainty scalar at $\tau = 0.1$. The former is retained in the final optimizer so that the matched Black–Litterman strategies differ through their views rather than the implied-return scale; the latter assigns finite but restrained uncertainty to the prior and is treated as a calibration rather than a fitted parameter.

For a clean matched comparison, both Black–Litterman variants use the same admissible sleeve set

$$\mathcal{A}_t^{BL} = \{g \in \mathcal{G}_t : n_{g,t}^H \geq 12 \text{ and } n_{g,t}^M \geq 12\}. \quad (2.53)$$

The view matrix P_t contains one row for each sleeve in $\mathcal{A}_t^{BL}$. For sleeve g ,

$$P_{gi,t} = \begin{cases} 1/N_{g,t}, & i \in \mathcal{U}_{g,t}, \\ 0, & i \notin \mathcal{U}_{g,t}. \end{cases} \quad (2.54)$$

Each row therefore represents an equal-weight sleeve portfolio. Views are expressed on 15 economically interpretable portfolios rather than on as many as 71 individual ETF means, which reduces noise and aligns the information layer with the benchmark and regime allocation.

The historical view on sleeve g is its trailing mean excess return:

$$Q_{g,t}^{H,(W_\mu)} = \frac{1}{W_\mu} \sum_{\ell=1}^{W_\mu} (R_{g,t-\ell}^{(t)} - r_{f,t-\ell}), \quad W_\mu = 24 \text{ in the baseline.} \quad (2.55)$$

The macro view uses the predicted regime-conditioned sleeve excess return:

$$Q_{g,t}^{M,(W_\mu)} = \sum_{k=1}^3 \pi_{k,t|t-1} \tilde{\mu}_{g,k,t}^{e,(W_\mu)}, \quad (2.56)$$

where $\tilde{\mu}_{g,k,t}^{e,(W_\mu)}$ is constructed from excess returns using the same state probabilities, expected-return window, and pooling rule as (2.47). The macro variables do not impose qualitative trading rules such as automatically buying commodities when one indicator rises. They affect the view only through estimated state probabilities and the observed historical behaviour of each sleeve within those states.

View uncertainty is estimated from genuine prior forecast errors. For view type $v \in \{H, M\}$,

$$e_{g,\tau}^v = Q_{g,\tau}^v - \left(R_{g,\tau}^{(\tau)} - r_{f,\tau} \right). \quad (2.57)$$

A sleeve enters the common view matrix only after at least 12 errors are available for both view types. Thereafter, each version estimates its own diagonal uncertainty element from up to the most recent 36 type-specific errors:

$$\Omega_{gg,t}^v = \max \left\{ \text{Var} \left(e_{g,\tau}^v : \tau \in \mathcal{H}_{g,t}^v \right), 10^{-8} \right\}, \quad (2.58)$$

where $\mathcal{H}_{g,t}^v$ contains the latest $\min\{36, n_{g,t}^v\}$ stored errors and $n_{g,t}^v \geq 12$. Using more than the minimum 12 observations makes confidence less dependent on one unusual year while preserving adaptation. Off-diagonal elements are zero because estimating a full forecast-error covariance from such short histories would add another unstable matrix. A sleeve without sufficient error history remains in the reference portfolio but receives no active view.

For either view type, the posterior excess-return vector is

$$\hat{\mu}_t^{BL,v} = \left[(\tau \hat{\Sigma}_t^{\text{QIS}})^{-1} + P_t^\top (\Omega_t^v)^{-1} P_t \right]^{-1} \left[(\tau \hat{\Sigma}_t^{\text{QIS}})^{-1} \Pi_t + P_t^\top (\Omega_t^v)^{-1} Q_t^v \right]. \quad (2.59)$$

The final ETF allocation is

$$w_t^{BL,v} = \arg \max_{w_t \in \mathcal{W}_t} \left\{ w_t^\top \hat{\mu}_t^{BL,v} - \frac{\delta}{2} w_t^\top \hat{\Sigma}_t^{\text{QIS}} w_t \right\}. \quad (2.60)$$

The QIS covariance remains the portfolio-risk matrix; Black–Litterman changes expected excess returns only. Using the same δ in reverse optimization and final allocation means that, in the absence of active views and binding constraints, the policy reference portfolio is recovered. Consequently, comparing $w_t^{BL,H}$ with $w_t^{BL,M}$ isolates whether the macro-regime view provides more useful information than a trailing-return view.

2.3.5 Matched Sleeve-Level Mean–Variance Comparison

The second H2 comparison allows macro information to affect both expected returns and covariance. It is conducted at the 15-sleeve level because dividing historical observations across three states cannot support separate 71-by-71 conditional covariance estimates. The lower dimension is also economically useful: the model selects among geographic equity, sector, factor, Treasury-duration,

commodity, and real-estate decisions, while diversification within each selected sleeve remains transparent.

Let x_t denote the vector of active sleeve weights. If the weight of sleeve g is divided equally among its $N_{g,t}$ eligible ETFs, the 20% ETF cap implies

$$0 \leq x_{g,t} \leq \bar{x}_{g,t} = \min\{1, 0.20N_{g,t}\}. \quad (2.61)$$

The sleeve feasible set is

$$\mathcal{X}_t = \left\{x_t : \mathbf{1}^\top x_t = 1, \quad 0 \leq x_{g,t} \leq \bar{x}_{g,t}\right\}. \quad (2.62)$$

The unconditional control uses the W_μ -month sleeve mean and 60-month sleeve QIS covariance:

$$x_t^U = \arg \max_{x_t \in \mathcal{X}_t} \left\{x_t^\top \hat{\mu}_t^{U, (W_\mu)} - \frac{\gamma}{2} x_t^\top \hat{\Sigma}_t^{U, \text{QIS}} x_t\right\}, \quad \gamma = 5. \quad (2.63)$$

The regime-conditioned portfolio changes only the moments supplied to the same objective:

$$x_t^{RS} = \arg \max_{x_t \in \mathcal{X}_t} \left\{x_t^\top \hat{\mu}_t^{RS, (W_\mu)} - \frac{\gamma}{2} x_t^\top \hat{\Sigma}_t^{RS} x_t\right\}. \quad (2.64)$$

Both allocations are mapped back to ETFs by equal weighting within each sleeve:

$$w_{i,t} = \frac{x_{g,t}}{N_{g,t}}, \quad i \in \mathcal{U}_{g,t}. \quad (2.65)$$

The two portfolios therefore share the same sleeves, feasible set, risk aversion, monthly rebalancing, and within-sleeve allocation. Their difference is whether expected returns and risk are estimated unconditionally or integrated across predicted macro-financial states.

2.4 Hypothesis 3: Retail-Investor Implementation

Hypothesis 3 does not introduce another optimized model. It carries the policy benchmark, hierarchical equal-weight benchmark, four H1 strategies, two Black–Litterman strategies, and two sleeve-level MVO strategies into a common retail wealth path. The strategy set is fixed before H1 and H2 results are observed; weak gross performers are not removed. H3 changes only the implementation environment and asks two related questions: how much contri-

butions and trading costs alter each strategy's own outcome, and where each net outcome lies relative to a distribution of simple Random DCA paths.

2.4.1 Monthly Contributions and Trading Costs

Wealth is measured in contribution units to avoid attaching the results to an arbitrary currency scale. Initial wealth is 20 units and the investor contributes one unit at the beginning of every month. Multiplying the complete schedule by USD 500 gives the equivalent interpretation of USD 10,000 initial capital and USD 500 monthly contributions. Fractional shares and proportional costs make all percentage results invariant to this common scaling.

Let $V_{s,t-1}$ be end-of-month wealth after the month- $t - 1$ return and let $C = 1$ be the contribution at the beginning of month t . Available wealth before trading is

$$A_{s,t} = V_{s,t-1} + C. \quad (2.66)$$

Let $H_{i,s,t}^-$ be the pre-trade market value of ETF i after the contribution has been added as cash, and let $H_{i,s,t}^+$ be its post-trade value. The signed trade is

$$\Delta H_{i,s,t} = H_{i,s,t}^+ - H_{i,s,t}^-. \quad (2.67)$$

The baseline transaction cost is five basis points on the absolute value of every purchase and sale:

$$TC_{s,t} = c \sum_{i \in \mathcal{U}_t} |\Delta H_{i,s,t}|, \quad c = 0.0005. \quad (2.68)$$

Applying the same rate to both sides of a trade avoids treating sales or purchases as costless. Five basis points is a transparent low-cost ETF baseline at the lower end of the 5–50 basis-point scenarios examined by Bessler *et al.* (2017); it is chosen to isolate turnover under proportional costs, not to reproduce a complete historical broker tariff. No minimum commission, time-varying bid–ask spread, market-impact function, or broker-specific tariff is added, because those features would require security- and investor-specific assumptions that cannot be applied consistently across the long sample.

Post-trade holdings satisfy

$$H_{i,s,t}^+ = w_{i,t}^s V_{s,t}^+, \quad V_{s,t}^+ = A_{s,t} - TC_{s,t}. \quad (2.69)$$

Since transaction cost depends on the trade required to reach weights defined on post-cost wealth, $V_{s,t}^+$ solves

$$V_{s,t}^+ = A_{s,t} - c \sum_{i \in \mathcal{U}_t} \left| w_{i,t}^s V_{s,t}^+ - H_{i,s,t}^- \right|. \quad (2.70)$$

The scalar equation is solved iteratively and checked against the budget identity. Fractional shares allow the target weights to be reached without integer-rounding residuals. Any remaining cash is below the numerical tolerance and is not treated as a strategic investment.

After trading, wealth at the end of month t is

$$V_{s,t} = \sum_{i \in \mathcal{U}_t} H_{i,s,t}^+ (1 + R_{i,t}). \quad (2.71)$$

Every strategy receives the same initial wealth, contribution amount, contribution dates, universe, and cost coefficient. Differences in terminal wealth therefore arise from returns and the amount and timing of trading generated by the target-weight path.

2.4.2 Random DCA Benchmark

Random DCA is a distributional retail benchmark rather than a passive market index. It represents an investor who follows the same contribution schedule and cost rule but chooses holdings through a simple random procedure. At each semiannual review date, approximately one third of the eligible ETFs are sampled uniformly without replacement. The one-third fraction is a simulation-design choice that creates diversified but meaningfully different subsets without converging toward full-universe equal weighting. At least five ETFs are selected so that an equal-weight draw remains compatible with the 20% cap.

The selected subset remains fixed for six months. This semiannual review interval is a simulation-design choice representing a low-maintenance contribution strategy: monthly reselection would create unnecessary turnover, whereas annual review would adapt more slowly as the eligible universe expands. At the review date, the portfolio is rebalanced to equal weights across the selected ETFs. In the intervening months, the new contribution is divided equally among the same selected ETFs, while existing holdings are allowed to drift and are not sold merely to restore equal weights.

The experiment generates $B = 1,000$ independently drawn paths using a fixed seed stored in the replication code. All paths experience the same histori-

cal return sequence and differ only in ETF selection. The resulting distribution is conditional on the realized sample; it is not a sampling distribution over alternative histories of the economy. The exact subset-size rule, review timing, and cost recursion are provided in Section E.7.

Random DCA serves a different purpose from the policy and hierarchical benchmarks. A strategy may outperform one fixed benchmark yet remain worse than many simple allocations that could have been chosen without estimation. Its percentile within the Random DCA distribution therefore might indicate whether systematic allocation produced an unusually strong outcome relative to uninformed ETF selection.

2.5 Backtesting and Performance Evaluation

The preceding sections define all target-weight rules before the evaluation procedure is introduced. Consolidating the backtest here avoids repeating the same timing, wealth, and performance equations in H1 and H2 and makes clear that every model is evaluated on the same dates.

2.5.1 Common Starting Date and Monthly Sequence

The empirical analysis uses two code-generated evaluation periods. Standalone H1 begins at the first month in which its same-universe benchmarks and unconditional strategies are feasible. The broad H1/H2 comparison and H3 begin later because they also require a valid recursive HMM and sufficient genuine forecast-error history for the matched Black–Litterman views.

Let t_0^{H1} be the first standalone H1 month satisfying the following conditions:

1. every ETF in $\mathcal{U}_{t_0}$ has 60 prior monthly returns;
2. equities, Treasuries, commodities, and listed real estate are represented;
3. at least five ETFs are eligible, so full investment is feasible under the 20% cap;
4. the policy and hierarchical equal-weight portfolios satisfy the ETF cap.

Formally,

$$t_0^{H1} = \min \left\{ t : \mathcal{G}_{E,t} \neq \emptyset, \mathcal{G}_{B,t} \neq \emptyset, \mathcal{G}_{C,t} \neq \emptyset, \mathcal{G}_{RE,t} \neq \emptyset, \right. \\ \left. N_t \geq 5, \max_i w_{i,t}^{\text{Policy}} \leq 0.20, \max_i w_{i,t}^{\text{HEW}} \leq 0.20 \right\}. \quad (2.72)$$

The resulting standalone H1 period is December 2009 through June 2026, comprising 199 holding months. The common H1/H2 and H3 period begins in March 2012 and ends in June 2026, comprising 172 holding months. The macro archive supplies its 120th complete observation in February 2011, so the first HMM holding month is March 2011. Twelve subsequent holding months accumulate genuine paired Black–Litterman forecast errors; the common performance period therefore begins in March 2012. Common-period H1 results are obtained by slicing the already generated H1 paths and rebasing wealth to one. H1 models are not re-estimated from the later date, and H3 replays the ten common-period target paths without re-estimating H1 or H2.

Before the common period, a recursive warm-up pass estimates the HMM whenever 120 macro observations are available and stores Black–Litterman forecast errors whenever a sleeve can form a genuine view. These quantities are generated using their historical information boundaries. The warm-up allows view reliability to be measured from the beginning of the common evaluation period without importing later forecast errors.

For every month in the applicable evaluation period, the backtest follows the same sequence:

1. close month $t - 1$ and record existing holdings and wealth;
2. update $\mathcal{I}_{t-1}$ with ETF prices, exchange rates, risk-free yields, and macro releases available by that date;
3. construct $\mathcal{U}_t$, $\mathcal{G}_t$, and the current sleeve membership counts;
4. estimate the 60-month ETF and sleeve QIS covariances and all strategy-specific expected-return inputs;
5. re-estimate the HMM, assign economic state labels, and calculate $\pi_{k,t|t-1}$ for macro-informed strategies;
6. solve for every strategy's target weights and verify all budget and bound constraints;

7. for H3, add the monthly contribution, execute the required trades, and deduct transaction costs;
8. hold the resulting portfolio through month t ; and
9. record returns, wealth, weights, costs, and model diagnostics.

When a new ETF becomes eligible, it enters at step 3 and can receive weight under every strategy's existing rule. The backtest does not restart and no historical return is rewritten. Since the processed ETF histories have no internal missing observations, each eligible ETF supplies the required balanced 60-month estimation window.

Numerical solutions are accepted only when all weights are finite, the budget residual is below 10^{-8} , and every bound violation is below 10^{-8} . Convex problems are solved from a feasible starting point; nonlinear ERC and maximum-diversification problems use the previous feasible allocation and the current policy allocation as separate initializations. If repeated attempts fail, the policy portfolio is used for that strategy-month and the event is retained in the diagnostic output. The fallback is fixed before performance is observed.

2.5.2 Frictionless Evaluation for H1 and H2

Hypotheses 1 and 2 evaluate the strategies as portfolio-allocation rules before retail cash flows and transaction costs are introduced. Each strategy therefore begins with one unit of wealth at the start of the applicable reported period, receives no subsequent contributions, and is rebalanced without trading frictions. The corresponding wealth index evolves as

$$G_{s,t} = G_{s,t-1} (1 + R_{p,t}^s). \quad (2.73)$$

This construction isolates the performance of the allocation rule from the scale and timing of investor-specific cash flows. The evaluation combines return-based measures, path-dependent risk measures, and an implementation diagnostic. The first group summarizes the average return and variability of the monthly strategy returns over the common evaluation period. Let T_s denote the number of evaluation months. The annualized arithmetic return is

$$\bar{R}_s^{ann} = 12 \left(\frac{1}{T_s} \sum_{t=t_0}^T R_{p,t}^s \right), \quad (2.74)$$

while annualized volatility is

$$\sigma_s^{ann} = \sqrt{12} \text{sd}(R_{p,t}^s). \quad (2.75)$$

Risk-adjusted performance is measured by the annualized Sharpe ratio. Because the monthly risk-free return varies through time, the calculation is based on the complete monthly excess-return series rather than on one constant annual rate:

$$SR_s = \sqrt{12} \frac{\overline{R_{p,t}^s - r_{f,t}}}{\text{sd}(R_{p,t}^s - r_{f,t})}. \quad (2.76)$$

The preceding measures summarize the distribution of monthly returns but do not fully describe the path through which wealth is accumulated. Maximum drawdown is therefore calculated from the frictionless wealth index and captures the largest peak-to-trough decline experienced during the evaluation period:

$$MDD_s = \min_t \left\{ \frac{G_{s,t}}{\max_{\tau \leq t} G_{s,\tau}} - 1 \right\}. \quad (2.77)$$

Final cumulative wealth is reported alongside maximum drawdown so that the terminal outcome and the severity of the path leading to it can be evaluated jointly.

In addition to performance, the analysis records how intensively each strategy changes its allocation. Turnover is measured relative to the portfolio weights after they have drifted with the preceding month's asset returns. For strategy s , the pre-trade weight of ETF i immediately before formation of the month- t portfolio is

$$w_{i,s,t}^- = \frac{w_{i,s,t-1} (1 + R_{i,t-1})}{1 + R_{p,t-1}^s}. \quad (2.78)$$

The one-way drift-adjusted turnover required to move from the pre-trade portfolio to the new target allocation is

$$TO_{s,t} = \frac{1}{2} \sum_{i \in \mathcal{U}_{t-1} \cup \mathcal{U}_t} |w_{i,s,t} - w_{i,s,t}^-|. \quad (2.79)$$

An ETF that is absent from either of the two eligible universes is assigned a weight of zero on the corresponding side of the calculation. The factor one half prevents purchases and sales associated with the same rebalancing event

from being counted twice. Initial portfolio formation is excluded from average turnover because no previously invested portfolio exists. In H1 and H2, turnover is reported as a diagnostic of allocation stability rather than as a direct cost deduction; its monetary consequences are introduced separately in the H3 retail implementation.

The H1 and H2 result tables therefore report annualized return, annualized volatility, the Sharpe ratio, maximum drawdown, final cumulative wealth, and average turnover. Together, these measures distinguish return generation, risk exposure, wealth-path severity, and the degree of portfolio adjustment required by each strategy. The study remains an out-of-sample comparative backtest, and small differences between strategies are not interpreted as formal statistical significance unless supported by a separate inference procedure.

2.5.3 Contribution-Based Evaluation for H3

H3 focuses on the investor's accumulated net wealth. Total contributed capital through the final month is

$$K_T = V_0 + \sum_{t=t_0}^T C_t. \quad (2.80)$$

The principal outcomes are terminal net wealth $V_{s,T}$, net investment gain $V_{s,T} - K_T$, and cumulative direct transaction cost

$$TC_s^{cum} = \sum_{t=t_0}^T TC_{s,t}. \quad (2.81)$$

For each target-weight strategy, a gross contribution-based counterfactual is generated with the same target weights and cash flows but $c = 0$. The difference between gross and net terminal wealth measures total terminal cost drag, including the forgone compounding of costs paid earlier. This comparison is distinct from cumulative direct cost because one dollar paid early also loses its subsequent investment return.

The net time-weighted return treats the contribution as an external beginning-of-period flow:

$$R_t^{TW,s} = \frac{V_{s,t}}{V_{s,t-1} + C_t} - 1. \quad (2.82)$$

Its annualized geometric value is

$$R_s^{TW,ann} = \left[\prod_{t=t_0}^T (1 + R_t^{TW,s}) \right]^{12/T_s} - 1. \quad (2.83)$$

The monthly money-weighted return r_s^{MW} is the internal rate that equates the present value of contributions with terminal wealth:

$$0 = -V_0 - \sum_{j=1}^{T_s} \frac{C_j}{(1 + r_s^{MW})^{j-1}} + \frac{V_{s,T}}{(1 + r_s^{MW})^{T_s}}. \quad (2.84)$$

The annualized money-weighted return is $(1 + r_s^{MW})^{12} - 1$. Reporting both measures separates the return of the strategy's investment path from the return experienced under the specified cash-flow schedule.

2.5.4 Random DCA Distributional Comparison

For a performance measure M for which higher values are preferable, the Random DCA percentile of strategy s is

$$\text{Pct}_s(M) = \frac{1}{B} \sum_{b=1}^B \mathbf{1}\{M_b^{RDA} \leq M_s\}. \quad (2.85)$$

A value of 0.90 means that the strategy exceeds 90% of the simulated Random DCA paths on that measure. For loss measures, the inequality is reversed so that a higher percentile always indicates a more favourable outcome.

The upper-tail conditional exceedance frequency is

$$\hat{p}_s^{RDA}(M) = \frac{1 + \sum_{b=1}^B \mathbf{1}\{M_b^{RDA} \geq M_s\}}{B + 1}. \quad (2.86)$$

The plus-one correction prevents a reported frequency of exactly zero in a finite simulation. This quantity describes how often simulated Random DCA paths meet or exceed the implemented outcome within the fixed design. It is not a classical population-level p -value because all paths share the same realized market history and differ only through the pre-specified random-selection mechanism. Its interpretation is therefore conditional on the eligible universe, contribution schedule, cost rule, and six-month review design.

2.5.5 Stationary-Bootstrap Inference for H1 and H2

Sampling uncertainty in the H1 and H2 Sharpe-ratio differences is assessed through the stationary bootstrap of Politis & Romano (1994). For each sample month, one joint observation contains all strategy returns relevant to the hypothesis family and the corresponding time-varying risk-free return. Each bootstrap replication resamples one common sequence of month indices for the complete joint vector. Strategies and the risk-free series are therefore never re-sampled independently, preserving contemporaneous cross-strategy dependence and the paired nature of each comparison.

The resampled indices are constructed from circular blocks with geometrically distributed lengths. At each position, a new block begins at a uniformly selected historical month with probability $1/L$; otherwise, the index advances to the next historical month and wraps to the beginning after the final observation. The primary expected block length is $L = 12$ months, with $L = 6$ and $L = 24$ reported as sensitivity specifications. Each block length uses 10,000 deterministic-seed replications. In each pseudo-history, the annualized Sharpe ratio is recomputed using (2.76), and the paired difference is

$$\Delta_{A,B}^{*(b)} = SR_A^{*(b)} - SR_B^{*(b)}. \quad (2.87)$$

Percentile intervals are formed from the 2.5th and 97.5th percentiles of the bootstrap differences. The positive-difference share records the fraction of pseudo-histories in which strategy A has the higher Sharpe ratio and is interpreted as a stability measure rather than a classical p -value. For the directional alternative, let $\hat{\Delta}$ be the observed difference and define $T_b^* = \Delta_b^* - \hat{\Delta}$. The plus-one centred-bootstrap tail probability is

$$p = \frac{1 + \sum_{b=1}^B \mathbf{1}\{T_b^* \geq \hat{\Delta}\}}{B + 1}. \quad (2.88)$$

For H1, each strategy must exceed both Policy and HEW. Its inferential sub-hypothesis therefore uses an intersection–union construction: the raw strategy-level p -value is the larger of its two benchmark-specific tail probabilities. Holm’s step-down procedure (Holm 1979) is then applied across the four H1 strategies. The two matched H2 tail probabilities are adjusted separately as the H2 family. These inferential results qualify rather than redefine the frozen sign-based descriptive classifications.

The stationary bootstrap and Random DCA answer different questions. Random DCA varies ETF selection while holding the realized market history fixed; the stationary bootstrap varies the representation of observed market episodes while keeping the fitted strategy paths fixed. The bootstrap is conditional on the selected ETF universe and does not re-estimate eligibility, HMM states, QIS covariances, expected returns, Black–Litterman views, or portfolio weights within each replication. It therefore does not capture complete model-estimation uncertainty, remove survivorship or selection risk, establish causality, generate structurally different macroeconomic histories, or imply future persistence. The procedure is used only to qualify the H1 and H2 Sharpe-ratio comparisons. It is not used to assess H3 or to test contribution-based terminal wealth because H3 evaluates net time-weighted return and terminal wealth under chronologically ordered contributions and transaction costs, whereas the bootstrap statistic is a paired Sharpe-ratio difference calculated from the frozen frictionless H1 and H2 return paths. Furthermore, H3 utilizes the same optimized saved target-weight paths as the common-period frictionless analysis so stationary bootstrap inference is adding value to H3.

2.5.6 Robustness Specifications and Model Diagnostics

Robustness analysis is intentionally narrow so that it does not become an ex post search over many parameter combinations. The baseline expected-return window is 24 months. Mean-dependent models are re-estimated with 12- and 48-month windows while the eligible universe, 60-month covariance window, QIS estimator, constraints, risk-aversion parameters, cost assumptions, and evaluation dates remain fixed. The window change applies consistently to ETF MVO, unconditional sleeve MVO, historical Black–Litterman views, and regime-conditioned expected-return means. Covariance-only H1 strategies are unaffected.

The flat ETF-level $1/N_t$ portfolio is reported alongside the two core hierarchical benchmarks. In H2, one additional diagnostic re-estimates the regime engine without CFNAI-MA3, using only VIX, the term spread, and the credit spread. This financial-only specification is not a third core H2 strategy. Its purpose is to show whether the direct macroeconomic activity variable contributes information beyond financial conditions.

VT is reported in a separate external-reference analysis over the standalone H1 and common periods. Its frictionless path holds one fund continuously. Its

H3 reference uses the same initial wealth, monthly contributions, and proportional purchase cost as the core strategies, but it has no strategic rebalancing and no terminal liquidation. VT does not alter the formal hypothesis assessments.

The three-state baseline is fixed before the portfolio backtest. Information criteria for alternative state counts may be reported as descriptive diagnostics, but they do not select the state count after performance is observed. Chapter 3 also reports HMM transition matrices, state probabilities, state mean characteristics, effective regime sample sizes, pooling weights, covariance eigenvalues, optimizer status, turnover, and fallback months. These diagnostics distinguish an economically weak model from a numerical failure.

2.5.7 Summary of Fixed Methodological Choices

Table 2.5 consolidates the choices that remain fixed throughout the core analysis. The table appears after the full methodology so that each item refers to a concept already defined rather than introducing unexplained parameters.

The empirical architecture is therefore continuous across the chapter. H1 evaluates unconditional allocation on a common growing universe. H2 changes only the information layer through two matched comparisons. H3 retains the complete strategy set and changes only the wealth-accounting environment. Chapter 3 follows the same sequence so that every reported result corresponds to one pre-specified hypothesis and comparison.

Table 2.5: Baseline methodological specification

<i>Design element</i>	<i>Baseline specification</i>
Decision frequency	Monthly; weights for month t use information available by the end of month $t - 1$
ETF universe	71 ETFs organized into 15 permanent economic sleeves and four top-level asset classes
Eligibility	Complete 60-month USD return history before formation; identical $\mathcal{U}_t$ in every core comparison
Base currency	U.S. dollars; local-currency adjusted prices are converted before monthly returns are calculated
H1 and H2 wealth path	One unit of initial wealth; no contributions or transaction costs

Continued on next page

Table 2.5 – continued

<i>Design element</i>	<i>Baseline specification</i>
Common constraints	Long-only, no leverage, fully invested, and $w_{i,t} \leq 0.20$
Expected-return input	24-month arithmetic mean; 12- and 48-month robustness windows
Covariance input	60-month Quadratic-Inverse Shrinkage covariance matrix in every specification
Benchmarks	Policy allocation of 60% equity, 20% Treasuries, 15% commodities, and 5% real estate; hierarchical equal weight as the neutral core benchmark
Black–Litterman	Policy reference portfolio; $\delta = 2$, $\tau = 0.1$; 24-month baseline sleeve views; minimum 12 and maximum 36 prior forecast errors for view uncertainty
Regime variables	Real-time CFNAI-MA3, monthly average log VIX, monthly average 10-year minus 3-month term spread, and monthly average log Baa–Treasury spread
Regime model	Three-state diagonal Gaussian HMM; 120-month minimum history; monthly expanding recursive estimation
Regime moments	24-month conditional means and 60-month conditional covariances, pooled toward unconditional inputs with $\kappa = 36$
H3 contribution scale	Initial wealth of 20 contribution units and one beginning-of-month contribution unit
H3 trading cost	Five basis points of absolute traded value; fractional shares; no minimum commission
Random DCA	1,000 paths; semiannual random selection of approximately one third of eligible ETFs; equal weighting within each draw

Source: Author’s methodological specification.

Chapter 3

Empirical Results

This chapter reports the out-of-sample evidence for the three hypotheses defined in Section 1.8. The empirical sequence follows the design established in Chapter 2. It first evaluates the unconditional allocation strategies over the longer standalone H1 period, then examines the matched macro-conditioned comparisons over the common period, and finally applies the same common-period target paths to the retail contribution and transaction-cost setting. The chapter also reports Vanguard Total World Stock ETF (VT) as a separate external global-equity reference. VT provides market context but does not enter the formal hypothesis assessments.

The baseline data snapshot ends on 30 June 2026. The processed inputs, configuration, and generated results are frozen at this endpoint. Any later data refresh constitutes a new empirical vintage and must be stored under a separate manifest. Reported differences are descriptive realized-sample results unless explicitly identified as conditional simulation evidence. No causal interpretation is assigned to the portfolio comparisons.

3.1 Evaluation Sample and Presentation of Results

The analysis uses two evaluation periods, summarized in Table 3.1. Standalone H1 begins in December 2009, when the growing universe first supports the full unconditional strategy comparison, and ends in June 2026. This period contains 199 holding months. H2 requires a valid recursive HMM and genuine forecast-error histories for the matched Black–Litterman comparison. Its common evaluation period therefore begins in March 2012 and contains 172 holding

months through June 2026. The broad comparison of all ten core strategies and the H3 retail implementation use this same common period. Common-period H1 values slice the already generated H1 paths; the models are not re-estimated from the later start.

H1 and H2 are evaluated as frictionless allocation rules from one unit of initial wealth. H3 changes the accounting basis rather than the target-weight rules: each strategy begins with 20 contribution units, receives one unit at the beginning of every holding month, and incurs a five-basis-point cost on absolute traded value. The H1 tables therefore emphasize return, risk, wealth, turnover, and concentration, while H2 emphasizes matched macro-minus-control differences. H3 reports gross and net contribution-based wealth, direct cost, terminal cost drag, TWR, MWR, and position within the conditional Random DCA distribution.

Table 3.1: Evaluation periods and frozen empirical vintages

Experiment	Period		Months	Scope	Accounting
Standalone H1	Dec. 2009–June 2026		199	6 core + Flat 1/N	Frictionless
H2/common context	Mar. 2012–June 2026		172	10 core strategies	Frictionless
H3	Mar. 2012–June 2026		172	10 core + 1,000 RDA	Net contributions
VT external reference	Both periods		199 / 172	Single global-equity ETF	Frictionless and H3

Note: Standalone H1 uses the longer feasible history. H2, the broad ten-strategy comparison, and H3 use the common period. *Source:* Author’s calculations from the frozen H1, H2, H3, and VT run metadata.

The analysis begins with H1 over the longer standalone period, before moving to the common-period evidence required for H2 and H3.

3.2 Hypothesis 1: Unconditional Portfolio Allocation

Hypothesis 1 examines whether unconditional classical and risk-based allocation strategies achieve higher out-of-sample risk-adjusted performance than the Policy and HEW benchmarks. The Sharpe ratio is the primary assessment metric, while return, volatility, drawdown, wealth, turnover, and concentration describe the associated trade-offs.

3.2.1 Overall Performance

The standalone results show that the value of unconditional allocation was strategy-specific rather than uniform. As reported in Table 3.2, GMV produced the highest Sharpe ratio, at 0.759, compared with 0.560 for Policy and 0.557 for HEW. ERC and MD also exceeded both benchmarks, with Sharpe ratios of 0.687 and 0.688, respectively. These three strategies therefore provide evidence consistent with improved risk-adjusted performance relative to the two simple allocations.

Table 3.2: Standalone out-of-sample performance of unconditional ETF allocation strategies

Strategy	Return	Volatility	Sharpe	Drawdown	Wealth	Turnover	Eff. N
Policy	7.82%	11.40%	0.560	-22.77%	3.277	0.240	23.1
HEW	8.10%	11.95%	0.557	-23.91%	3.395	0.241	28.2
GMV	6.64%	6.81%	0.759	-16.23%	2.887	0.545	7.1
ERC	9.17%	11.24%	0.687	-21.24%	4.107	0.505	27.4
MD	6.63%	7.51%	0.688	-14.18%	2.857	0.580	7.5
MVO	8.86%	14.31%	0.519	-22.31%	3.646	2.352	5.5
Flat 1/ N	9.16%	14.12%	0.546	-25.06%	3.854	0.214	57.0

Note: December 2009–June 2026, 199 holding months. Return and volatility are annualized; turnover is annualized one-way drift-adjusted turnover. Flat 1/ N is a robustness benchmark. *Source:* Author’s calculations from the frozen baseline H1 contracts.

ETF-level MVO provides the central adverse result. Its Sharpe ratio was 0.519, below both Policy and HEW, despite its comparatively high return. The additional return was accompanied by sufficiently high volatility to leave risk-adjusted performance weaker than that of either core benchmark. Thus, the performance table does not support a general ranking in which statistical optimization dominates transparent benchmark allocation.

Flat 1/ N is included in Table 3.2 only as a secondary robustness benchmark. Its construction gives greater aggregate weight to economic exposures represented by more ETF tickers, so it does not replace Policy or HEW in the formal H1 comparison. Its results likewise do not alter the strategy-specific pattern observed for the four principal H1 models. The following subsection separates the risk and wealth trade-offs underlying the principal H1 Sharpe comparisons.

3.2.2 Risk, Drawdown, and Wealth Dynamics

The risk and wealth evidence clarifies that the three favourable Sharpe comparisons arose through different return–risk combinations. GMV attained its lead-

ing Sharpe primarily through low volatility rather than high terminal wealth. MD similarly reduced volatility and produced the smallest drawdown among the reported H1 strategies, but its comparatively low return constrained wealth accumulation. Accordingly, the favourable Sharpe ratios of GMV and MD should not be interpreted as return or wealth dominance.

ERC offered a different balance. It combined a Sharpe ratio above both core benchmarks with the highest final frictionless wealth among the core H1 strategies, ending at 4.107 wealth units. The cumulative paths in Figure 3.1 show how these terminal outcomes developed over the standalone period and make clear that the ranking by final wealth differs from the ranking by Sharpe ratio. In particular, the covariance-based strategies that most strongly reduced risk did not finish with the greatest wealth.

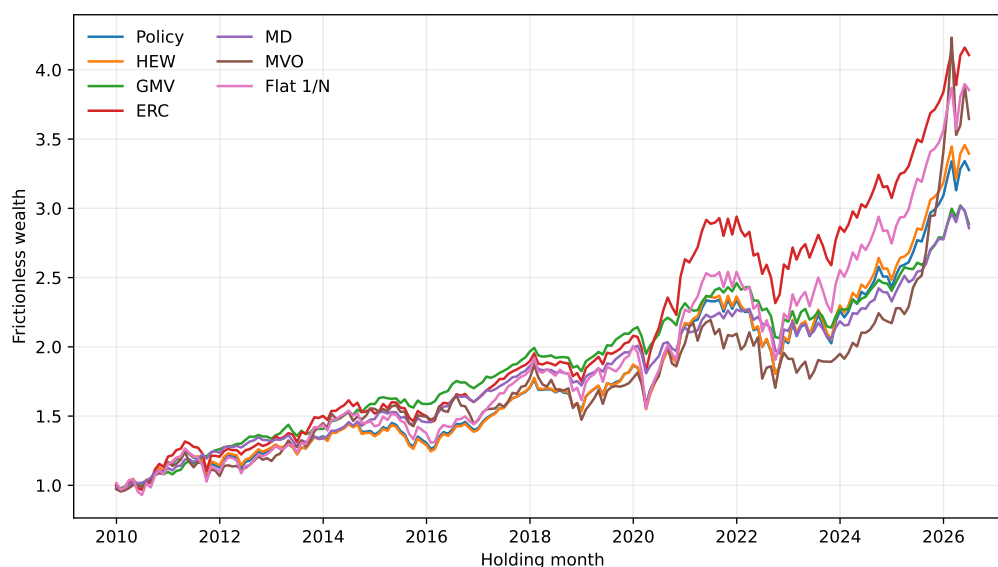


Figure 3.1: Cumulative frictionless wealth of the standalone H1 strategies, December 2009–June 2026. All paths begin at one. Flat $1/N$ is a robustness benchmark rather than a formal same-universe H1 comparator.

Source: Author's calculations from the frozen baseline H1 wealth contract.

MVO accumulated more wealth than Policy and HEW, but its wealth result does not reverse the adverse primary comparison. Its return path entailed materially higher volatility, and its Sharpe ratio remained below both benchmarks. The combined evidence therefore separates terminal accumulation from risk-adjusted performance rather than treating either measure as sufficient on its own.

3.2.3 Portfolio Composition, Concentration, and Turnover

The allocation diagnostics reveal substantial differences in how the H1 strategies generated their outcomes. ERC remained comparatively broad, consistent with its objective of distributing portfolio risk across holdings. By contrast, GMV and MD were concentrated around allocations that frequently reached the ETF weight cap. Their low-volatility outcomes were therefore achieved through selective covariance-based exposures rather than uniformly dispersed capital weights.

MVO was the most concentrated principal strategy and the most turnover-intensive. Its annualized one-way drift-adjusted turnover was 2.352, compared with 0.240 for Policy and 0.241 for HEW. Figure 3.2 presents these turnover and effective-holdings contrasts. Because the H1 comparison is frictionless, turnover is an allocation-stability diagnostic rather than a deducted transaction cost; nevertheless, the magnitude indicates a materially greater implementation burden than for the simple benchmarks.



Figure 3.2: Turnover and concentration diagnostics for the standalone H1 strategies. The upper panel reports the 12-month moving average of one-way drift-adjusted turnover; the lower panel reports effective holdings.

Source: Author's calculations from the frozen baseline H1 turnover and concentration contracts.

Flat 1/N remained the broadest allocation because it spread capital across eligible tickers, although that breadth is partly mechanical in an unbalanced ETF universe. Taken together, the diagnostics show that higher risk-adjusted performance did not correspond to one common portfolio structure: ERC remained broad, whereas GMV and MD obtained their favourable Sharpe ratios with substantially greater concentration.

3.2.4 Assessment of Hypothesis 1

Hypothesis 1. “Classical and risk-based portfolio allocation strategies achieve higher out-of-sample risk-adjusted performance than simple benchmark allocations when evaluated on the same growing international and cross-sector ETF universe.”

Under the explicit sign-based rule, Hypothesis 1 receives partial descriptive support. GMV, ERC, and MD each achieved a higher standalone Sharpe ratio than both Policy and HEW, but ETF-level MVO did not. The favourable evidence is therefore concentrated in the covariance-based and risk-based methods rather than in optimization generally. Moreover, the gains involved distinct trade-offs: GMV and MD improved risk-adjusted performance through lower risk while sacrificing terminal wealth, ERC combined a Sharpe improvement with the highest core H1 wealth, and MVO paired high return with weak risk-adjusted performance, high concentration, and substantially greater turnover.

The stationary-bootstrap analysis provides a separate inferential qualification; the complete results are reported in Section 3.7. ERC’s 95% intervals remain above zero against both benchmarks, and its H1 intersection–union p -value remains below 5% after Holm adjustment ($p = 0.032$). The GMV and MD intervals cross zero against both benchmarks, with adjusted p -values of 0.567, so their positive realized rankings remain imprecisely estimated. MVO’s observed differences are negative and its intervals also cross zero. These results do not alter the partial descriptive classification, but they show that the strongest inferential evidence within H1 is concentrated in ERC.

This assessment remains conditional on the realized standalone sample, the same growing universe, and the frictionless evaluation design. Non-rejection does not establish equality, and the bootstrap does not support a causal claim or uniform superiority over simple benchmark allocation.

3.3 Hypothesis 2: Added Value of Macroeconomic and Financial Information

Hypothesis 2 examines whether macro-conditioned allocation strategies achieve higher out-of-sample risk-adjusted performance than otherwise comparable strategies based only on historical returns or unconditional return moments. The assessment uses the common period from 31 March 2012 to 30 June 2026, comprising 172 holding months. The two matched comparisons provide the primary evidence: macro-view versus historical-view Black–Litterman and regime-conditioned versus unconditional sleeve MVO. Comparisons with the H1 strategies are reported only as secondary common-period context.

3.3.1 Estimated Macro-Financial Regimes

The estimated states were economically distinguishable in the final recursive fit. As reported in Table 3.3, the favourable state combined positive activity with relatively low log VIX and log credit-spread values, whereas the adverse state combined below-average activity with elevated financial stress. In standardized units, the adverse-state mean was -0.970 for real-time CFNAI-MA3, 1.352 for log VIX, and 1.196 for the log credit spread. The intermediate state occupied neither extreme of the stress-based ordering and was distinguished in part by a more positive term-spread mean.

Table 3.3: Economic characteristics and persistence of the final recursive HMM fit

State	CFNAI	Log VIX	Log credit	Term spread	Persistence	Duration
Favourable	0.295	-0.390	-0.867	-0.831	0.985	68.1
Intermediate	0.185	-0.286	0.437	0.769	0.961	25.4
Adverse	-0.970	1.352	1.196	0.556	0.917	12.1

Note: State means are standardized. Duration is in months. The fit uses information from March 2001 through May 2026 for the June 2026 holding decision. *Source:* Author’s calculations from the frozen baseline HMM parameters and transition matrix.

The final recursive transition estimates implied substantial persistence in all three states. Expected durations were 68.1, 25.4, and 12.1 months for the favourable, intermediate, and adverse states, respectively. These values indicate that the estimated states represented persistent macro-financial configurations rather than isolated monthly observations. They remain diagnostics from the final recursive estimate, however, rather than fixed population parameters.

Table 3.4 reports the information supporting the state-specific return moments. Effective sample size measures the concentration of the HMM probability weights rather than a count of calendar months. The favourable state had median effective samples of 15.14 months for means and 35.26 months for covariances. The corresponding intermediate-state values were 5.39 and 27.16 months, while the adverse state had only 1.14 and 8.54 months. State-specific adverse estimates were therefore supported by substantially less effective information.

Table 3.4: Effective regime samples and pooling toward unconditional moments

<i>Panel A: effective samples and pooling weights</i>				
State	N_{eff}^{μ}	N_{eff}^{Σ}	λ^{μ}	λ^{Σ}
Favourable	15.14 [0.00, 24.00]	35.26 [1.00, 51.25]	0.30 [0.00, 0.40]	0.49 [0.03, 0.59]
Intermediate	5.39 [1.70, 24.00]	27.16 [2.66, 50.02]	0.13 [0.05, 0.40]	0.43 [0.07, 0.58]
Adverse	1.14 [0.00, 3.02]	8.54 [2.78, 9.33]	0.03 [0.00, 0.08]	0.19 [0.07, 0.21]
Overall	2.68 [0.00, 20.71]	10.08 [2.53, 44.24]	0.07 [0.00, 0.37]	0.22 [0.07, 0.55]

<i>Panel B: fallback shares</i>		
State	Mean fallback	Covariance fallback
Favourable	15.22%	11.96%
Intermediate	9.24%	0.00%
Adverse	9.78%	0.00%
Overall	11.41%	3.99%

Note: Entries in Panel A are medians with interquartile ranges in brackets, based on one deduplicated observation per formation date and state across 184 formation months, including the 12-month pre-evaluation forecast-error warm-up. Pooling follows $\lambda = N_{\text{eff}} / (N_{\text{eff}} + 36)$. The maximum possible weights are 0.40 for means and 0.625 for covariances under the 24- and 60-month windows. *Source:* Author's calculations from the frozen baseline regime-moment contract.

Pooling weights measure how much state-specific information entered the final portfolio inputs. Median mean-pooling weights were 0.296, 0.130, and 0.031 for the favourable, intermediate, and adverse states, respectively; median covariance-pooling weights were 0.495, 0.430, and 0.192. The remaining weight came from unconditional history. As shown in Figure 3.3, the strength of the conditional adjustment also changed materially through time. This regularization is deliberate: it limits reliance on sparsely supported regimes rather than indicating numerical failure. The theoretical maxima of 0.40 for means and 0.625 for covariances further ensure that even the best-supported state cannot fully replace the unconditional estimate.

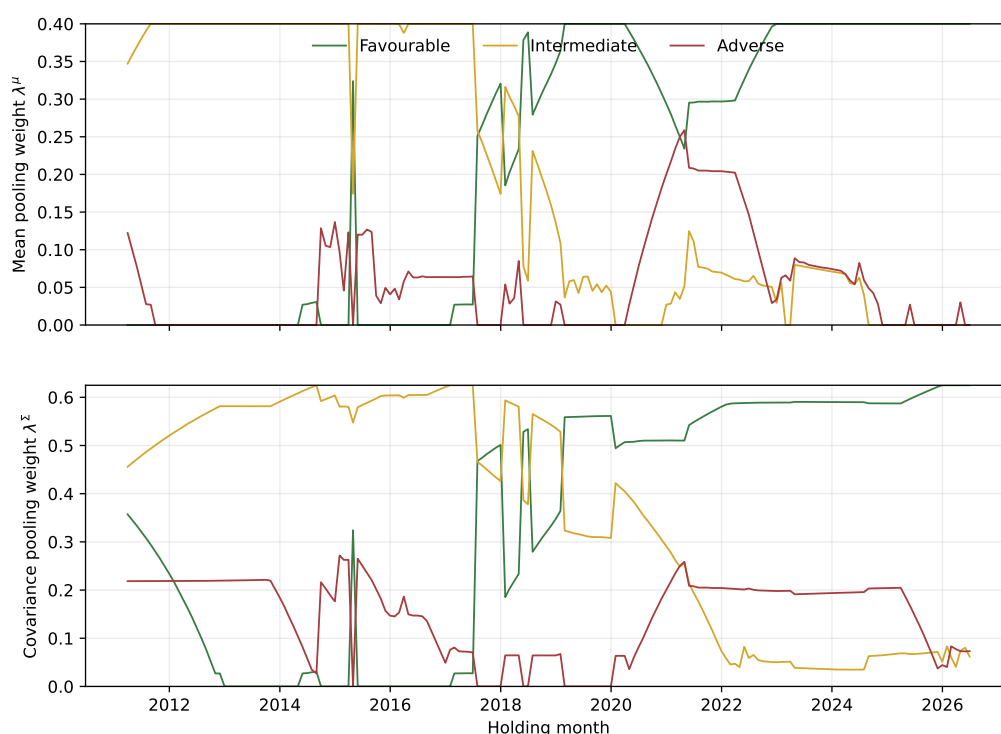


Figure 3.3: Pooling weights applied to state-specific mean and covariance estimates, by HMM state and holding month. The remaining weight is assigned to the corresponding unconditional estimate. Panel bounds equal the theoretical maxima implied by the 24-month mean window, 60-month covariance window, and $\kappa = 36$.

Source: Author's calculations from the deduplicated frozen baseline regime-moment contract.

The downstream allocation effect was moderate in a typical month. The median one-way weight distance was 0.004 between BL-M and BL-H and 0.020 between Sleeve-MVO-RC and Sleeve-MVO-U; the corresponding means were 0.040 and 0.058. Maximum distances of 0.397 and 0.600 show that larger reallocations occurred episodically. An economically interpretable state classification therefore did not automatically imply a large persistent portfolio change, although the conditional engine could materially alter weights in selected months.

Future research could strengthen adverse-state estimation in two complementary ways. First, state-specific moments could use an expanding, time-decayed history, allowing earlier stress episodes to contribute while assigning less weight to older observations. Second, a weekly financial-only HMM could test whether higher-frequency market variables identify stress transitions more precisely; however, the pooling constant and effective-sample interpretation

would need to be rescaled because weekly observations do not provide proportionally independent information.

The recursively estimated one-step predicted probabilities in Figure 3.4 show how the dominant state changed through time. The intermediate state carried most of the probability during much of the early sample, the favourable state dominated from approximately 2017 until early 2020, and the adverse state became dominant around the 2020–2021 stress episode. The favourable state again received most of the predicted probability thereafter. The transitions are probabilistic rather than deterministic break dates, as shown by the short intervals in which probability mass was distributed across states.

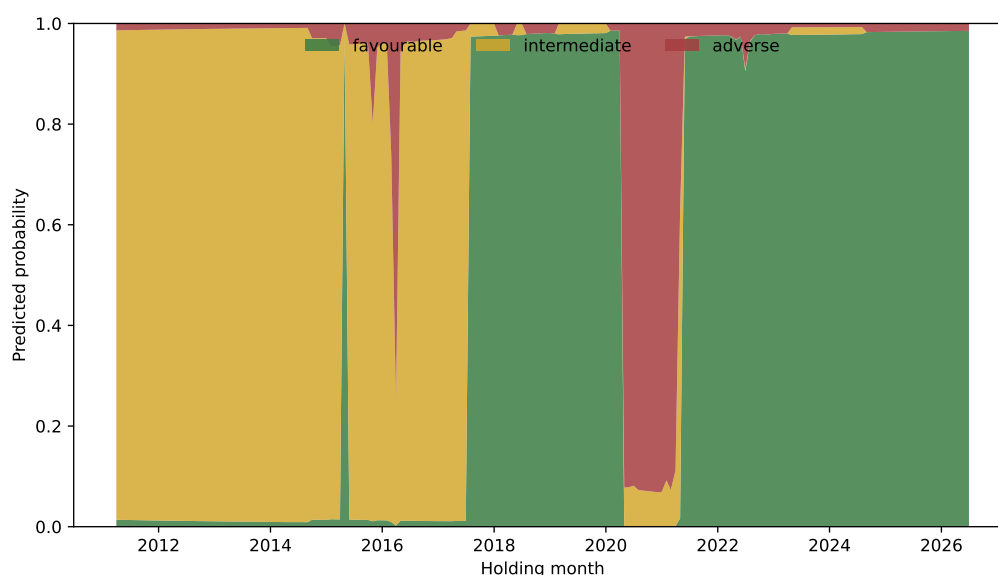


Figure 3.4: One-step predicted probabilities of the favourable, intermediate, and adverse macro-financial states from the recursive baseline HMM. The probabilities available at each formation date sum to one and include the pre-evaluation warm-up period.

Source: Author's calculations from the frozen baseline HMM probability contract.

3.3.2 Historical-View versus Macro-View Black–Litterman

The first matched comparison asks whether regime-conditioned sleeve views improve on views formed from historical returns while the remaining Black–Litterman architecture is held constant. In the baseline specification, Macro-view Black–Litterman (BL-M) achieved an annualized Sharpe ratio of 0.298, compared with 0.282 for Historical-view Black–Litterman (BL-H). The matched increment was therefore 0.016. BL-M's maximum drawdown was marginally

more negative. Table 3.5 reports the remaining baseline differences together with the pre-specified robustness results.

The Sharpe increment remained positive but small with a 12-month expected-return window, at 0.008, and was 0.028 in the additional financial-only HMM diagnostic. In contrast, it changed sign under the 48-month window, where the BL-M minus BL-H difference was -0.038 . The baseline result consequently provides only modest evidence in favour of macro-based views. Its reversal at 48 months shows that the direction of the comparison depends on the expected-return specification and precludes a claim of robust BL-M dominance.

Table 3.5: Matched macro-conditioning increments across baseline and pre-specified variants

Specification	Pair	ΔR	$\Delta\sigma$	ΔSR	ΔDD	ΔW	ΔTO
Baseline (24m)	BL-M – BL-H	+0.21	-0.06	0.016	-0.22	+0.062	+0.073
Baseline (24m)	RC – U MVO	+0.84	+0.06	0.058	+4.21	+0.370	+0.028
12m robustness	BL-M – BL-H	+0.13	+0.05	0.008	+0.13	+0.054	-0.024
12m robustness	RC – U MVO	+0.01	+0.01	0.000	-1.10	+0.005	+0.055
48m robustness	BL-M – BL-H	-0.61	-0.20	-0.038	+2.62	-0.204	+0.170
48m robustness	RC – U MVO	+0.77	+0.06	0.056	+4.09	+0.398	+0.055
Financial-only diagnostic	BL-M – BL-H	+0.39	-0.03	0.028	+0.12	+0.111	+0.126
Financial-only diagnostic	RC – U MVO	+0.84	-0.04	0.061	+4.42	+0.377	+0.190

Note: March 2012–June 2026, 172 holding months. Return, volatility, and drawdown differences are percentage points. Positive drawdown differences indicate a less-negative drawdown. The financial-only row is an HMM diagnostic, not an expected-return-window robustness specification. *Source:* Author’s calculations from the frozen baseline and robustness H2 contracts.

3.3.3 Unconditional versus Regime-Conditioned Sleeve MVO

The second matched comparison evaluates whether regime-conditioned sleeve moments improve on unconditional moments within the same sleeve-level allocation architecture. The baseline evidence is stronger than in the Black–Litterman comparison. Regime-conditioned sleeve MVO (Sleeve-MVO-RC) exceeded Unconditional sleeve MVO (Sleeve-MVO-U) by 0.058 in annualized Sharpe ratio. It also produced a higher annualized return and final frictionless wealth, a less-negative maximum drawdown, and slightly higher annualized volatility. Table 3.5 reports the exact secondary differences.

The robustness results qualify, but do not overturn, this baseline comparison. The Sharpe increment was effectively zero under the 12-month window, at 0.000, but remained close to the baseline result under the 48-month window,

at 0.056, and in the additional financial-only HMM diagnostic, at 0.061. The matched difference was therefore non-negative across the pre-specified variants, although it was not uniformly material. Moreover, the positive financial-only result does not identify a separate contribution from CFNAI-MA3 and should not be interpreted as evidence about the causal role of any individual state variable.

3.3.4 Broader Common-Period Context

The matched comparisons provide the primary H2 evidence; the following common-period ranking supplies secondary context only. For secondary context, Table 3.6 compares all ten core strategies over the same 172-month period. ERC recorded the highest common-period Sharpe ratio, while Sleeve-MVO-RC produced the highest final frictionless wealth. By contrast, BL-H and BL-M had the two lowest common-period Sharpe ratios, at 0.282 and 0.298, and the two lowest final wealth levels, at 1.930 and 1.992 units. The sleeve strategies were therefore competitive along selected dimensions, whereas the Black–Litterman variants remained weak in the broader comparison. These rankings do not replace the matched tests because the H1 and H2 architectures differ in more than their use of macro-financial information.

Table 3.6: Common-period performance of the ten core strategies

Strategy	Return	Volatility	Sharpe	Drawdown	Wealth
ERC	8.81%	10.63%	0.673	-21.24%	3.251
GMV	5.93%	7.02%	0.606	-16.23%	2.254
MD	5.91%	7.65%	0.555	-14.18%	2.233
Sleeve-MVO-RC	9.36%	14.05%	0.548	-22.05%	3.306
Policy	7.54%	11.04%	0.533	-22.77%	2.693
HEW	7.78%	11.72%	0.523	-23.91%	2.756
MVO	9.17%	14.48%	0.519	-22.31%	3.185
Sleeve-MVO-U	8.52%	13.99%	0.490	-26.25%	2.936
BL-M	5.79%	13.87%	0.298	-28.10%	1.992
BL-H	5.58%	13.93%	0.282	-27.88%	1.930

Note: March 2012–June 2026, 172 holding months. This broad ranking is secondary to the matched H2 comparisons. *Source:* Author’s calculations from the frozen common-period H1/H2 contracts.

The 20% cap applies to each ETF rather than to an aggregate economic exposure. Table 3.7 and Figure 3.5 therefore describe the asset-class and sleeve

composition permitted by the existing constraints. Policy and HEW maintained the broadest sleeve allocations, with median effective sleeve counts of 14.3 and 15.0. GMV and MD assigned median weights of approximately 40% to Treasuries and had about 5.8 effective sleeves, while ERC retained a median equity weight of 65.5% and 8.5 effective sleeves.

The MVO and Black–Litterman strategies were more concentrated. MVO’s median largest-sleeve weight was 51.8%, and its largest sleeve exceeded 40% in 73.3% of months. The corresponding shares were 35.5% for both Black–Litterman variants and 74.4% and 76.7% for Sleeve-MVO-U and Sleeve-MVO-RC, respectively. Sleeve-MVO-RC had a median effective sleeve count of 2.3 despite averaging 8.2 effective ETF holdings, while ERC averaged 29.1 effective ETF holdings but 8.5 effective sleeves. ETF-level diversification can therefore coexist with materially narrower economic-exposure diversification. These diagnostics describe the allocations produced by the models; they do not establish that concentration caused the observed performance differences.

Table 3.7: Common-period asset-class exposure and sleeve concentration

<i>Panel A: median target weight by asset class</i>				
Strategy	Equity	Treasuries	Commodities	Real estate
Policy	60.00%	20.00%	15.00%	5.00%
HEW	66.67%	13.33%	13.33%	6.67%
GMV	38.68%	40.00%	21.32%	0.00%
ERC	65.48%	20.00%	9.39%	4.51%
MD	38.89%	40.00%	23.19%	0.00%
MVO	80.00%	0.00%	16.00%	0.00%
BL-H	71.55%	0.00%	9.37%	4.99%
BL-M	71.33%	0.00%	9.00%	4.82%
Sleeve-MVO-U	70.06%	2.43%	13.61%	0.00%
Sleeve-MVO-RC	75.11%	0.00%	11.23%	0.00%

<i>Panel B: sleeve-level concentration</i>				
Strategy	Median largest sleeve	Maximum largest sleeve	Median effective sleeves	Months above 40% largest sleeve
Policy	10.00%	12.00%	14.3	0.00%
HEW	6.67%	10.00%	15.0	0.00%
GMV	20.38%	48.94%	5.8	19.19%
ERC	20.00%	42.00%	8.5	2.91%
MD	21.29%	39.07%	5.8	0.00%
MVO	51.81%	100.00%	2.7	73.26%
BL-H	34.61%	76.69%	4.2	35.47%
BL-M	35.76%	73.95%	4.1	35.47%
Sleeve-MVO-U	55.67%	100.00%	2.4	74.42%
Sleeve-MVO-RC	59.05%	100.00%	2.3	76.74%

Note: March 2012–June 2026, 172 holding months. Weights are saved month-end targets.

Effective sleeves equal $1/\sum_g x_{g,t}^2$. The final column is the share of months in which the largest sleeve exceeds 40%. *Source:* Author’s calculations from the frozen common-period target-weight and universe-classification contracts.

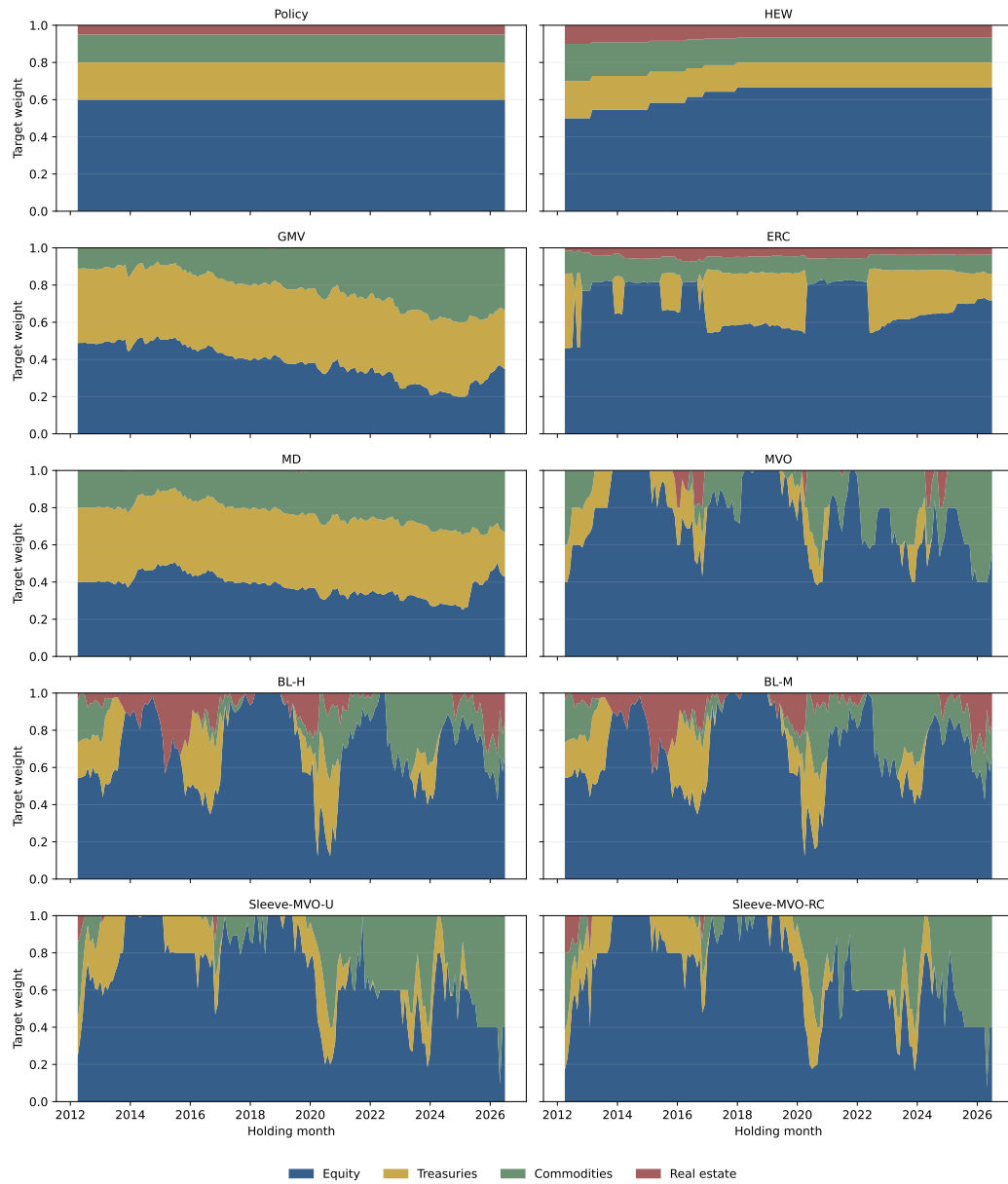


Figure 3.5: Common-period target weights in Equity, Treasuries, Commodities, and Real estate for the ten core strategies, March 2012–June 2026. Each panel sums to one at every holding month.

Source: Author’s calculations from the frozen common-period target-weight and universe-classification contracts.

3.3.5 Assessment of Hypothesis 2

Hypothesis 2. “Macro-conditioned allocation strategies achieve higher out-of-sample risk-adjusted performance than otherwise comparable strategies based only on historical returns or unconditional return moments.”

Under the explicit baseline assessment rule, Hypothesis 2 receives full descriptive support because both matched baseline Sharpe differences are positive. This formal classification is economically qualified. The Black–Litterman increment is small and reverses under the 48-month expected-return specification, whereas the regime-conditioned sleeve increment is larger and remains non-negative across the reported variants. The baseline evidence therefore supports both predicted directions, but the substantive and robustness evidence is concentrated in the sleeve-level comparison.

The stationary-bootstrap analysis further qualifies this result; the complete H2 inference is reported in Section 3.7. The 95% interval for BL-M minus BL-H is $[-0.020, 0.065]$, with a Holm-adjusted p -value of 0.206. The interval for Sleeve-MVO-RC minus Sleeve-MVO-U remains positive at $[0.006, 0.121]$, and 98.88% of pseudo-histories retain the positive direction. Its raw one-sided p -value of 0.038 would reject the comparison-specific null before adjustment, but its H2 Holm-adjusted value of 0.076 does not reject at the 5% familywise level. Thus, the sleeve comparison is more stable than the Black–Litterman increment, although neither H2 directional null is rejected after Holm correction.

This inferential qualification does not alter H2’s full descriptive support under the frozen sign-based rule. Non-rejection does not establish equality, and the conditional analysis does not imply causal or population-level evidence.

3.4 Hypothesis 3: Retail-Investor Implementation

Where H1 and H2 evaluate frictionless strategy returns, H3 applies the common-period target paths under regular contributions and proportional transaction costs. Hypothesis 3 examines whether previously favourable frictionless comparisons remain positive in net time-weighted return and terminal wealth after regular contributions and proportional transaction costs are incorporated. The evidence covers the common period from 31 March 2012 to 30 June 2026, comprising 172 holding months.

The normalized target compositions are the same as those examined in Table 3.7 and Figure 3.5. Contributions and transaction costs change invested wealth and trading intensity in H3, but they do not change the saved target weight proportions.

3.4.1 Contributions, Trading Costs, and Net Wealth

Each strategy begins with 20 contribution units and receives 172 monthly contributions of one unit at the beginning of each month, so total contributed capital equals 192 units. Table 3.8 reports the resulting contribution-based outcomes for all ten core target-weight strategies.

Terminal net wealth is the final portfolio value after trading costs, whereas net investment gain subtracts the 192 units of contributed capital. Net time-weighted return (TWR) measures the compounded investment path while treating contributions as external flows, whereas money-weighted return (MWR) reflects the specified timing of initial capital and monthly contributions.

MVO produced the highest terminal net wealth, at 421.963 units. ERC followed closely at 421.342 units, while Regime-conditioned sleeve MVO (Sleeve-MVO-RC) reached 419.592 units. Their MWR ordering matched terminal wealth, whereas the TWR ordering differed: ERC ranked first, followed by Sleeve-MVO-RC and MVO. Thus, the small terminal-wealth lead of MVO did not imply the highest return on the underlying net investment path.

The direction of both matched H2 advantages survived the contribution-based net implementation. Historical-view Black–Litterman (BL-H) trailed Macro-view Black–Litterman (BL-M) by 7.266 units, while Unconditional sleeve MVO (Sleeve-MVO-U) trailed Sleeve-MVO-RC by 33.371 units. Net TWR and MWR preserved the same ordering in both pairs. The evidence therefore indicates economically meaningful survival for the regime-conditioned sleeve

Table 3.8: Retail implementation outcomes under monthly contributions and five-basis-point trading costs

<i>Panel A: wealth and costs</i>					
Strategy	Gross wealth	Net wealth	Net gain	Direct cost	Cost drag
MVO	430.926	421.963	229.963	4.716	8.963
ERC	422.912	421.342	229.342	0.818	1.570
Sleeve-MVO-RC	426.943	419.592	227.592	4.328	7.351
HEW	394.527	393.718	201.718	0.462	0.809
Sleeve-MVO-U	392.803	386.221	194.221	4.049	6.583
Policy	382.253	381.464	189.464	0.463	0.789
BL-M	327.677	319.459	127.459	5.201	8.218
MD	318.196	316.497	124.497	1.213	1.699
BL-H	320.002	312.193	120.193	5.024	7.808
GMV	307.525	305.868	113.868	1.209	1.656

<i>Panel B: returns, trading intensity, and Random DCA context</i>					
Strategy	Net TWR	MWR	Traded-value ratio	RDA percentile	RDA exceedance
MVO	8.16%	9.34%	0.381	49.80%	50.25%
ERC	8.52%	9.32%	0.058	48.70%	51.35%
Sleeve-MVO-RC	8.50%	9.27%	0.323	46.40%	53.65%
HEW	7.30%	8.55%	0.036	13.40%	86.61%
Sleeve-MVO-U	7.61%	8.33%	0.320	8.40%	91.61%
Policy	7.13%	8.19%	0.036	5.50%	94.51%
BL-M	4.62%	6.14%	0.498	0.00%	100.00%
MD	5.70%	6.03%	0.099	0.00%	100.00%
BL-H	4.39%	5.87%	0.488	0.00%	100.00%
GMV	5.77%	5.64%	0.099	0.00%	100.00%

Note: March 2012–June 2026. Total contributed capital is 192 units. RDA measures are conditional on the fixed 1,000-path simulation design; exceedance frequencies are plus-one corrected. *Source:* Author's calculations from the frozen H3 contracts.

advantage and a more modest survival for the macro-view Black–Litterman advantage; it does not establish practical dominance for macro-conditioned allocation generally.

The favourable H1 comparisons did not carry into H3 uniformly. ERC remained above Policy and HEW in terminal net wealth and also retained higher net TWR and MWR. GMV and MD, however, ended below both benchmarks, so their favourable standalone Sharpe comparisons did not translate into favourable contribution-based terminal outcomes. MVO’s high net wealth likewise should not be described as survival of a favourable H1 primary result because its standalone Sharpe comparison had been adverse.

3.4.2 Gross-to-Net Performance Differences

The gross-to-net comparison in Table 3.8 separates cumulative direct transaction cost from terminal cost drag. Direct transaction cost is the cash paid on executed purchases and sales, whereas terminal cost drag is the difference between gross and net terminal wealth and includes investment growth forgone because costs were paid earlier. The table also reports the actual traded-value ratio, an implementation-intensity measure distinct from the one-way drift-adjusted turnover used in the frictionless H1 and H2 analysis.

Trading costs materially narrowed the differences among the three strongest strategies. Gross terminal wealth was 430.926 units for MVO, 426.943 units for Sleeve-MVO-RC, and 422.912 units for ERC. After costs, ERC moved ahead of Sleeve-MVO-RC and finished only 0.620 units below MVO.

ERC incurred cumulative direct cost of 0.818 units and terminal cost drag of 1.570 units, compared with 4.716 and 8.963 units for MVO. Their actual traded-value ratios were 0.058 and 0.381, respectively. ERC therefore nearly matched MVO’s terminal net wealth with much lower realized trading intensity and cost drag.

The same trade-off is visible among the more active H2 strategies. Sleeve-MVO-RC was also more implementation-intensive than the simple strategies, while BL-H and BL-M recorded the highest actual traded-value ratios, at 0.488 and 0.498. By contrast, Policy, HEW, and ERC had actual traded-value ratios no greater than 0.058 and much smaller direct costs. These results show that the five-basis-point charge was not economically uniform across strategies: target paths requiring more traded value were penalized more heavily, while terminal cost drag additionally depended on the timing of those costs

and subsequent returns. The net outcomes establish the economic comparison; the Random DCA distribution provides a separate conditional distributional comparison for implemented terminal outcomes.

3.4.3 Comparison with the Random DCA Distribution

Figure 3.6 locates each target strategy within the 1,000-path Random DCA terminal-net-wealth distribution. The simulated median was 422.137 units and the mean was 424.212 units. MVO, ERC, and Sleeve-MVO-RC occupied the 49.8th, 48.7th, and 46.4th percentiles, respectively. Their plus-one upper-tail conditional exceedance frequencies were 50.25%, 51.35%, and 53.65%. All three therefore lay near the simulated median and below the simulated mean, so the comparison does not show that they were unusual within the fixed simulation design.

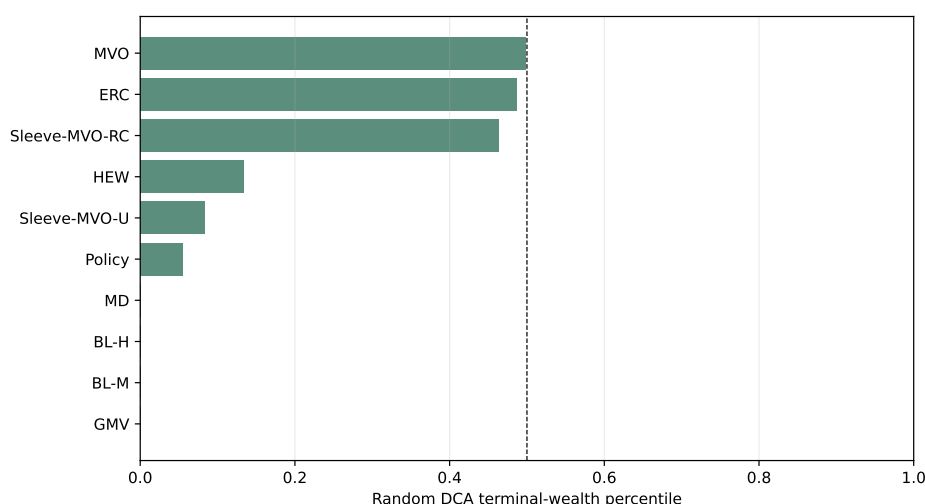


Figure 3.6: Terminal net wealth percentiles of the ten target-weight strategies within the conditional 1,000-path Random DCA distribution, March 2012–June 2026. The Random DCA mean and median terminal wealth are 424.212 and 422.137 contribution units; its 5th and 95th percentiles are 378.906 and 473.780. Higher strategy percentiles indicate more favourable terminal wealth within the fixed simulation design.

Source: Author's calculations from the frozen H3 and Random DCA contracts.

The remaining strategies were less competitive within the fixed distribution. HEW, Sleeve-MVO-U, and Policy occupied the 13.4th, 8.4th, and 5.5th percentiles, respectively. GMV, MD, BL-H, and BL-M finished below all 1,000

simulated paths on terminal net wealth. The zero empirical percentiles for these four strategies describe their position within a finite conditional simulation; they do not imply an unconditional probability of zero or describe all possible retail-investor outcomes.

This evidence is conditional on one realized market history, the growing eligible ETF universe, the ETF-level sampling rule, the six-month review schedule, the contribution path, and the five-basis-point cost assumption. It therefore describes positions within a fixed simulation design rather than a classical population-level test. The maintenance rules also differ. The target-weight strategies pursue monthly allocations, whereas Random DCA selects and equal-weights a subset only at semiannual reviews and directs interim contributions to the existing subset without selling merely to restore equal weights. The comparison consequently reflects both allocation outcomes and differing review intensity; it does not isolate security selection alone.

3.4.4 Assessment of Hypothesis 3

Hypothesis 3. “Strategy-specific performance advantages observed in the frictionless comparisons remain positive in net time-weighted return and terminal wealth after monthly contributions and proportional transaction costs are incorporated.”

H3 assesses five favourable frictionless comparisons: GMV, ERC, and MD against both Policy and HEW, together with BL-M versus BL-H and Sleeve-MVO-RC versus Sleeve-MVO-U. ERC and both matched H2 comparisons retained higher net time-weighted return and terminal wealth, whereas GMV and MD did not remain above both simple benchmarks. Three of the five favourable comparisons therefore survived contributions and transaction costs.

Hypothesis 3 therefore receives partial support. ERC and both matched H2 comparisons retain their favourable directions in net time-weighted return and terminal wealth, whereas GMV and MD do not. This formal classification remains compatible with the broader conclusion that H3 does not support a general practical-dominance claim. The supplementary Random DCA comparison shows that even the strongest implemented outcomes were not unusual within the realized history and fixed simulation design; it does not enter the support classification.

3.5 External Market Benchmark: Vanguard VT

Vanguard Total World Stock ETF (VT) provides an external global-equity market reference for the core strategy results (The Vanguard Group 2026). It represents a single-fund, all-equity opportunity set rather than a strategy formed on the same growing ETF universe and under the same allocation design as the core models. The comparisons below are therefore descriptive market context, not matched benchmark evidence, and VT is not used to assess Hypotheses 1–3.

Over the standalone H1 period from 31 December 2009 to 30 June 2026, VT’s annualized arithmetic return and final frictionless wealth were higher than those of the six core H1 strategies. As reported in Table 3.9, its Sharpe ratio remained below those of GMV, MD, and ERC, while its drawdown was more negative than that of every core H1 strategy. The standalone paths in Figure 3.7 place the higher terminal accumulation alongside the larger equity drawdown rather than treating either feature as a general performance ranking.

The common-period results are reported separately to avoid mixing evaluation windows. From 31 March 2012 to 30 June 2026, VT’s Sharpe ratio and final frictionless wealth were higher than those of all ten core strategies, as shown in Table 3.9 and Figure 3.7. This comparison does not make VT a matched H2 comparator: its continuously all-equity exposure and single-fund construction differ from the opportunity sets and allocation rules of the core strategies.

Table 3.9: External market context from Vanguard Total World Stock ETF

<i>Panel A: frictionless results</i>						
Period	Return	Volatility	Sharpe	Drawdown	Wealth	
Standalone H1	11.19%	14.80%	0.659	-25.51%	5.298	
Common period	11.53%	13.95%	0.708	-25.51%	4.514	

<i>Panel B: common-period contribution-based results</i>						
Gross wealth	Net wealth	Direct cost	Cost drag	Net TWR	MWR	RDA percentile
527.921	527.657	0.096	0.264	11.08%	11.84%	99.90%

Note: VT is an external global-equity reference and does not enter formal H1–H3 assessments. Wrapper-level concentration does not describe the fund’s underlying security diversification. *Source:* Author’s calculations from the frozen VT external-reference contracts.

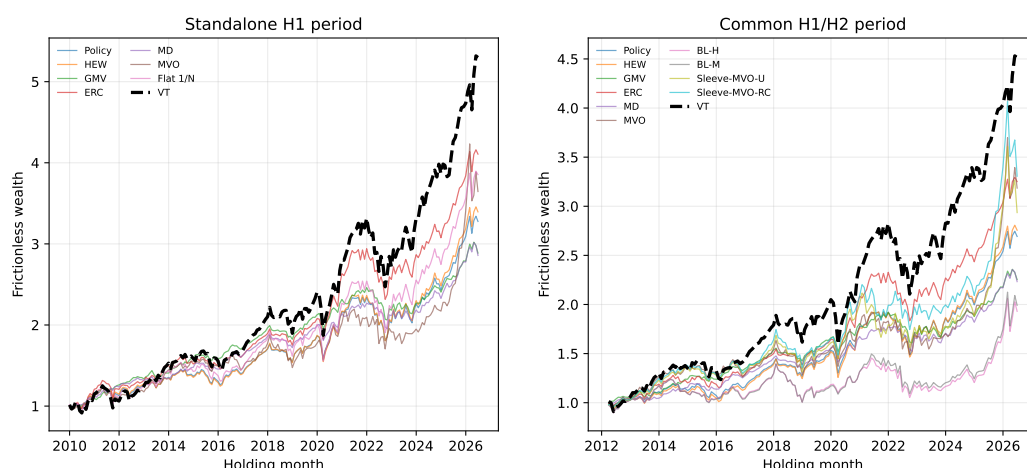


Figure 3.7: VT and the formal strategy wealth paths over the standalone H1 and common H1/H2 periods. VT is distinguished as an external all-equity reference and does not enter the formal hypothesis comparisons.

Source: Author's calculations from the frozen VT and core frictionless wealth contracts.

Under the common H3 contribution schedule and purchase-cost assumption, VT's terminal net contribution-based wealth was slightly below its gross contribution-based wealth. Terminal cost drag exceeded cumulative direct transaction cost because it also reflects the subsequent compounding forgone when costs were paid. Table 3.9 reports the complete accounting measures, while Figure 3.10 shows the VT net-wealth path alongside the core strategy paths.

VT is broadly diversified across thousands of underlying companies, but it is not a balanced multi-asset benchmark. As shown in Figure 3.8, North America represented 64.9% of the fund's regional equity exposure, compared with 13.7% for Europe, 11.1% for the Pacific region, and 10.0% for emerging markets. The concentration is also visible among its largest positions. The ten securities in Figure 3.9 represented 20.42% of net assets as of 30 June 2026 and were dominated by large technology and technology-adjacent platform companies.¹

¹The reported regional and security composition is an end-of-sample snapshot that illustrates VT's concentration near the end of the evaluation period. It is not a historical holdings attribution for the complete 2012–2026 period, because the fund's regional and security weights changed over time

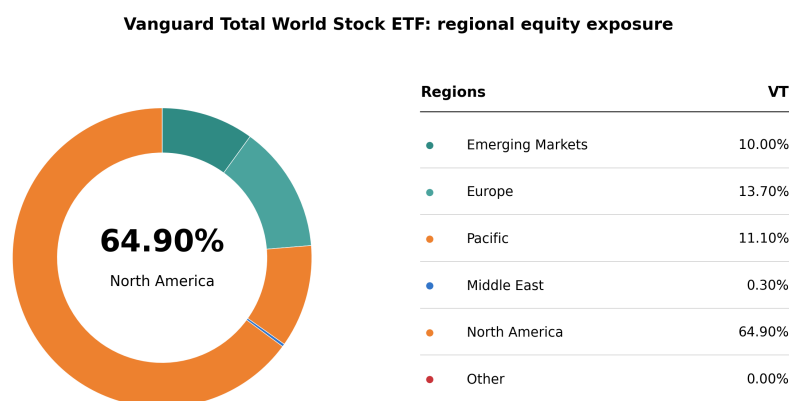


Figure 3.8: Regional composition of Vanguard Total World Stock ETF. North America accounts for 64.9% of the fund's equity exposure; the remaining exposure is distributed mainly across Europe, the Pacific region, and emerging markets.

Source: Author's adaptation of the Vanguard VT portfolio-composition page, Regions panel, accessed 25 July 2026.

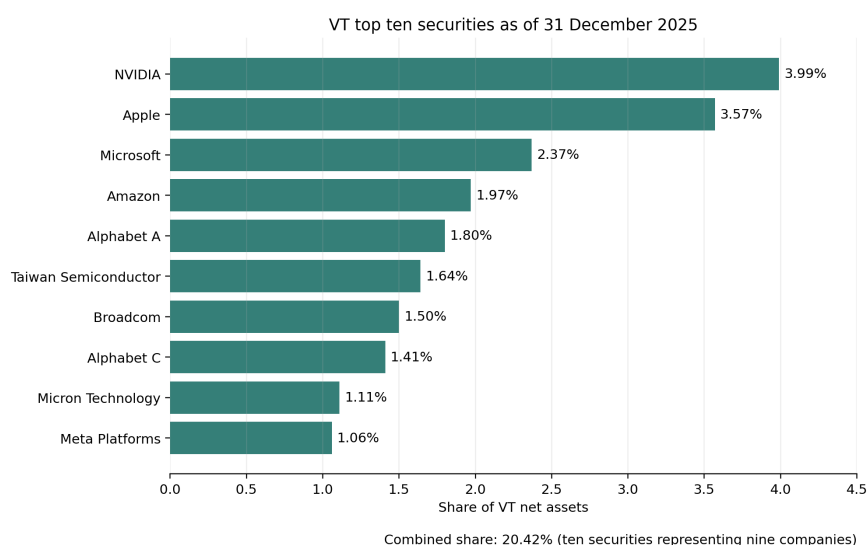


Figure 3.9: Ten largest VT securities as of 30 June 2026. Together they represented 20.42% of net assets and were concentrated in large technology and technology-adjacent platform companies. Alphabet appears through two share classes.

Source: Author's adaptation of the Vanguard VT portfolio-composition page, Holding details panel, as of 30 June 2026.

Moreover, VT also holds equities only, whereas the thesis strategies can allocate to Treasuries, commodities, and listed real estate. Its realized per-

formance therefore reflects a different risk budget from that of the multi-asset strategies. The end-of-sample snapshot shows substantial North American and mega-cap exposure near the sample endpoint, but it does not establish that the same weights prevailed throughout the evaluation period or quantify their contribution to VT's realized return. VT's stronger common-period return, Sharpe ratio, and terminal wealth should therefore be interpreted as evidence that a broad all-equity fund was a demanding practical reference in this realized sample, not as evidence that its end-of-sample holdings caused its historical outperformance.

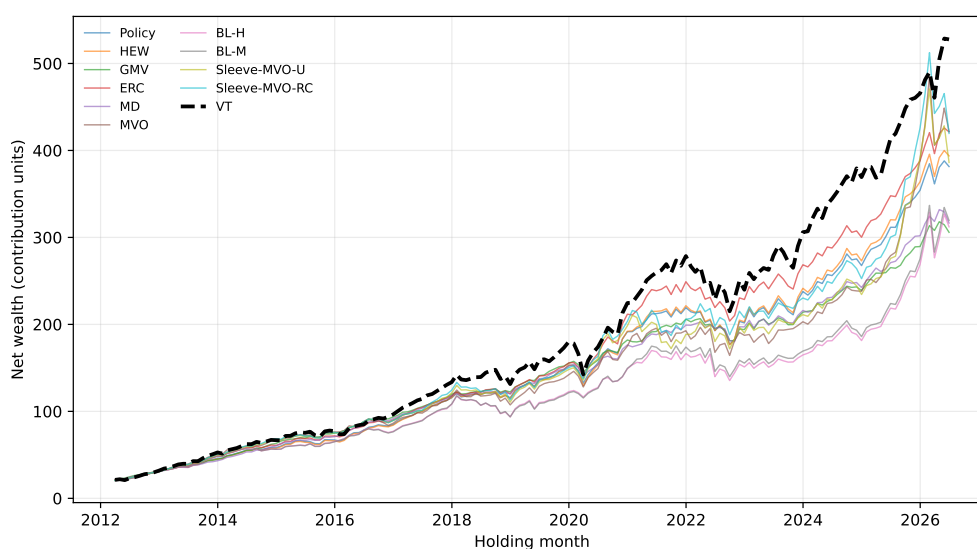


Figure 3.10: VT and the ten formal strategy net contribution-based wealth paths, March 2012–June 2026. All paths use the common initial wealth, beginning-of-month contributions, fractional shares, and proportional trading-cost assumption.

Source: Author's calculations from the frozen VT and H3 wealth contracts.

Finally, Figure 3.11 places VT's terminal net wealth at the 99.9th percentile of the conditional Random DCA distribution. As qualified in the H3 conditional-design discussion in Section 3.4, this position is specific to the fixed Random DCA experiment and is neither a classical population-level significance result nor an unconditional probability. VT and Random DCA also differ in their investment opportunity sets and maintenance rules. The result consequently remains external market context and does not enter any hypothesis assessment.

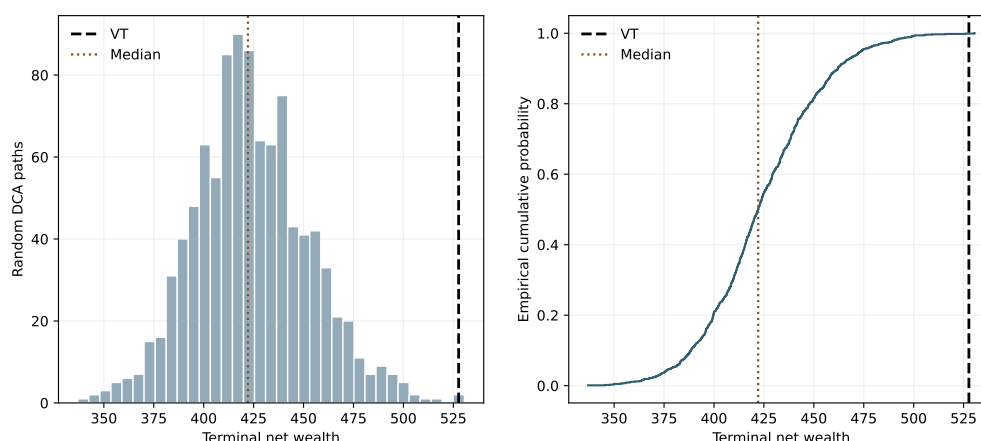


Figure 3.11: VT within the conditional Random DCA terminal-net-wealth distribution. The dashed line identifies VT and the dotted line identifies the Random DCA median. The comparison is conditional on the realized market history and the fixed Random DCA design.

Source: Author's calculations from the frozen VT and Random DCA contracts.

Overall, in this realized sample, added allocation complexity did not deliver a compelling general advantage over a single broad all-equity fund. This practical comparison remains qualified because VT operated under a different risk budget and experienced larger drawdowns than the defensive covariance-based strategies.

3.6 Robustness Analysis

The H2 matched table reports the two pre-specified expected-return-window robustness specifications and the separately labelled financial-only HMM diagnostic alongside the baseline. The principal conclusion is architecture-dependent. The regime-conditioned sleeve comparison remains non-negative across all reported variants, although its 12-month Sharpe increment is effectively zero. The macro-view Black–Litterman increment is smaller and changes sign under the 48-month window. The flat ETF-level $1/N$ result in Table 3.2 supplies the secondary H1 benchmark check; its ticker-count sensitivity prevents it from replacing Policy and HEW as the formal comparators.

The alternative HMM state-count information criteria are descriptive only. They decline through five states in the final-date diagnostic, but the baseline three-state model was fixed before portfolio performance was evaluated.

The diagnostic therefore indicates that three states may compress finer macro-financial distinctions; it does not authorize ex post state-count selection.

The transaction-cost sensitivity in Table E.5 replays the ten frozen target-weight paths at 5, 25, and 50 basis points without re-estimating H1 or H2 and without rerunning Random DCA. Higher costs penalize the active models most strongly: relative to 5 basis points, terminal net wealth at 50 basis points falls by 70.881 units for MVO, 63.251 for BL-M, 60.354 for BL-H, and 59.749 for Sleeve-MVO-RC, compared with 13.772 for ERC, 7.181 for HEW, and 7.005 for Policy. ERC becomes the highest-terminal-wealth strategy at both 25 and 50 basis points, strengthening its relative implementation advantage. Nevertheless, ERC remains above Policy and HEW, BL-M remains above BL-H, and Sleeve-MVO-RC remains above Sleeve-MVO-U in both net time-weighted return and terminal wealth at every tested rate. The explicit H3 classification therefore remains partial support: the same three of five favourable frictionless comparisons survive, while GMV and MD still do not remain above both simple benchmarks.

3.7 Stationary-Bootstrap Inference

The stationary-bootstrap results in Table 3.10 and Figure 3.12 provide a conditional inferential layer for the frozen H1 and H2 return paths. Under the primary 12-month expected block length, only ERC rejects its H1 intersection-union null after Holm adjustment. GMV and MD retain positive observed differences but have pairwise intervals crossing zero. Within H2, the BL-M increment is imprecise, whereas the Sleeve-MVO-RC interval remains positive; neither H2 comparison rejects after the two-test Holm adjustment. The descriptive classifications remain unchanged because they are determined by the explicit sign-based rules rather than by ex post inferential thresholds.

Taken together, Table 3.10 shows that only ERC rejects its directional null after familywise correction. Within H2, BL-M provides no precise inferential evidence, while Sleeve-MVO-RC is favourable under its unadjusted test but falls short after Holm correction. This does not contradict H2's full descriptive support: the frozen assessment rule classifies H2 from the signs of the two realized matched Sharpe differences, whereas the bootstrap separately evaluates their sampling uncertainty.

Table 3.10: Stationary-bootstrap qualification of H1 and H2 Sharpe comparisons

Family	Comparison	$\hat{\Delta}$	Raw p	Holm p	5% decision
H1	GMV vs. both benchmarks	+0.199	0.189	0.567	Do not reject
H1	ERC vs. both benchmarks	+0.127	0.008	0.032	Reject
H1	MD vs. both benchmarks	+0.128	0.220	0.567	Do not reject
H1	MVO vs. both benchmarks	-0.041	0.623	0.623	Do not reject
H2	BL-M – BL-H	+0.016	0.206	0.206	Do not reject
H2	Sleeve-MVO-RC – U	+0.058	0.038	0.076	Do not reject

Note: Primary expected block length of 12 months and 10,000 replications. For H1, the observed difference is the smaller benchmark increment and the raw p -value is the larger of the two pairwise values. Holm adjustment is applied separately within the four H1 and two H2 comparisons. Decisions qualify but do not redefine the descriptive support classifications. *Source:* Author's calculations from the frozen H1 and H2 strategy-return paths.

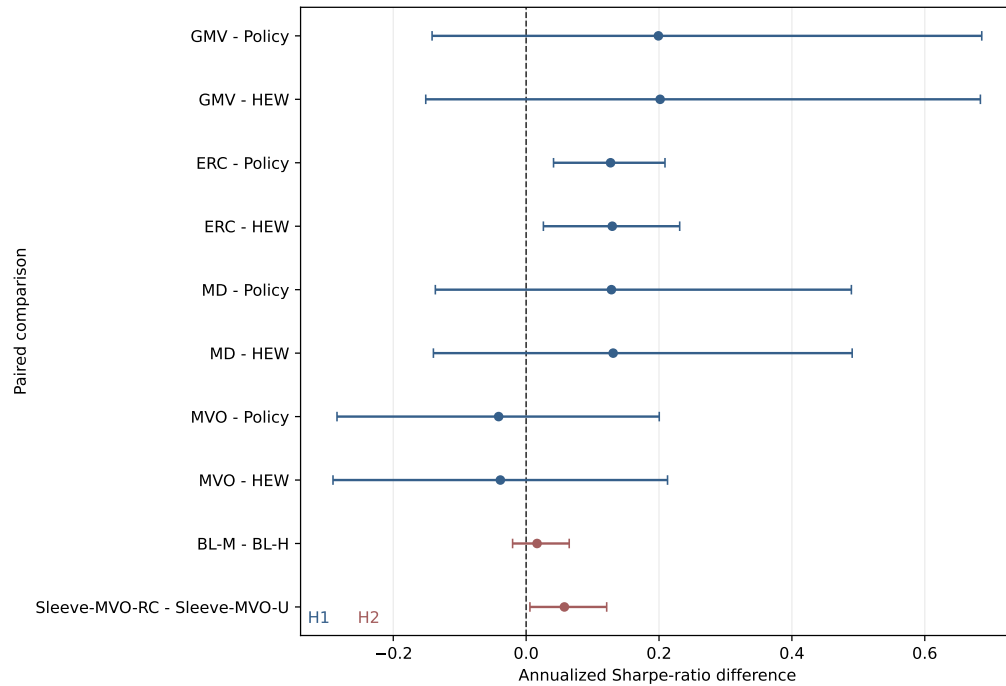


Figure 3.12: Observed paired Sharpe-ratio differences and 95% stationary-bootstrap percentile intervals under the primary 12-month expected block length. Blue rows belong to H1 and red rows to H2; intervals crossing zero indicate that the ranking reverses in some plausible pseudo-histories.

Source: Author's calculations from 10,000 joint stationary-bootstrap replications of the frozen H1 and H2 strategy-return paths.

The detailed pairwise results and block-length sensitivity are reported in Table E.3 and Table E.4. The qualitative interval pattern is unchanged for

expected block lengths of 6, 12, and 24 months: ERC remains positive against both benchmarks, GMV and MD continue to cross zero, BL-M continues to cross zero, and Sleeve-MVO-RC remains positive relative to Sleeve-MVO-U.

These results quantify sampling uncertainty conditional on the fitted return paths. They do not re-estimate the underlying portfolio models, alter the sign-based H1 or H2 classifications, or provide a test of H3 terminal wealth. The distinction from Random DCA follows Subsection 2.5.5: Random DCA varies ETF selection under one realized history, whereas the stationary bootstrap varies the representation of observed market episodes for fixed strategies.

3.8 Summary of Empirical Findings

H1 receives partial support. GMV, ERC, and MD improve standalone Sharpe ratios relative to Policy and HEW, but their economic implications differ. GMV and MD obtain their advantages mainly through lower risk and lower terminal wealth, whereas ERC combines an improved Sharpe ratio with the highest core H1 final wealth. ETF-level MVO does not improve Sharpe and entails substantially greater turnover and concentration. Unconditional allocation therefore adds value selectively rather than uniformly.

H2 receives full descriptive support because both baseline matched Sharpe differences are positive. The economic evidence is nevertheless heterogeneous and concentrated in the regime-conditioned sleeve architecture. Sleeve-MVO-RC improves baseline Sharpe, drawdown, and wealth relative to Sleeve-MVO-U and retains a non-negative Sharpe increment across the reported variants. The BL-M improvement over BL-H is modest, reverses under the 48-month window, and remains weak in the broader ten-strategy comparison.

H3 receives partial support because ERC and both matched H2 comparisons retain higher net time-weighted return and terminal wealth, whereas GMV and MD do not remain above both simple benchmarks. The formal classification does not imply general practical dominance. The strongest target-weight strategies remain near the center of the conditional Random DCA distribution, while four sophisticated strategies finish below all 1,000 simulated paths on terminal wealth.

VT provides an important but non-matched qualification. The single global-equity fund produces higher common-period Sharpe and wealth than the formal strategies and lies at the 99.9th percentile of the Random DCA distribution. Its continuously all-equity opportunity set also produces a larger drawdown than

the risk-focused H1 portfolios, and its wrapper-level simplicity is not directly comparable with the multi-asset allocation problem. Taken together, the evidence suggests that complex allocation can improve selected dimensions, especially risk control and regime-conditioned sleeve allocation, but simple transparent alternatives remain highly competitive and implementation demands materially affect practical value.

Chapter 4

Discussion

4.1 Research Question and Empirical Contribution

This thesis examined whether increasingly sophisticated portfolio-allocation methods, and macro-financial conditioning in particular, add practical value relative to simple ETF allocation rules for a constrained retail investor. The question was evaluated in a growing universe of 71 exchange-traded funds representing international and country equity markets, sectors, factors, U.S. Treasuries, commodities, and listed real estate. The analysis imposed a common monthly information boundary, long-only and fully invested portfolios, a 20% ETF cap, common eligibility rules, and a frozen data vintage ending in June 2026. These controls were intended to ensure that differences in realized performance reflected allocation rules rather than unequal access to assets, later information, or different constraints.

The empirical design separated three questions that are often combined. First, it tested whether unconditional classical and risk-based strategies improved out-of-sample risk-adjusted performance relative to transparent Policy and hierarchical equal-weight benchmarks. Second, it isolated the incremental contribution of macro-financial information through matched Black–Litterman and sleeve-level mean–variance comparisons. Third, it applied the resulting target paths under regular contributions and proportional transaction costs and compared their terminal outcomes with a conditional distribution of 1,000 Random DCA paths. This sequence distinguished statistical allocation performance from practical implementation and prevented a favourable result in one dimension from being treated as general strategy dominance.

The principal conclusion is that sophistication added value selectively, but

not uniformly. Covariance-based and risk-based allocation improved particular risk-adjusted outcomes, and macro-financial conditioning was useful within the regime-conditioned sleeve architecture. Nevertheless, simple transparent allocations remained competitive, expected-return-driven models were fragile, and the strongest implemented strategies did not occupy the upper tail of the Random DCA distribution. The evidence therefore rejects a single ranking of methods that applies across risk, wealth, trading intensity, and implementation complexity.

4.2 Evaluation of the Hypotheses

4.2.1 Unconditional Allocation

Hypothesis 1 proposed that classical and risk-based allocation strategies would achieve higher out-of-sample risk-adjusted performance than simple benchmarks on the same growing ETF universe. The standalone evaluation covered 199 holding months from December 2009 through June 2026. The hypothesis received partial support.

Global minimum variance, equal risk contribution, and maximum diversification each produced a higher Sharpe ratio than both Policy and hierarchical equal weight. GMV recorded the highest standalone Sharpe ratio, at 0.759, while MD produced the smallest maximum drawdown, at -14.18% . These outcomes show that covariance information can improve risk control even when it does not increase return or final wealth. ERC provided the strongest balance among the unconditional methods: its Sharpe ratio was 0.687 and its final frictionless wealth of 4.107 units was the highest among the six core H1 strategies. It also remained broadly diversified rather than obtaining its result through a small set of capped positions.

The favourable evidence did not extend to ETF-level mean–variance optimization. MVO generated a relatively high return and final wealth, but its volatility, concentration, and annualized turnover were sufficiently high that its Sharpe ratio of 0.519 remained below both core benchmarks. GMV and MD presented the opposite trade-off: their low risk raised Sharpe ratios but constrained wealth accumulation. H1 therefore demonstrates that the meaning of “better performance” depends on the investor’s objective. Risk reduction, terminal accumulation, diversification, and trading stability did not identify the same preferred method.

4.2.2 Macroeconomic and Financial Information

Hypothesis 2 proposed that macro-conditioned strategies would achieve higher out-of-sample risk-adjusted performance than otherwise comparable strategies based only on historical returns or unconditional moments. The matched evaluation covered 172 holding months from March 2012 through June 2026. Under the explicit baseline rule, the hypothesis received full descriptive support because both matched Sharpe differences were positive. Its economic interpretation remains qualified because the evidence was concentrated in the sleeve-level allocation model.

The three-state hidden Markov model produced economically interpretable and persistent favourable, intermediate, and adverse states without numerical fallback. Regime-conditioned sleeve MVO improved the baseline Sharpe ratio from 0.490 for the unconditional sleeve model to 0.548. It also increased return and final wealth and reduced maximum drawdown. The Sharpe increment remained non-negative in the reported variants, although it was effectively zero with a 12-month expected-return window. Macro-financial conditioning therefore added value when it affected both predictive sleeve means and covariances within a lower-dimensional economic allocation problem, but the magnitude was not uniform across specifications.

The Black–Litterman comparison was substantially weaker. BL-M improved the baseline Sharpe ratio only from 0.282 to 0.298 relative to BL-H, and the advantage reversed under the 48-month expected-return window. Both variants ranked at the bottom of the common-period strategy comparison. The appendix analysis shows why this result need not contradict favourable Black–Litterman evidence in earlier research. Historical and macro sleeve views were highly similar and had little next-month predictive correlation, while the ETF-level optimizer translated modest posterior-return changes into large departures from the Policy portfolio. Overlapping sleeve errors, a diagonal uncertainty matrix, high effective concentration, and annualized turnover near 3.0 further limited performance. The 12-month robustness specification materially improved both Black–Litterman Sharpe ratios to approximately 0.50, but this sensitivity reinforces rather than resolves the dependence on expected-return estimation. It cannot justify replacing the pre-specified 24-month baseline after observing the outcomes.

The combined H2 evidence has an important interpretation. An economically coherent state model is not sufficient by itself. Additional information

creates portfolio value only when the allocation architecture uses it in a stable way and when conditional estimates improve the relevant investment inputs. In this study, the regime signal was more useful for lower-dimensional sleeve means and covariances than for absolute Black–Litterman views mapped into a large, overlapping ETF universe.

4.2.3 Retail Implementation

Hypothesis 3 proposed that previously favourable frictionless comparisons would remain positive in net time-weighted return and terminal wealth after regular contributions and proportional transaction costs. Three of five favourable comparisons survived, so the hypothesis received partial support. This formal classification does not support a general practical-dominance claim.

The implemented strategies began with 20 contribution units, received one unit at the beginning of each of 172 holding months, and paid five basis points on absolute traded value. Selected advantages survived this accounting change. ERC retained favourable net return and wealth relative to the simple benchmarks while incurring low trading costs. The direction of both matched H2 comparisons also survived: BL-M finished above BL-H, and regime-conditioned sleeve MVO finished above its unconditional control. These results show that contributions and costs did not mechanically eliminate every benefit identified in the frictionless analysis.

Implementation intensity nevertheless changed the economic ranking. MVO produced the highest terminal net wealth, at 421.963 units, but ERC followed at 421.342 units while incurring only 0.818 units of direct transaction cost, compared with 4.716 units for MVO. ERC also achieved the highest net time-weighted return among the three leading outcomes. The difference illustrates why gross performance alone is insufficient for a retail investor: two strategies can finish with almost identical wealth while requiring very different trading intensity, cost, and monitoring effort.

The conditional Random DCA comparison provided supplementary distributional context for H3. MVO, ERC, and regime-conditioned sleeve MVO occupied only the 49.8th, 48.7th, and 46.4th percentiles of terminal net wealth, respectively. They were near the simulated median and below the simulated mean rather than in the upper tail. GMV, MD, BL-H, and BL-M finished below all 1,000 simulated paths. These percentiles apply only to the realized market history, growing universe, contribution schedule, and Random DCA

rules used in the experiment. They are not population probabilities. Within that fixed design, they do not show that the strongest sophisticated outcomes were unusual relative to feasible simple investment paths.

4.3 Implications for a Retail Investor

The results do not identify a universally optimal ETF strategy. They instead suggest several decision principles. First, an investor whose primary objective is risk control may benefit from covariance-based methods such as GMV or maximum diversification, but should not infer that a higher Sharpe ratio guarantees greater terminal wealth. Second, ERC offered the most balanced outcome in this study. It combined broad diversification, favourable standalone and common-period Sharpe ratios, strong contribution-based wealth, and low implementation cost. This does not establish universal ERC superiority, but it shows that a risk-based rule can preserve much of the value of optimization without relying on unstable expected-return forecasts.

Third, expected-return-driven allocation requires particular caution. ETF-level MVO and both Black–Litterman variants reacted strongly to uncertain return inputs and produced high concentration or turnover. Regularization, long-only constraints, and weight caps did not fully remove this sensitivity. A technically sophisticated forecast is valuable only if it predicts relevant return differences and if the optimizer converts those differences into stable exposures. Model interpretability alone does not satisfy either condition.

Fourth, macro-financial information appears more useful as a structured risk and allocation input than as a direct source of ETF return views. The strongest H2 result allowed estimated regimes to affect sleeve-level expected returns and covariance, while the weaker Black–Litterman result changed expected returns alone. For practical portfolio design, this favours parsimonious economic groupings and explicit control of estimation error over high-dimensional tactical tilts.

Finally, simplicity remained economically important. Policy and hierarchical equal-weight portfolios required little turnover and were transparent, while VT showed that a single global-equity fund could achieve strong realized return and wealth over the common period. VT also carried a different risk profile: it was continuously invested in equities, had a large North American allocation and substantial exposure to major technology-related companies, and experienced larger drawdowns than the risk-focused H1 methods. It is therefore not a matched substitute for the multi-asset strategies. Its strong outcome neverthe-

less reinforces the central practical standard of the thesis: additional portfolio complexity must deliver benefits large enough to justify its estimation, trading, and governance demands.

4.4 Limitations

The findings are conditional on one realized market history and one frozen empirical design. The growing universe contains surviving ETFs that could be identified with sufficiently long data histories; it does not reconstruct every fund that was liquidated before universe selection. Fund-level survivorship bias may therefore remain. The selected U.S. and European listings also represent a hypothetical international investor with access to both markets. Taxes, product-distribution restrictions, account-specific brokerage permissions, and investor-specific currency-conversion charges are outside the analysis.

Although each portfolio decision respects the contemporaneous information set, the final research architecture was developed with knowledge of the historical evaluation sample. The reported performance should therefore be interpreted as recursive historical out-of-sample evidence rather than as evidence from a prospectively untouched holdout.

The return data use adjusted closes as total-return proxies and convert non-dollar listings to U.S. dollars. H3 applies one proportional transaction-cost rate, fractional shares, and no terminal liquidation. It excludes time-varying bid–ask spreads, market impact, taxes, minimum commissions, custody charges, and integer-share constraints. The five-basis-point assumption provides a common implementation layer rather than a complete tariff for a named broker or jurisdiction.

The macro-financial model is deliberately parsimonious. Three Gaussian states with diagonal emission covariances may compress finer distinctions in the data, as suggested by the descriptive state-count criteria. State labels are economically interpretable but do not establish causal effects of individual macro variables. Expected-return and covariance estimates also remain uncertain, particularly when 15 sleeve views are mapped into a growing universe of overlapping ETFs. Robustness results reveal sensitivity to the expected-return window, especially for Black–Litterman, but the narrow pre-specified variants do not cover every plausible modelling choice.

The Random DCA distribution is a conditional benchmark rather than a universal representation of retail behaviour. Its paths share the realized return

history and eligibility calendar, sample ETFs under a fixed rule, review selections every six months, and use maintenance rules that differ from monthly target-weight strategies. The comparison therefore combines allocation outcomes with review intensity. It cannot establish how the models would rank under alternative economic histories or investor behaviour.

4.5 Future Research and Final Assessment

Future research could first examine whether the central findings reproduce under a new, prospectively frozen data vintage. Repetition across later market conditions would be more informative than adding many ex post parameter combinations to the current sample. A broader validation could also compare different jurisdictions, tax regimes, and listing constraints while preserving the common-universe and information-boundary controls developed here.

A complementary extension could examine descriptive performance stability across fixed, non-overlapping market subperiods, such as the pre-COVID expansion, the COVID shock and recovery, and the subsequent inflation and higher-rate period. This analysis could apply the existing performance measures to the frozen return paths of the ten core strategies and VT without re-estimating any portfolio model. It would show whether the full-sample H1 and H2 rankings were persistent or concentrated in particular market environments, while treating evidence from the comparatively short COVID subperiod with appropriate caution.

The Black–Litterman evidence suggests a second direction. A prospective replication of the lower-dimensional architecture used by Bessler *et al.* (2017) could compare broad, non-overlapping sleeve assets with the present ETF-level implementation. Such a study could evaluate posterior covariance, correlated view uncertainty, and an explicit portfolio-volatility constraint without relabelling the resulting model as a robustness test of the frozen H2 baseline. The objective would be to identify which architectural choices stabilize the transmission from views to weights, not to search retrospectively for a favourable specification.

Further work on practical implementation could incorporate broker-specific minimum commissions, bid–ask spreads, integer shares, taxes, and residual cash at explicitly stated account sizes. Random DCA could also be studied under prospectively defined sampling rules that separate broad economic exposure from ticker-count representation. These extensions would require new account-

ing states and benchmark definitions and should therefore remain distinct from the present baseline.

Overall, the thesis finds no general evidence that greater allocation sophistication produces a superior retail-investor outcome. Unconditional optimization added value selectively; macro-financial conditioning was useful mainly within the regime-conditioned sleeve model; and implementation costs and benchmark choice materially affected practical interpretation. ERC provided the most balanced combination of risk-adjusted performance, wealth, diversification, and low trading intensity among the formal strategies, while simple allocations and the external VT reference remained difficult comparators. The broader lesson is not that optimization or macroeconomic information lacks value. It is that their value depends on stable inputs, disciplined comparison, and benefits sufficiently large to compensate for the complexity they introduce.

Conclusion

This thesis examined whether sophisticated ETF portfolio-allocation methods add practical value relative to simple rules for a constrained retail investor. Unconditional strategies, matched macro-conditioned models, and contribution-based implementations were evaluated in a growing 71-ETF universe under common information, eligibility, constraint, and validation rules. The evidence is interpreted as recursive historical out-of-sample under this strict information boundary. Implemented outcomes were also compared with 1,000 Random DCA paths generated under the same realized history and contribution schedule.

The evidence shows selective rather than general gains from sophistication. Global minimum variance, equal risk contribution, and maximum diversification improved standalone Sharpe ratios relative to Policy and hierarchical equal weight, but through different trade-offs. GMV and maximum diversification mainly reduced risk, whereas ERC combined favourable risk-adjusted performance with strong wealth accumulation and broad diversification. ETF-level mean–variance optimization generated high wealth but remained concentrated and turnover-intensive.

Macro-financial conditioning added value only within a suitable architecture. Regime-conditioned sleeve MVO improved Sharpe ratio, drawdown, and wealth relative to its unconditional control, although not uniformly across robustness specifications. Macro-view Black–Litterman improved only modestly on its historical-view counterpart and reversed under the 48-month window. Interpretable macro states therefore did not automatically produce useful return views when weak and correlated forecasts entered a high-dimensional optimizer.

The retail implementation further qualified the frictionless evidence. MVO achieved the highest terminal net wealth, yet ERC finished only 0.620 contribution units lower, incurred much less direct cost, and achieved the highest net time-weighted return among the leading strategies. ERC consequently provided

the most balanced formal outcome across risk-adjusted performance, wealth, diversification, and implementation burden.

Hypothesis 1 received partial support, Hypothesis 2 received full descriptive support under the baseline rule, and Hypothesis 3 received partial support. The economic strength of H2 remained heterogeneous and specification-sensitive, while the H3 classification did not establish general practical dominance. The strongest strategies remained near the center of the conditional Random DCA distribution rather than showing unusually strong outcomes within that fixed design. VT provided favourable external market context, but its all-equity opportunity set and larger drawdown prevent a matched comparison. Overall, complexity is valuable only when stable gains justify its estimation, trading, and governance demands.

Summary

This thesis evaluates whether advanced ETF portfolio-allocation methods provide practical value relative to simple rules for a constrained retail investor. A growing universe of 71 ETFs is examined in a common recursive historical out-of-sample framework. Unconditional strategies are compared with transparent benchmarks, while macro-conditioned models are evaluated against matched controls. All strategies are then implemented with monthly contributions and transaction costs, and their terminal outcomes are located within a distribution of 1,000 Random DCA paths.

The empirical design addresses three questions: whether unconditional optimized strategies outperform simple benchmarks (H1), whether macro-financial conditioning improves otherwise comparable models (H2), and whether favourable frictionless differences persist after monthly contributions and transaction costs are introduced (H3).

Under the explicit assessment rules, H1 and H3 receive partial support, whereas H2 receives full descriptive support because both baseline matched Sharpe-ratio differences are positive. The economic strength of the H2 evidence nevertheless remains uneven and specification-sensitive. Equal risk contribution provides the strongest balance of risk-adjusted performance, wealth accumulation, diversification, and implementation cost. The leading implemented strategies remain near the centre of the Random DCA distribution rather than its upper tail. Robust and transparent allocation rules therefore remain difficult benchmarks, and additional complexity is worthwhile only when its benefits survive estimation uncertainty and implementation demands.

Shrnutí

Tato diplomová práce hodnotí, zda pokročilé metody alokace ETF přinášejí drobnému investorovi praktickou hodnotu ve srovnání s jednoduchými pravidly.

Rostoucí soubor 71 ETF je zkoumán v jednotném rekurzivním historickém rámci testování mimo vzorek. Bezpodmínečné strategie jsou porovnány s transparentními benchmarky, zatímco makrofinančně podmíněné modely jsou hodnoceny vůči odpovídajícím modelům bez tohoto podmínění. Všechny strategie jsou následně implementovány s měsíčními vklady a transakčními náklady a jejich konečné výsledky jsou zasazeny do rozdělení 1 000 trajektorií Random DCA.

Empirický design zkoumá tři otázky: zda nepodmíněné optimalizované strategie překonávají jednoduché benchmarky (H1), zda makrofinanční podmínění zlepšuje jinak srovnatelné modely (H2) a zda příznivé rozdíly z prostředí bez frikčí přetrvávají i po zahrnutí měsíčních vkladů a transakčních nákladů (H3).

Podle explicitních hodnoticích pravidel získávají H1 a H3 částečnou podporu, zatímco H2 získává plnou deskriptivní podporu, protože oba párové rozdíly Sharpeova poměru jsou v základní specifikaci kladné. Ekonomická síla důkazů pro H2 však zůstává nerovnoměrná a citlivá na specifikaci. Strategie rovnoměrného příspěvku k riziku nabízí nejlepší kombinaci rizikově očištěné výkonnosti, růstu majetku, diverzifikace a implementačních nákladů. Vedoucí strategie zůstávají poblíž středu rozdělení Random DCA, nikoli v jeho horním chvostu. Robustní a transparentní pravidla proto zůstávají obtížně překonatelnými benchmarky a vyšší složitost má smysl pouze tehdy, pokud její přínosy přetrvávají i po zohlednění nejistoty odhadů a implementačních nároků.

Bibliography

- ANG, A. & G. BEKAERT (2003): “How Do Regimes Affect Asset Allocation?” *Technical Report 10080*, National Bureau of Economic Research.
- BESSLER, W., H. OPFER, & D. WOLFF (2017): “Multi-Asset Portfolio Optimization and Out-of-Sample Performance: An Evaluation of Black–Litterman, Mean-Variance, and Naive Diversification Approaches.” *The European Journal of Finance* **23(1)**: pp. 1–30.
- BESSLER, W., G. TAUSHANOV, & D. WOLFF (2021): “Factor Investing and Asset Allocation Strategies: A Comparison of Factor Versus Sector Optimization.” *Journal of Asset Management* **22(6)**: pp. 488–506.
- BIGLOVA, A., S. ORTOBELLI, S. T. RACHEV, & S. STOYANOV (2004): “Different Approaches to Risk Estimation in Portfolio Theory.” *The Journal of Portfolio Management* **31(1)**: pp. 103–112.
- BLACK, F. & R. LITTERMAN (1992): “Global Portfolio Optimization.” *Financial Analysts Journal* **48(5)**: pp. 28–43.
- BRANDT, M. W., P. SANTA-CLARA, & R. VALKANOV (2009): “Parametric Portfolio Policies: Exploiting Characteristics in the Cross-Section of Equity Returns.” *The Review of Financial Studies* **22(9)**: pp. 3411–3447.
- CAMPBELL, J. Y., Y. L. CHAN, & L. M. VICEIRA (2003): “A Multivariate Model of Strategic Asset Allocation.” *Journal of Financial Economics* **67(1)**: pp. 41–80.
- CAMPBELL, J. Y. & L. M. VICEIRA (2001): “Who Should Buy Long-Term Bonds?” *American Economic Review* **91(1)**: pp. 99–127.
- CARHART, M. M. (1997): “On Persistence in Mutual Fund Performance.” *The Journal of Finance* **52(1)**: pp. 57–82.

- CHOUEIFATY, Y. & Y. COIGNARD (2008): "Toward Maximum Diversification." *The Journal of Portfolio Management* **35**(1): pp. 40–51.
- COCHRANE, J. H. (2008): "The Dog That Did Not Bark: A Defense of Return Predictability." *The Review of Financial Studies* **21**(4): pp. 1533–1575.
- DE BONDT, W. F. M. & R. THALER (1985): "Does the Stock Market Overreact?" *The Journal of Finance* **40**(3): pp. 793–805.
- DEMIGUEL, V., L. GARLAPPI, & R. UPPAL (2009): "Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?" *The Review of Financial Studies* **22**(5): pp. 1915–1953.
- EMEKTTER, R., B. JIRASAKULDECH, & R. BEAVES (2024): "Efficient Asset Allocation for Individual Investors in the ETF World." *Financial Planning Research Journal* **8**(1): pp. 79–98.
- FAMA, E. F. & K. R. FRENCH (1992): "The Cross-Section of Expected Stock Returns." *The Journal of Finance* **47**(2): pp. 427–465.
- FAMA, E. F. & K. R. FRENCH (1993): "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* **33**(1): pp. 3–56.
- FAMA, E. F. & K. R. FRENCH (2015): "A Five-Factor Asset Pricing Model." *Journal of Financial Economics* **116**(1): pp. 1–22.
- FERSON, W. E. & C. R. HARVEY (1991): "The Variation of Economic Risk Premiums." *Journal of Political Economy* **99**(2): pp. 385–415.
- FLAVIN, T. J. & M. R. WICKENS (2003): "Macroeconomic Influences on Optimal Asset Allocation." *Review of Financial Economics* **12**(2): pp. 207–231.
- GORTON, G. B. & K. G. ROUWENHORST (2006): "Facts and Fantasies about Commodity Futures." *Financial Analysts Journal* **62**(2): pp. 47–68.
- GUIDOLIN, M. & A. TIMMERMAN (2007): "Asset Allocation under Multivariate Regime Switching." *Journal of Economic Dynamics and Control* **31**(11): pp. 3503–3544.
- HOESLI, M., J. LEKANDER, & W. WITKIEWICZ (2004): "International Evidence on Real Estate as a Portfolio Diversifier." *Journal of Real Estate Research* **26**(2): pp. 161–206.

- HOLM, S. (1979): "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* **6(2)**: pp. 65–70.
- HUANG, M.-Y. & J.-B. LIN (2011): "Do ETFs Provide Effective International Diversification?" *Research in International Business and Finance* **25(3)**: pp. 335–344.
- JAGANNATHAN, R. & T. MA (2003): "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps." *The Journal of Finance* **58(4)**: pp. 1651–1683.
- JEGADEESH, N. & S. TITMAN (1993): "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *The Journal of Finance* **48(1)**: pp. 65–91.
- JORION, P. (1986): "Bayes-Stein Estimation for Portfolio Analysis." *The Journal of Financial and Quantitative Analysis* **21(3)**: pp. 279–292.
- KOLLÁR, M. & C. SCHMIEDER (2019): "Macro-Financial Cycle and Optimal Asset Allocation." *Technical report*, European Investment Fund.
- LEDOIT, O. & M. WOLF (2004): "Honey, I Shrunk the Sample Covariance Matrix." *The Journal of Portfolio Management* **30(4)**: pp. 110–119.
- LEDOIT, O. & M. WOLF (2012): "Nonlinear shrinkage estimation of large-dimensional covariance matrices." *The Annals of Statistics* **40(2)**: pp. 1024–1060.
- LEDOIT, O. & M. WOLF (2017): "Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks." *The Review of Financial Studies* **30(12)**: pp. 4349–4388.
- LEDOIT, O. & M. WOLF (2022a): "The power of (non-)linear shrinking: A review and guide to covariance matrix estimation." *Journal of Financial Econometrics* **20(1)**: pp. 187–218.
- LEDOIT, O. & M. WOLF (2022b): "Quadratic shrinkage for large covariance matrices." *Bernoulli* **28(3)**: pp. 1519–1547.
- LEHMANN, B. N. (1990): "Fads, Martingales, and Market Efficiency." *The Quarterly Journal of Economics* **105(1)**: pp. 1–28.

- LETTAU, M. & S. LUDVIGSON (2001): "Consumption, Aggregate Wealth, and Expected Stock Returns." *The Journal of Finance* **56**(3): pp. 815–849.
- LETTAU, M. & A. MADHAVAN (2018): "Exchange-Traded Funds 101 for Economists." *Journal of Economic Perspectives* **32**(1): pp. 135–154.
- LINTNER, J. (1965): "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets." *The Review of Economics and Statistics* **47**(1): pp. 13–37.
- MA, Y., L. MACLEAN, K. XU, & Y. ZHAO (2011): "A Portfolio Optimization Model with Regime-Switching Risk Factors for Sector Exchange Traded Funds." *Pacific Journal of Optimization* **7**(1): pp. 281–298.
- MAILLARD, S., T. RONCALLI, & J. TEILETCHE (2010): "The Properties of Equally Weighted Risk Contribution Portfolios." *The Journal of Portfolio Management* **36**(4): pp. 60–70.
- MARKOWITZ, H. (1952): "Portfolio selection." *The Journal of Finance* **7**(1): pp. 77–91.
- MERTON, R. C. (1980): "On Estimating the Expected Return on the Market: An Exploratory Investigation." *Journal of Financial Economics* **8**(4): pp. 323–361.
- MOSSIN, J. (1966): "Equilibrium in a capital asset market." *Econometrica* **34**(4): pp. 768–783.
- NEZAFAT, P. & C. SLAVÍK (2021): "Asset Prices and Business Cycles with Liquidity Shocks." SSRN Working Paper No. 3963053.
- PASTOR, L. (2000): "Portfolio Selection and Asset Pricing Models." *The Journal of Finance* **55**(1): pp. 179–223.
- PASTOR, L. & R. F. STAMBAUGH (2000): "Comparing Asset Pricing Models: An Investment Perspective." *Journal of Financial Economics* **56**(3): pp. 335–381.
- PETRONIO, F., T. LANDO, A. BIGLOVA, & S. ORTOBELLI (2014): "Optimal Portfolio Performance with Exchange-Traded Funds." *Central European Review of Economic Issues* **17**(1): pp. 5–12.

- POLITIS, D. N. & J. P. ROMANO (1994): “The Stationary Bootstrap.” *Journal of the American Statistical Association* **89(428)**: pp. 1303–1313.
- PUELZ, D., C. M. CARVALHO, & P. R. HAHN (2015): “Optimal ETF Selection for Passive Investing.” arXiv Preprint arXiv:1510.03385 [q-fin.ST].
- ROCKAFELLAR, R. T. & S. URYASEV (2000): “Optimization of Conditional Value-at-Risk.” *The Journal of Risk* **2(3)**: pp. 21–41.
- SHARPE, W. F. (1964): “Capital asset prices: A theory of market equilibrium under conditions of risk.” *The Journal of Finance* **19(3)**: pp. 425–442.
- SHARPE, W. F. (1966): “Mutual fund performance.” *The Journal of Business* **39(1)**: pp. 119–138.
- THE VANGUARD GROUP (2026): “Vanguard Total World Stock ETF (VT) Fund Profile.” Vanguard Total World Stock Index Fund ETF Shares (VT), accessed 25 July 2026.
- TOBIN, J. (1958): “Liquidity preference as behavior towards risk.” *The Review of Economic Studies* **25(2)**: pp. 65–86.
- TU, J. & G. ZHOU (2011): “Markowitz Meets Talmud: A Combination of Sophisticated and Naive Diversification Strategies.” *Journal of Financial Economics* **99(1)**: pp. 204–215.
- WELCH, I. & A. GOYAL (2008): “A Comprehensive Look at the Empirical Performance of Equity Premium Prediction.” *The Review of Financial Studies* **21(4)**: pp. 1455–1508.
- YADAV, R., A. SHARMA, & A. MEHRA (2026): “Shrinkage Estimators for Mean and Covariance: Evidence on Portfolio Efficiency Across Market Dimensions.” arXiv Preprint arXiv:2601.20643 [q-fin.PM].

Appendix A

Supplementary Notes on Mean-Variance Optimization

A.1 Constrained and Utility-Based Formulations

This appendix shows the relationship between the constrained mean-variance problem and the utility-maximization formulation. Let w denote the vector of portfolio weights, μ the vector of expected asset returns, Σ the covariance matrix of asset returns, and $\mathbf{1}$ a vector of ones. The covariance matrix is assumed to be positive definite.

First, consider the constrained minimization problem. For a given target expected return μ_p , the investor chooses the portfolio with the lowest possible variance:

$$\begin{aligned} \min_w \quad & w' \Sigma w \\ \text{subject to} \quad & w' \mu = \mu_p, \\ & \mathbf{1}' w = 1. \end{aligned} \tag{A.1}$$

The Lagrangian for this problem is

$$\mathcal{L} = w' \Sigma w - \lambda (w' \mu - \mu_p) - \nu (\mathbf{1}' w - 1), \tag{A.2}$$

where λ and ν are Lagrange multipliers. The first-order conditions are

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w} &= 2\Sigma w - \lambda\mu - \nu\mathbf{1} = 0, \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= -(w'\mu - \mu_p) = 0, \\
\frac{\partial \mathcal{L}}{\partial \nu} &= -(\mathbf{1}'w - 1) = 0.
\end{aligned} \tag{A.3}$$

The first condition in Equation (A.3) gives

$$w = \frac{1}{2}\Sigma^{-1}(\lambda\mu + \nu\mathbf{1}). \tag{A.4}$$

Define the following scalars:

$$A = \mathbf{1}'\Sigma^{-1}\mathbf{1}, \quad B = \mathbf{1}'\Sigma^{-1}\mu, \quad C = \mu'\Sigma^{-1}\mu, \quad D = AC - B^2. \tag{A.5}$$

Substituting $w = \frac{1}{2}\Sigma^{-1}(\lambda\mu + \nu\mathbf{1})$ into the two constraints gives the following system for the Lagrange multipliers:

$$\begin{aligned}
B\lambda + A\nu &= 2, \\
C\lambda + B\nu &= 2\mu_p.
\end{aligned} \tag{A.6}$$

Solving this system gives

$$\lambda = \frac{2(A\mu_p - B)}{D}, \quad \nu = \frac{2(C - B\mu_p)}{D}. \tag{A.7}$$

Substituting these multipliers back into the expression for w gives the constrained mean-variance solution:

$$w(\mu_p) = \frac{C - B\mu_p}{D}\Sigma^{-1}\mathbf{1} + \frac{A\mu_p - B}{D}\Sigma^{-1}\mu. \tag{A.8}$$

Equation (A.8) gives the minimum-variance portfolio that achieves the target expected return μ_p . Varying μ_p traces out the mean-variance frontier.

The same risk-return trade-off can also be expressed using a utility maximization problem. Instead of fixing a target return and minimizing variance, the investor maximizes expected return while penalizing portfolio variance:

$$\begin{aligned}
\max_w \quad & w'\mu - \frac{\gamma}{2}w'\Sigma w \\
\text{subject to} \quad & \mathbf{1}'w = 1.
\end{aligned} \tag{A.9}$$

In Equation (A.9), the parameter $\gamma > 0$ captures the investor's degree of

risk aversion. A higher value of γ means that the investor places a stronger penalty on portfolio variance. A lower value of γ means that the investor is more willing to accept risk in exchange for higher expected return.

The Lagrangian for the utility problem is

$$\mathcal{L} = w'\mu - \frac{\gamma}{2}w'\Sigma w - \eta(\mathbf{1}'w - 1), \quad (\text{A.10})$$

where η is the Lagrange multiplier on the budget constraint. The first-order conditions are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w} &= \mu - \gamma \Sigma w - \eta \mathbf{1} = 0, \\ \frac{\partial \mathcal{L}}{\partial \eta} &= -(\mathbf{1}'w - 1) = 0. \end{aligned} \quad (\text{A.11})$$

The first condition in Equation (A.11) gives

$$w = \frac{1}{\gamma} \Sigma^{-1} (\mu - \eta \mathbf{1}). \quad (\text{A.12})$$

Substituting this expression into the budget constraint gives a one-equation system for the multiplier η :

$$B - A\eta = \gamma. \quad (\text{A.13})$$

Therefore,

$$\eta = \frac{B - \gamma}{A}. \quad (\text{A.14})$$

The utility-maximizing portfolio is therefore

$$w(\gamma) = \frac{1}{A} \Sigma^{-1} \mathbf{1} + \frac{1}{\gamma} \left(\Sigma^{-1} \mu - \frac{B}{A} \Sigma^{-1} \mathbf{1} \right). \quad (\text{A.15})$$

This solution has a direct interpretation. The first term is the global minimum-variance portfolio. The second term tilts the portfolio toward assets with higher expected returns, adjusted for covariance. The strength of this tilt depends on γ . A higher γ reduces the tilt toward risky expected-return opportunities, while a lower γ increases it.

The two formulations are equivalent in the following sense. The utility-maximizing portfolio in Equation (A.15) has expected return

$$\mu_p(\gamma) = \mu'w(\gamma) = \frac{B}{A} + \frac{D}{\gamma A}. \quad (\text{A.16})$$

If this value of $\mu_p(\gamma)$ is substituted into the constrained solution in Equation (A.8), the result is exactly the utility solution in Equation (A.15). Therefore, the constrained minimization problem and the utility maximization problem describe the same efficient set from two different perspectives. The constrained problem asks which portfolio gives the lowest variance for a chosen expected return. The utility problem asks which point on the same frontier is preferred by an investor with risk aversion γ .

The intuition of the model is therefore simple but powerful. Diversification allows investors to combine assets in a way that can reduce risk without necessarily reducing expected return by the same proportion. This happens because assets do not always move together. As a result, a diversified portfolio might offer a better risk-return combination than a single asset. This insight remains one of the central foundations of modern portfolio theory.

A.2 Tobin Separation and the Tangency Portfolio

Tobin's separation result can be understood as an extension of the mean-variance framework in which investors can combine a risky asset or risky portfolio with a safe asset. In Tobin's original liquidity-preference setting, cash plays the role of the safe asset. The investor allocates a share A_1 to cash and a share A_2 to an interest-bearing asset, so that

$$A_1 + A_2 = 1. \quad (\text{A.17})$$

If the interest-bearing asset offers current yield r and an uncertain capital gain or loss g , the portfolio return can be written as

$$R = A_2(r + g). \quad (\text{A.18})$$

Assuming $\mathbb{E}(g) = 0$, expected return and risk are

$$\mathbb{E}(R) = A_2 r, \quad \sigma_R = A_2 \sigma_g, \quad (\text{A.19})$$

where σ_g is the standard deviation of the capital gain or loss. Eliminating A_2 gives

$$\mathbb{E}(R) = \frac{r}{\sigma_g} \sigma_R. \quad (\text{A.20})$$

This equation shows that combinations of cash and the risky asset form a

straight line in risk-return space. In the modern portfolio setting, the same idea is expressed by combining a risk-free asset with return r_f and a risky portfolio.

Let μ denote the vector of expected returns on risky assets and let Σ denote the covariance matrix of risky-asset returns. Define the vector of expected excess returns as

$$\mu_e = \mu - r_f \mathbf{1}. \quad (\text{A.21})$$

For any risky portfolio with weights w , the expected excess return and variance are

$$w' \mu_e, \quad w' \Sigma w. \quad (\text{A.22})$$

The tangency portfolio is the risky portfolio that maximizes the Sharpe ratio:

$$\max_w \frac{w' \mu_e}{\sqrt{w' \Sigma w}} \quad \text{subject to} \quad \mathbf{1}' w = 1. \quad (\text{A.23})$$

Since multiplying the risky weights by a positive scalar does not change the Sharpe ratio, the direction of the tangency portfolio is proportional to $\Sigma^{-1} \mu_e$. Imposing the budget constraint $\mathbf{1}' w_T = 1$ gives

$$w_T = \frac{\Sigma^{-1}(\mu - r_f \mathbf{1})}{\mathbf{1}' \Sigma^{-1}(\mu - r_f \mathbf{1})}. \quad (\text{A.24})$$

Equation (A.24) gives the composition of the optimal risky portfolio. It depends on expected returns, covariances, and the risk-free rate, but not on the investor's risk aversion. Risk aversion determines only how much of this risky portfolio the investor holds.

Suppose the investor allocates a share y of wealth to the tangency portfolio T and the remaining share $1 - y$ to the risk-free asset. The return on the complete portfolio is

$$R_p = (1 - y)r_f + yR_T, \quad (\text{A.25})$$

where R_T is the return on the tangency portfolio. The expected return is

$$\mathbb{E}(R_p) = r_f + y[\mathbb{E}(R_T) - r_f], \quad (\text{A.26})$$

and, because the risk-free asset has zero variance and zero covariance with the risky portfolio, the variance is

$$\sigma_p^2 = y^2 \sigma_T^2. \quad (\text{A.27})$$

A.2.1 Optimal Allocation Between the Risk-Free Asset and the Tangency Portfolio

The investor is assumed to maximize the following mean-variance utility function:

$$U = \mathbb{E}(R_p) - \frac{\gamma}{2} \sigma_p^2, \quad (\text{A.28})$$

where $\gamma > 0$ measures the investor's degree of risk aversion. Substituting the expressions for expected return and variance gives

$$U(y) = r_f + y[\mathbb{E}(R_T) - r_f] - \frac{\gamma}{2} y^2 \sigma_T^2. \quad (\text{A.29})$$

The first-order condition is

$$\frac{dU}{dy} = [\mathbb{E}(R_T) - r_f] - \gamma y \sigma_T^2 = 0. \quad (\text{A.30})$$

Solving for y gives

$$y^* = \frac{\mathbb{E}(R_T) - r_f}{\gamma \sigma_T^2}. \quad (\text{A.31})$$

The second-order condition is

$$\frac{d^2 U}{dy^2} = -\gamma \sigma_T^2 < 0, \quad (\text{A.32})$$

because $\gamma > 0$ and $\sigma_T^2 > 0$. Therefore, the utility function is concave in y , and the first-order condition gives the unique unconstrained optimum.

The result has a clear economic interpretation. The optimal share invested in the tangency portfolio increases when the expected excess return $\mathbb{E}(R_T) - r_f$ is higher. It decreases when risk aversion γ is higher and when the variance σ_T^2 of the tangency portfolio is higher. Thus, the tangency portfolio determines the composition of the risky part, while y^* determines the scale of risky exposure.

If borrowing and lending at the risk-free rate are unrestricted, y^* may be greater than one. In that case, the investor borrows at the risk-free rate and invests more than 100 percent of wealth in the tangency portfolio. If borrowing and short selling are not allowed, the allocation is usually restricted to

$$0 \leq y \leq 1. \tag{A.33}$$

Under this restriction, the optimal allocation becomes

$$y^* = \min \left\{ 1, \max \left[0, \frac{\mathbb{E}(R_T) - r_f}{\gamma \sigma_T^2} \right] \right\}. \tag{A.34}$$

Appendix B

Supplementary Notes on ETF Factor Portfolios

B.1 Description of Factor and Anomaly Portfolios

This appendix briefly describes the factor and anomaly portfolios used in the ETF replication literature. These portfolios are not individual securities, but systematic long–short strategies constructed from large cross-sections of stocks. They are often used in empirical asset pricing because they capture broad sources of return variation that are difficult for individual investors to implement directly.

Market factor. The market factor represents the excess return of the broad equity market over the risk-free rate. It captures the compensation investors require for bearing aggregate market risk and is the central factor in CAPM and later multi-factor asset-pricing models (Sharpe 1964; Fama & French 1993).

Size factor. The size factor captures the return difference between small-cap and large-cap stocks. It is commonly represented by SMB, or “small minus big”, and reflects the empirical finding that smaller firms have historically earned different average returns from larger firms (Fama & French 1992; 1993).

Value factor. The value factor captures the return difference between high book-to-market stocks and low book-to-market stocks. It is commonly represented by HML, or “high minus low”. This factor reflects the value premium, where firms with relatively low market prices compared with book value have historically behaved differently from growth firms (Fama & French 1992; 1993).

Profitability factor. The profitability factor captures the return difference between firms with robust profitability and firms with weak profitability. In the Fama–French five-factor model, this factor is motivated by the evidence that more profitable firms tend to earn higher average returns, controlling for other characteristics (Fama & French 2015).

Investment factor. The investment factor captures the return difference between firms with conservative investment policies and firms with aggressive investment policies. In the Fama–French five-factor model, firms that invest more conservatively tend to earn higher average returns than firms that expand assets aggressively (Fama & French 2015).

Momentum factor. The momentum factor captures the tendency of stocks with strong recent performance to continue outperforming in the short to medium term, while recent losers tend to continue underperforming. This anomaly was documented by Jegadeesh & Titman (1993) and later incorporated into fund performance analysis by Carhart (1997).

Short-term reversal factor. The short-term reversal factor captures the tendency of very recent winners to underperform and very recent losers to outperform over short horizons. This effect is often interpreted as evidence of temporary price pressure, liquidity effects, or short-run overreaction (Lehmann 1990).

Long-term reversal factor. The long-term reversal factor captures the tendency of stocks with poor long-run past performance to outperform previous long-run winners. This pattern is commonly associated with overreaction and subsequent correction in stock prices (De Bondt & Thaler 1985).

Overall, these factor and anomaly portfolios represent important dimensions of systematic return variation in equity markets. Since most of them are constructed as theoretical long–short portfolios and require trading many individual stocks, they are difficult for retail investors to implement directly. This motivates the use of ETFs as practical and investable proxies for broad factor exposures.

Appendix C

Data Construction and Core ETF Universe

C.1 Synthetic U.S. Treasury-Bill Series

The synthetic risk-free series converts the final daily 3-month Treasury constant maturity yield observed by each portfolio-formation date into the monthly return defined in (2.3). The series is used for excess-return calculations and does not enter the investable ETF set. The code first selects the final non-missing `DGS3MO` observation available by the end of month $t - 1$ and then applies the annual-to-monthly transformation. No later yield is carried backward.

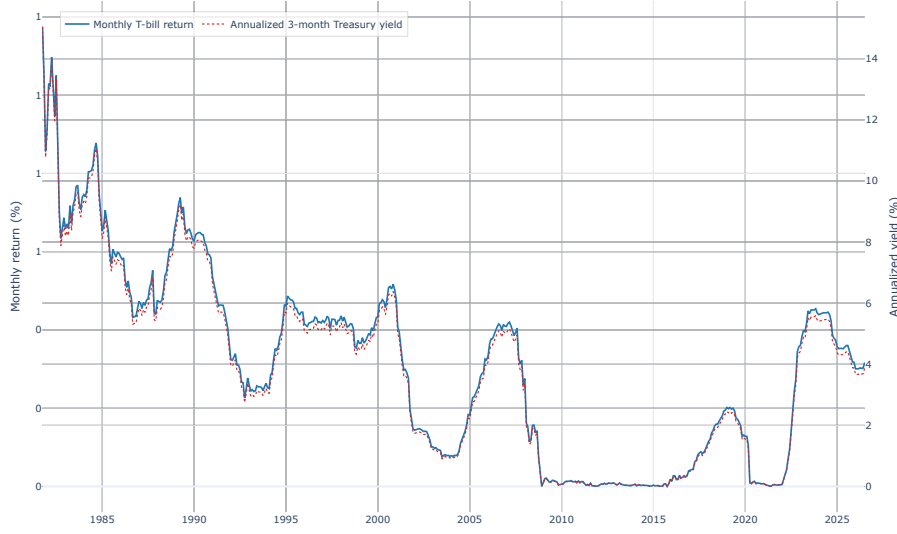


Figure C.1: Synthetic monthly U.S. Treasury-bill return constructed from the FRED daily 3-month Treasury constant maturity yield

Source: Author's calculations from FRED series DGS3M0.

C.2 Real-Time Macroeconomic and Financial Series

The H2 regime engine uses one real-time monthly activity measure and three daily financial series aggregated to monthly frequency. This appendix specifies the information timing used to construct the vector in (2.31).

Let d_t be the final portfolio-formation date in calendar month t . For CFNAI-MA3, let $\mathcal{V}(d_t)$ be the set of publication vintages released no later than d_t . For vintage $v \in \mathcal{V}(d_t)$, let $q(v)$ denote the latest reference month for which the vintage contains a reported value. The real-time activity observation is

$$A_t^{RT} = CFNAIMA3_{v_t}(q(v_t)), \quad v_t = \max \mathcal{V}(d_t). \quad (C.1)$$

Thus, the backtest uses the newest value that had actually been released by formation. The reference month may precede the formation month because CFNAI is published with a delay. Historical values are reconstructed from the Chicago Fed's past-release archive and, where the relevant dates are covered, from ALFRED or the FRED API's vintage-date functionality. The replication file retains the source release associated with every observation. Today's revised

history is never shifted backward and treated as if it had been observed in real time. Before the backtest is run, the code verifies complete archival coverage over the required HMM sample. Any uncovered month is flagged and can enter only a separately identified robustness specification; it is not silently replaced with revised data.

For VIX, the term spread, and the Baa–Treasury spread, the code downloads all daily observations available in each calendar month and applies (2.28)–(2.30). The monthly averages are calculated from the available business-day observations; no weekend or holiday value is manufactured. VIX and BAA10Y are checked to be strictly positive before logarithms are taken. The three financial series and the real-time CFNAI observation are then merged by information month. The common HMM sample begins with the first month for which all four values are available.

For reproducibility, the stored macro input file contains at least the following fields: information month, source series, economic reference month where applicable, vintage date, release date, raw value, transformed monthly value, and download timestamp. This structure makes it possible to verify that a portfolio decision did not use a later macroeconomic revision.

C.3 List of 71 ETFs

The following tables report the complete ETF universe. The “first return” column gives the first valid monthly USD return in the processed history. Eligibility occurs later because (2.4) requires a complete trailing 60-month return window.

Table C.1: Core ETF universe, part 1

<i>Ticker</i>	<i>Fund name</i>	<i>Permanent sleeve</i>	<i>Class</i>	<i>First return</i>
EEM	Emerging Markets	Emerging-market and regional equity	Equity	2003-05
EWA	Australia	Developed-country equity	Equity	1996-04
EWG	Germany	Developed-country equity	Equity	1996-04
EWI	Italy	Developed-country equity	Equity	1996-04
EWJ	Japan	Developed-country equity	Equity	1996-04
EWK	France	Developed-country equity	Equity	1996-04
EWL	Taiwan	Developed-country equity	Equity	1996-04
EWU	United Kingdom	Developed-country equity	Equity	2000-07
EWY	South Korea	Developed-country equity	Equity	1996-04
ILF	Latin America 40	Emerging-market and regional equity	Equity	2000-06
INDA	India	Emerging-market and regional equity	Equity	2001-11
MCHI	China	Emerging-market and regional equity	Equity	2012-03
SPY	S&P 500	U.S. broad equity	Equity	2011-04
VEA	FTSE Developed Markets	Broad developed ex-U.S. equity	Equity	1993-02
VWG	FTSE Europe	Broad developed ex-U.S. equity	Equity	2007-08
VOX	Communication Services	Broad developed ex-U.S. equity	Equity	2005-04
XLB	Materials	U.S. sectors	Equity	2004-10
XLE	Energy	U.S. sectors	Equity	1999-01
XLF	Financial	U.S. sectors	Equity	1999-01
XLI	Industrial	U.S. sectors	Equity	1999-01
XLK	Technology	U.S. sectors	Equity	1999-01
XLV	Consumer Staples	U.S. sectors	Equity	1999-01
XLW	Utilities	U.S. sectors	Equity	1999-01
XLV	Health Care	U.S. sectors	Equity	1999-01
XLV	Consumer Discretionary	U.S. sectors	Equity	1999-01
EXH1.DE	STOXX Europe 600 Banks	European sectors	Equity	2008-02
EXH2.DE	STOXX Europe 600 Insurance	European sectors	Equity	2008-02
EXH3.DE	STOXX Europe 600 Oil & Gas	European sectors	Equity	2008-02
EXH4.DE	STOXX Europe 600 Basic Resources	European sectors	Equity	2008-02
EXH5.DE	STOXX Europe 600 Chemicals	European sectors	Equity	2008-02
EXH6.DE	STOXX Europe 600 Construction & Materials	European sectors	Equity	2008-02
EXH7.DE	STOXX Europe 600 Financial Services	European sectors	Equity	2008-02
EXH9.DE	STOXX Europe 600 Automobiles & Parts	European sectors	Equity	2008-02
EXV1.DE	STOXX Europe 600 Food & Beverage	European sectors	Equity	2008-02
EXV2.DE	STOXX Europe 600 Health Care	European sectors	Equity	2008-02

Source: Author's classification based on the processed ETF universe and monthly return histories.

Table C.2: Core ETF universe, part 2

<i>Ticker</i>	<i>Fund name</i>	<i>Permanent sleeve</i>	<i>Class</i>	<i>First return</i>
EXV3.DE	STOXX Europe 600 Industrial Goods & Services	European sectors	Equity	2008-02
EXV4.DE	STOXX Europe 600 Media	European sectors	Equity	2008-02
EXV5.DE	STOXX Europe 600 Personal & Household Goods	European sectors	Equity	2008-02
EXV6.DE	STOXX Europe 600 Retail	European sectors	Equity	2008-02
EXV7.DE	STOXX Europe 600 Technology	European sectors	Equity	2008-02
EXV8.DE	STOXX Europe 600 Telecommunications	European sectors	Equity	2008-02
EXV9.DE	STOXX Europe 600 Utilities	European sectors	Equity	2008-02
CHIQ	China Consumer Discretionary	Asia/China sectors	Equity	2010-01
CQQQ	China Technology	Asia/China sectors	Equity	2010-02
KWEB	CSI China Internet	Asia/China sectors	Equity	2013-09
HDV	High Dividend	U.S. factors	Equity	2011-04
MTUM	USA Momentum Factor	U.S. factors	Equity	2013-05
QUAL	USA Quality Factor	U.S. factors	Equity	2013-08
SIZE	USA Size Factor	U.S. factors	Equity	2013-05
USMV	USA Min Vol Factor	U.S. factors	Equity	2013-05
VLUE	USA Value Factor	U.S. factors	Equity	2013-05
IEFM.L	MSCI Europe Momentum Factor	European factors	Equity	2015-02
IEFQL	MSCI Europe Quality Factor	European factors	Equity	2015-02
IEFVL	MSCI Europe Value Factor	European factors	Equity	2015-02
MOVLL	MSCI Europe Minimum Volatility	European factors	Equity	2012-12
EFV	EAFE Min Vol Factor	International factors	Equity	2011-11
INTM	Intl Momentum Factor	International factors	Equity	2015-02
IQLT	Intl Quality Factor	International factors	Equity	2015-02
IVLU	Intl Value Factor	International factors	Equity	2015-02
SHY	1-3 Year Treasury Bond	Short-duration Treasuries	Equity	2002-08
TLT	20+ Year Treasury Bond	Long-duration Treasuries	Treasuries	2002-08
GDX	Gold Miners	Precious metals / gold-linked	Commodities	2006-06
GLD	Gold	Precious metals / gold-linked	Commodities	2004-12
SLV	Silver	Precious metals / gold-linked	Commodities	2006-05
DBB	DB Base Metals	Broad / industrial commodities	Commodities	2007-02
DBC	DB Commodity	Broad / industrial commodities	Commodities	2006-03
EXH8.DE	STOXX Europe 600 Real Estate	Broad real estate	Real estate	2008-02
IPRPL	European Property Yield	Listed real estate	Real estate	2009-02
VNQ	Real Estate	Listed real estate	Real estate	2004-10
VNQI	Global ex-U.S. Real Estate	Listed real estate	Real estate	2010-12

Source: Author's classification based on the processed ETF universe and monthly return histories.

C.4 Permanent Sleeve Classification

Every ETF belongs to one and only one permanent economic sleeve. The classification is fixed before the backtest and is used by the hierarchical benchmarks, the Black-Litterman view matrix, and the sleeve-level regime allocation. The four-way regional split ensures that broad U.S., broad developed ex-U.S., developed-country, and emerging-market decisions receive separate economic representation. EXH8.DE is classified as listed real estate rather than as a European sector. SHY and TLT form separate duration sleeves, and commodity exposures are divided into precious-metal and broad or industrial groups.

Table C.3: Permanent economic sleeves and member ETFs

<i>No.</i>	<i>Sleeve</i>	<i>Class</i>	<i>N</i>	<i>Member tickers</i>
1	U.S. broad equity	Equity	1	SPY
2	Broad developed ex-U.S. equity	Equity	2	VEA, V GK
3	Developed-country equity	Equity	7	VEA, EWG, EWI, EWJ, EWP, EWQ, EWU
4	Emerging-market and regional equity	Equity	6	EEM, EWT, EWY, ILF, INDA, MCHI
5	U.S. sectors	Equity	10	VOX, XLB, XLE, XLF, XLJ, XLK, XLP, XLU, XLV, XLY
6	European sectors	Equity	17	EXH1.DE, EXH2.DE, EXH3.DE, EXH4.DE, EXH5.DE, EXH6.DE, EXH7.DE, EXH9.DE, EXV1.DE, EXV2.DE, EXV3.DE, EXV4.DE, EXV5.DE, EXV6.DE, EXV7.DE, EXV8.DE, EXV9.DE
7	Asia/China sectors	Equity	3	CHIQ, CQQQ, KWEB
8	U.S. factors	Equity	6	HDV, MTUM, QUAL, SIZE, USMV, VLUE
9	European factors	Equity	4	IEFM.L, IEFQ.L, IEFV.L, MVOLL
10	International factors	Equity	4	EFV, IMTM, IQLT, IVLU
11	Short-duration Treasuries	Treasuries	1	SHY
12	Long-duration Treasuries	Treasuries	1	TLT
13	Precious metals and gold-linked	Commodities	3	GDX, GLD, SLV
14	Broad and industrial commodities	Commodities	2	DBB, DBC
15	Listed real estate	Real estate	4	EXH8.DE, IPRP.L, VNQ, VNQI
Total			71	

Source: Author's classification based on the processed ETF universe.

Appendix D

Quadratic-Inverse Shrinkage Covariance Estimation

This appendix describes the construction of the Quadratic-Inverse Shrinkage (QIS) covariance estimator used in the empirical analysis. The purpose is to make the implementation reproducible and to clarify how the regularized eigenvalues are obtained. The full large-dimensional asymptotic proofs are provided by Ledoit & Wolf (2022b); the presentation below focuses on the steps required for the empirical implementation.

D.1 Data Matrix and Sample Covariance

At a portfolio-formation date t , let $p = N_t$ for an ETF-level matrix and $p = |\mathcal{G}_t|$ for a sleeve-level matrix. Let $W_\Sigma = 60$ denote the covariance-estimation window. Arrange the monthly ETF returns in the matrix

$$Y_t = \begin{bmatrix} R_{t-W_\Sigma}^\top \\ \vdots \\ R_{t-1}^\top \end{bmatrix}, \quad Y_t \in \mathbb{R}^{W_\Sigma \times p}. \quad (\text{D.1})$$

After subtracting the column means, let X_t denote the centred return matrix. Because one degree of freedom is used for demeaning, the effective sample size is

$$n = W_\Sigma - 1. \quad (\text{D.2})$$

The sample covariance matrix can then be written as

$$S_t = \frac{1}{n} X_t^\top X_t. \quad (\text{D.3})$$

For notational simplicity, the time subscript t is omitted in the remaining derivation. The spectral decomposition of S is

$$S = U \Lambda U^\top, \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_p), \quad (\text{D.4})$$

where the eigenvalues are ordered as

$$0 \leq \lambda_1 \leq \dots \leq \lambda_p. \quad (\text{D.5})$$

Here, $U = [u_1, \dots, u_p]$ is the orthogonal matrix whose columns are the sample eigenvectors. Each eigenvector represents an estimated direction of joint variation in the ETF-return universe, while its eigenvalue measures the sample variance assigned to that direction.

D.2 Rotation-Equivariant Oracle

QIS belongs to the class of rotation-equivariant covariance estimators. Such an estimator retains the sample eigenvectors and changes only the eigenvalues:

$$\hat{\Sigma} = U D U^\top, \quad D = \text{diag}(d_1, \dots, d_p). \quad (\text{D.6})$$

To see which eigenvalues would be optimal if the population covariance matrix Σ were known, consider the squared Frobenius loss

$$L_F(\hat{\Sigma}, \Sigma) = \frac{1}{p} \|\hat{\Sigma} - \Sigma\|_F^2. \quad (\text{D.7})$$

Using $\hat{\Sigma} = U D U^\top$ and the orthogonality of U gives

$$\begin{aligned} p L_F &= \text{tr} \left[(U D U^\top - \Sigma)^2 \right] \\ &= \text{tr}(D^2) - 2 \text{tr}(D U^\top \Sigma U) + \text{tr}(\Sigma^2) \\ &= \sum_{i=1}^p d_i^2 - 2 \sum_{i=1}^p d_i u_i^\top \Sigma u_i + \text{tr}(\Sigma^2). \end{aligned} \quad (\text{D.8})$$

Differentiating with respect to d_i yields

$$\frac{\partial L_F}{\partial d_i} = \frac{2}{p} (d_i - u_i^\top \Sigma u_i). \quad (\text{D.9})$$

The loss-minimizing eigenvalue is therefore

$$d_i^{\text{oracle}} = u_i^\top \Sigma u_i. \quad (\text{D.10})$$

This is the true variance in the direction of the sample eigenvector u_i . It is not generally equal to the corresponding population eigenvalue because the sample eigenvectors themselves are estimated with error. Since the population covariance matrix Σ is unobservable, (D.10) cannot be implemented directly. QIS constructs feasible estimates of these oracle quantities from the complete sample eigenvalue spectrum.

D.3 Smoothed Local Shrinkage Functions

Define the concentration ratio as

$$c = \frac{p}{n}. \quad (\text{D.11})$$

Let

$$m = \min\{p, n\} \quad (\text{D.12})$$

denote the number of nonzero sample eigenvalues and define their index set as

$$\mathcal{J} = \{p - m + 1, \dots, p\}. \quad (\text{D.13})$$

For each nonzero sample eigenvalue, define the inverse eigenvalue

$$q_j = \lambda_j^{-1}, \quad j \in \mathcal{J}. \quad (\text{D.14})$$

The smoothing bandwidth recommended by Ledoit & Wolf (2022b) is

$$h = \left[\min \{c^2, c^{-2}\} \right]^{0.35} p^{-0.35}. \quad (\text{D.15})$$

For an inverse eigenvalue x , the smoothed Stein shrinker is

$$\hat{\theta}_h(x) = \frac{1}{m} \sum_{j \in \mathcal{J}} \frac{q_j(q_j - x)}{(q_j - x)^2 + h^2 q_j^2}. \quad (\text{D.16})$$

Its conjugate component is

$$\hat{\theta}_h^*(x) = \frac{1}{m} \sum_{j \in \mathcal{J}} \frac{h q_j^2}{(q_j - x)^2 + h^2 q_j^2}. \quad (\text{D.17})$$

The squared amplitude is

$$\hat{A}_h^2(x) = \hat{\theta}_h(x)^2 + \hat{\theta}_h^*(x)^2. \quad (\text{D.18})$$

The denominator in (D.16)–(D.17) depends on the distance between q_j and x . Consequently, nearby inverse eigenvalues exert a stronger influence, while the influence of distant eigenvalues declines. The bandwidth h prevents the local adjustment from becoming unstable when two sample eigenvalues are very close.

D.4 Nonsingular and Boundary Cases

The QIS implementation distinguishes between the cases $p \leq n$ and $p > n$, where p is the number of assets and n is the effective sample size used by the estimator. The case $p = n$ is treated as the boundary of the $p \leq n$ formula rather than as part of the singular $p > n$ branch. This convention follows the published reference implementation and is tested explicitly because the growing ETF universe reaches $p = n$ during several monthly estimation dates.

For $p \leq n$, let

$$c = \frac{p}{n} \leq 1.$$

When $p < n$, all sample eigenvalues are strictly positive under full rank. For each $i = 1, \dots, p$, QIS estimates the inverse covariance eigenvalue as

$$\begin{aligned} \tilde{d}_i^{-1} &= (1 - c)^2 q_i + 2c(1 - c) q_i \hat{\theta}_h(q_i) \\ &\quad + c^2 q_i \hat{A}_h^2(q_i). \end{aligned} \quad (\text{D.19})$$

The corresponding covariance eigenvalue is

$$\tilde{d}_i = \frac{1}{\tilde{d}_i^{-1}}. \quad (\text{D.20})$$

At the boundary $p = n$, the concentration ratio satisfies $c = 1$. Consequently, the terms proportional to $1 - c$ vanish and (D.19) reduces to

$$\tilde{d}_i^{-1} = q_i \hat{A}_h^2(q_i), \quad p = n. \quad (\text{D.21})$$

The separate singular formula is therefore required only when $p > n$.

The coefficients in (D.19) satisfy

$$(1 - c)^2 + 2c(1 - c) + c^2 = 1. \quad (\text{D.22})$$

When c is small, the inverse sample eigenvalue receives relatively more weight. As c approaches one, greater weight is assigned to the spectrum-dependent nonlinear adjustment because the sample covariance spectrum becomes less reliable. Unlike linear shrinkage toward a common matrix target, the QIS adjustment varies across eigenvalues because both $\hat{\theta}_h(q_i)$ and $\hat{A}_h^2(q_i)$ depend on the location of q_i within the sample spectrum.

D.5 Singular Case

When $p > n$, the sample covariance matrix has at least $p - n$ zero eigenvalues. The inverse of a zero sample eigenvalue is undefined, so QIS separates the null and non-null parts of the spectrum.

Let the positive sample eigenvalues be $\lambda_{p-n+1}, \dots, \lambda_p$ and define

$$\bar{q} = \frac{1}{n} \sum_{j=p-n+1}^p \lambda_j^{-1}. \quad (\text{D.23})$$

Each of the $p - n$ null sample eigenvalues is replaced by the common positive estimate

$$\tilde{d}_0 = \frac{1}{(c - 1)\bar{q}}. \quad (\text{D.24})$$

For each nonzero sample eigenvalue, the regularized covariance eigenvalue is

$$\tilde{d}_i = \frac{1}{q_i \hat{A}_h^2(q_i)}, \quad i = p - n + 1, \dots, p. \quad (\text{D.25})$$

The first two terms from (D.19) disappear in the singular case and the adjustment of the nonzero spectrum is driven by the squared-amplitude component. The zero sample eigenvalues are no longer interpreted as genuine zero-risk directions. They are assigned positive values determined by the concentration ratio and the harmonic behaviour of the nonzero sample eigenvalues.

This step is important for the empirical application. With 71 eligible ETFs and a 60-month estimation window, $p = 71$ and $n = 59$. The sample covariance matrix therefore contains at least 12 zero eigenvalues, whereas the QIS estimator assigns positive variance to all 71 directions.

D.6 Trace Preservation and Matrix Reconstruction

After the preliminary shrunk eigenvalues have been obtained, they are rescaled so that the QIS covariance matrix has the same trace as the sample covariance matrix. Define

$$a = \frac{\text{tr}(S)}{\sum_{i=1}^p \tilde{d}_i} = \frac{\sum_{i=1}^p \lambda_i}{\sum_{i=1}^p \tilde{d}_i}. \quad (\text{D.26})$$

The final regularized eigenvalues are

$$d_i^{\text{QIS}} = a \tilde{d}_i, \quad i = 1, \dots, p. \quad (\text{D.27})$$

The covariance estimator is reconstructed as

$$\hat{\Sigma}^{\text{QIS}} = U \text{diag} \left(d_1^{\text{QIS}}, \dots, d_p^{\text{QIS}} \right) U^\top. \quad (\text{D.28})$$

Trace preservation maintains the total sample variance,

$$\text{tr} \left(\hat{\Sigma}^{\text{QIS}} \right) = \text{tr}(S), \quad (\text{D.29})$$

while changing how that variance is distributed across the estimated risk directions.

D.7 Interpretation

QIS modifies the estimated amount of risk assigned to each sample eigenvector but does not change the eigenvectors themselves. It therefore remains within the rotation-equivariant class of covariance estimators. The estimator does not require an externally specified target matrix. Its regularization is instead determined by the observed eigenvalue spectrum, the relative position of each eigenvalue within that spectrum, and the concentration ratio p/n .

The procedure can be summarized in five steps:

1. estimate the sample covariance matrix;
2. obtain its eigenvalues and eigenvectors;
3. calculate local shrinkage quantities from the inverse nonzero eigenvalues;
4. replace the sample eigenvalues with positive regularized values;
5. reconstruct the covariance matrix using the original sample eigenvectors.

The estimator should not be interpreted as discovering the true factor structure of ETF returns. It regularizes the estimated covariance spectrum without labelling the eigenvectors as market, sector, duration, commodity, or other economic factors. Its purpose in this thesis is narrower: to reduce the optimizer's sensitivity to noisy and singular sample covariance matrices.

D.8 Python Implementation

The empirical implementation follows the Python reference implementation provided with the covariance-shrinkage code of Ledoit & Wolf (2022b). The relevant source file is `QIS.py` in the `pald22/covShrinkage` repository. The input is a data matrix with observations in rows and ETFs in columns. With the default setting, the function demeanes each return series and reduces the effective sample size by one. Its output is an invertible $p \times p$ covariance matrix.

The repository is distributed as source code rather than as a conventional installed Python package. For reproducibility, the QIS function is stored locally in the replication code, while the example file-loading and file-saving commands included at the bottom of the original script are removed. The retained statistical calculations are not otherwise modified. The exact source snapshot used in the analysis is recorded with the replication materials.

At each portfolio-formation date, the implementation follows the sequence

1. select the ETFs or sleeves required by the strategy at date t ;
2. extract their previous $W_{\Sigma} = 60$ monthly returns;
3. verify the balanced matrix implied by the eligibility rule and continuous post-inception histories;
4. pass the $W_{\Sigma} \times N_t$ return matrix to the QIS function;

5. restore the ETF names as row and column labels;
6. verify symmetry, finite entries, and positive eigenvalues;
7. pass the resulting matrix to the relevant portfolio optimizer.

Appendix E

Technical Portfolio Specifications

This appendix records the numerical and statistical details required to reproduce Chapter 2. The main chapter presents the equations needed to follow the empirical argument. The purpose here is different: to specify how constrained solutions are obtained, how recursive forecasts are stored, and how numerical failures are distinguished from genuine portfolio outcomes. Methods that are not part of the final strategy set are excluded.

E.1 Common Feasible Sets and Sleeve Mapping

At month t , let $\mathcal{U}_t$ be the eligible ETF universe and $N_t = |\mathcal{U}_t|$. All ETF-level optimization problems use

$$\mathcal{W}_t = \left\{ w_t : \mathbf{1}^\top w_t = 1, \quad 0 \leq w_{i,t} \leq w_{\max}, \quad w_{\max} = 0.20 \right\}. \quad (\text{E.1})$$

For sleeve-level strategies, x_t denotes the weights of the active sleeves $\mathcal{G}_t$. Equal weighting within sleeve g maps its sleeve allocation into ETF weights as

$$w_{i,t} = \frac{x_{g,t}}{N_{g,t}}, \quad i \in \mathcal{U}_{g,t}. \quad (\text{E.2})$$

The ETF cap therefore implies

$$0 \leq x_{g,t} \leq \bar{x}_{g,t} = \min\{1, w_{\max} N_{g,t}\}, \quad (\text{E.3})$$

and the sleeve feasible set is

$$\mathcal{X}_t = \left\{ x_t : \mathbf{1}^\top x_t = 1, \quad 0 \leq x_{g,t} \leq \bar{x}_{g,t} \right\}. \quad (\text{E.4})$$

This mapping makes the ETF- and sleeve-level models comparable. Every final allocation is long-only, fully invested, unlevered, and capped at 20% per ETF.

E.2 Numerical Solution of the Unconditional Models

Chapter 2 defines the four H1 portfolio models and their common feasible set. This section explains how those models are solved numerically and how candidate portfolios are validated. All models use the 60-month QIS covariance matrix, remain fully invested and long-only, and impose a maximum weight of 20% on each eligible ETF.

E.2.1 GMV and Mean–Variance Optimization

GMV and MVO are implemented in CVXPY as constrained quadratic optimization problems. Provided that the validated QIS covariance matrix is positive definite, both problems are convex and do not require a multiple-start procedure. CLARABEL is used as the primary solver. If it does not return a valid portfolio, the same problem is solved with OSQP as a predefined backup.

For MVO, the maximization problem in (2.27) is passed to the solver in the equivalent minimization form

$$\frac{\gamma}{2} w_t^\top \hat{\Sigma}_t^{\text{QIS}} w_t - w_t^\top \hat{\mu}_t^{(W_\mu)}. \quad (\text{E.5})$$

subject to the common ETF-level constraints. The baseline specification sets $W_\mu = 24$ and $\gamma = 5$. The expected-return window is changed to 12 or 48 months in the robustness specifications, while the 60-month QIS covariance estimator and all portfolio constraints remain unchanged.

Every returned portfolio is checked before it enters the backtest. The weights must be finite, sum to one within a strict numerical tolerance, remain non-negative, and satisfy the 20% ETF cap. The relevant objective is then recomputed from the returned weights and model inputs and must be finite. If neither CLARABEL nor OSQP produces a valid solution, the policy allocation is used as the predefined fallback for that strategy and month.

E.2.2 Equal Risk Contribution

The formal ERC objective is given in (2.24). Unlike GMV and MVO, the constrained ERC problem is nonlinear, and its numerical solution may depend on the initial portfolio. It is therefore solved with SciPy’s SLSQP algorithm from several feasible starting points. These include flat ETF-level equal weight, inverse-variance weights, the policy benchmark, the hierarchical equal-weight benchmark, and the previously accepted ERC portfolio. The benchmark allocations serve only as initial conditions for the numerical search and remain distinct benchmark strategies in the empirical analysis.

Any starting portfolio that does not satisfy the current universe and portfolio constraints is excluded. Each solution returned by SLSQP is then checked for finite weights, full investment, non-negativity, and compliance with the 20% ETF cap. The ERC objective in (2.24) is recomputed for every valid candidate, and the candidate with the smallest objective value is retained. If none of the starting points produces a valid solution, the policy allocation is used as the predefined fallback.

The feasibility of the capped portfolio is checked before optimization. Because no ETF may receive more than 20%, a fully invested portfolio requires at least five eligible ETFs. The growing-universe backtest begins only when this minimum and the other benchmark-feasibility conditions are satisfied.

E.2.3 Maximum Diversification

The diversification ratio is defined in (2.25). For numerical solution, maximizing this ratio is expressed as minimizing

$$-\log(w_t^\top \hat{\sigma}_t) + \frac{1}{2} \log(w_t^\top \hat{\Sigma}_t^{\text{QIS}} w_t), \quad (\text{E.6})$$

subject to $w_t \in \mathcal{W}_t$. The logarithmic transformation preserves the maximizing portfolio because the logarithm is strictly increasing, while providing a more convenient numerical scale for the optimization.

Maximum diversification is also solved with SLSQP from several feasible starting portfolios, including the simple benchmark allocations and the previously accepted maximum-diversification portfolio. The same general validation rules as for ERC are applied to each candidate. Accepted candidates are compared using the original diversification ratio rather than the transformed objec-

tive, and the portfolio with the highest ratio is retained. If no valid candidate is obtained, the policy portfolio is used as the predefined fallback.

E.3 Black–Litterman Implementation

The historical- and macro-view Black–Litterman strategies use identical code paths after their view vectors and forecast-error histories have been constructed. This restriction prevents the macro version from receiving a different prior, confidence rule, or optimizer.

E.3.1 Reference-Portfolio-Implied Returns

Let $w_t^{\text{ref}} = w_t^{\text{Policy}}$. The implied monthly excess-return vector and prior are

$$\Pi_t = \delta \hat{\Sigma}_t^{\text{QIS}} w_t^{\text{ref}}, \quad \delta = 2, \quad (\text{E.7})$$

$$\mu_t^e \sim \mathcal{N}(\Pi_t, \tau \hat{\Sigma}_t^{\text{QIS}}), \quad \tau = 0.1. \quad (\text{E.8})$$

Before these quantities are calculated, the reference weights are checked against the current ETF ordering and the budget and cap constraints. The implied-return vector is stored with ETF labels to prevent silent reordering when the universe grows.

E.3.2 Sleeve View Matrix and View Vectors

The matched comparison uses one common admissible set for both view types:

$$\mathcal{A}_t^{BL} = \left\{ g \in \mathcal{G}_t : n_{g,t}^H \geq 12 \text{ and } n_{g,t}^M \geq 12 \right\}. \quad (\text{E.9})$$

For every sleeve in $\mathcal{A}_t^{BL}$, row g of P_t is

$$P_{gi,t} = \begin{cases} 1/N_{g,t}, & i \in \mathcal{U}_{g,t}, \\ 0, & i \notin \mathcal{U}_{g,t}. \end{cases} \quad (\text{E.10})$$

The historical and macro views are

$$Q_{g,t}^{H,(W_\mu)} = \frac{1}{W_\mu} \sum_{\ell=1}^{W_\mu} \left(R_{g,t-\ell}^{(t)} - r_{f,t-\ell} \right), \quad (\text{E.11})$$

$$Q_{g,t}^{M,(W_\mu)} = \sum_{k=1}^3 \pi_{k,t|t-1} \tilde{\mu}_{g,k,t}^{e,(W_\mu)}. \quad (\text{E.12})$$

The baseline has $W_\mu = 24$; robustness uses 12 and 48 months. Both view types always use the same W_μ within a comparison.

E.3.3 Forecast-Error Reliability

After view $Q_{g,\tau}^v$ has been formed and month τ has been realized, the forecast error is

$$e_{g,\tau}^v = Q_{g,\tau}^v - \left(R_{g,\tau}^{(\tau)} - r_{f,\tau} \right). \quad (\text{E.13})$$

Only errors produced at their original historical formation dates are stored. They are never recomputed using a later sleeve composition, later HMM estimate, or revised macro vintage. Let $n_{g,t}^v$ be the number of stored errors available before date t and define

$$L_{g,t}^v = \min\{36, n_{g,t}^v\}. \quad (\text{E.14})$$

A sleeve is admitted to the common P_t only when both $n_{g,t}^H \geq 12$ and $n_{g,t}^M \geq 12$. Conditional on that common row set, each version retains its own view vector and uncertainty estimate. Its uncertainty is

$$\Omega_{gg,t}^v = \max \left\{ \frac{1}{L_{g,t}^v - 1} \sum_{\ell=1}^{L_{g,t}^v} \left(e_{g,t-\ell}^v - \bar{e}_{g,t}^v \right)^2, 10^{-8} \right\}, \quad (\text{E.15})$$

where $\bar{e}_{g,t}^v$ is the mean of the selected error history. Off-diagonal elements are zero. The floor protects the matrix inverse when the recent error variance is numerically negligible.

E.3.4 Posterior and Final Portfolio

The posterior excess-return vector is

$$\hat{\mu}_t^{BL,v} = \left[(\tau \hat{\Sigma}_t^{\text{QIS}})^{-1} + P_t^\top (\Omega_t^v)^{-1} P_t \right]^{-1} \left[(\tau \hat{\Sigma}_t^{\text{QIS}})^{-1} \Pi_t + P_t^\top (\Omega_t^v)^{-1} Q_t^v \right]. \quad (\text{E.16})$$

The implementation solves linear systems rather than constructing explicit matrix inverses whenever the numerical library permits. The final allocation solves

$$w_t^{BL,v} = \arg \max_{w_t \in \mathcal{W}_t} \left\{ w_t^\top \hat{\mu}_t^{BL,v} - \frac{\delta}{2} w_t^\top \hat{\Sigma}_t^{\text{QIS}} w_t \right\}. \quad (\text{E.17})$$

The stored diagnostics include the number of active views, each view and uncertainty, the distance between posterior and implied returns, the distance between final and reference weights, and the condition number of the linear system.

E.4 Black–Litterman Performance

The weak baseline Black–Litterman results require separate interpretation because the empirical design was partly motivated by the favourable evidence in Bessler *et al.* (2017). The comparison is informative, but it is not a replication. It also does not change the frozen H2 assessment or justify selecting a specification after observing its performance.

E.4.1 Absolute Performance Across Specifications

The matched-difference evidence in Table 3.5 isolates the incremental effect of replacing historical views with macro views. It does not show whether the absolute performance of the two Black–Litterman portfolios changes across specifications. Table E.1 reports those levels separately.

The absolute results qualify the especially weak 24-month baseline. With 12-month views, the Sharpe ratios rise to 0.500 for BL-H and 0.508 for BL-M, final wealth approaches 3.0, and drawdowns become less severe. The 48-month versions also exceed the baseline Sharpe ratios, although BL-M then underperforms BL-H. Thus, the architecture is not uniformly unsuccessful, but its performance is highly sensitive to the horizon used to estimate expected returns. The 12-month result is economically relevant because it matches the view window used by Bessler *et al.* (2017); it remains a pre-specified robustness result rather than an alternative baseline.

Table E.1: Absolute Black–Litterman performance across pre-specified variants

Specification	Strategy	Return	Volatility	Sharpe	Drawdown	Wealth	Turnover
Baseline (24m)	BL-H	5.58%	13.93%	0.282	-27.88%	1.930	2.917
Baseline (24m)	BL-M	5.79%	13.87%	0.298	-28.10%	1.992	2.990
12m robustness	BL-H	8.57%	13.80%	0.500	-23.23%	2.969	3.701
12m robustness	BL-M	8.70%	13.85%	0.508	-23.11%	3.023	3.677
48m robustness	BL-H	7.52%	13.87%	0.423	-29.73%	2.552	2.694
48m robustness	BL-M	6.91%	13.67%	0.384	-27.11%	2.348	2.864
Financial-only diagnostic	BL-H	5.58%	13.93%	0.282	-27.88%	1.930	2.917
Financial-only diagnostic	BL-M	5.97%	13.90%	0.310	-27.76%	2.041	3.043

Note: March 2012–June 2026, 172 holding months. Return and volatility are annualized; turnover is annualized one-way turnover. The 12- and 48-month rows change the expected-return window. The financial-only HMM is an additional diagnostic and leaves BL-H unchanged by construction.

Source: Author’s calculations from the frozen baseline, expected-return-window robustness, and financial-only H2 contracts.

E.4.2 Comparison with Bessler, Opfer, and Wolff

Bessler *et al.* (2017) report strong out-of-sample results for a sample-based Black–Litterman allocation. For their moderate investor, the Black–Litterman portfolio earns a net return of 9.58%, volatility of 8.65%, a net Sharpe ratio of 0.75, and a maximum drawdown of 9.46%. The matched mean–variance portfolio has a Sharpe ratio of 0.49 and a drawdown of 16.18%. Their result supports the usefulness of combining strategic implied returns with reliability-weighted views, but several design differences affect comparability.

The paper’s strong performance is primarily a shrinkage and risk-control result rather than evidence of highly accurate raw return forecasts. Its mixed Black–Litterman estimates reduce forecast mean-squared errors relative to the 12-month means, while the strategic prior, forecast-error reliability, posterior covariance, and volatility ceiling limit extreme allocations. The opportunity set also contains only five broad indices. By contrast, the present design maps up to 15 sleeve views into as many as 71 ETFs and deliberately retains the QIS covariance so that the matched comparison isolates the expected-return channel. It therefore asks a more demanding and different question.

E.4.3 Sources of Baseline Underperformance

Diagnostics derived from the frozen H2 outputs indicate that weak predictive information is the first limitation. Across active sleeve-month observations, both historical and macro views have a next-month forecast correlation close

Table E.2: Black–Litterman design comparison with Bessler, Opfer, and Wolff

Dimension	Bessler <i>et al.</i> (2017)	Present thesis
Investment universe	Five broad, non-overlapping stock, government-bond, high-yield-bond, and commodity indices	Growing universe of up to 71 ETFs organized into 15 partly overlapping economic sleeves
Evaluation period	January 1993–December 2011	March 2012–June 2026
Historical views	12-month mean returns	24-month baseline; 12- and 48-month robustness variants
Covariance input	36-month sample covariance	60-month ETF-level QIS covariance
Prior and confidence	Strategic weights, $\delta = 2$, $\tau = 0.1$, and diagonal forecast-error uncertainty	Policy weights, $\delta = 2$, $\tau = 0.1$, and diagonal uncertainty from 12–36 genuine prior errors
Strategic risk allocation	Moderate prior: 45% bonds, 40% stocks, and 15% commodities	Policy prior: 20% Treasuries, 60% equity, 15% commodities, and 5% real estate
Risk control	No short sales and a 10% expected-volatility ceiling for the moderate investor	No short sales and a 20% ETF cap, but no portfolio-volatility ceiling
Final risk matrix	Black–Litterman posterior covariance	Original ETF-level QIS covariance; Black–Litterman changes expected returns only
Primary comparison	Black–Litterman against MVO and static benchmarks	BL-M against matched BL-H; broader rankings are secondary

Source: Author’s summary of Bessler *et al.* (2017) and the present empirical design.

to zero, at approximately -0.05 and -0.04 , respectively. The two forecast systems are themselves almost identical: their correlation is 0.992 and their mean absolute difference is about 0.05 percentage points per month. This similarity explains why BL-M differs only modestly from BL-H and why the sign of the matched increment changes under the 48-month window.

The second limitation is the transmission from expected returns to portfolio weights. Although forecast-error uncertainty shrinks the sleeve views toward the prior, the constrained ETF optimizer produces an average one-way distance from Policy of approximately 79% for both Black–Litterman variants. About 73% of eligible ETF weights are numerically zero, average effective holdings fall to approximately 8.85 compared with 24.64 for Policy, and annualized turnover approaches 3.0. These departures reduce annualized return relative to Policy by approximately 1.96 percentage points for BL-H and 1.74 percentage points for BL-M while adding substantial active volatility.

Finally, diagonal Ω_t^v is a necessary simplification under short forecast-error histories, but it does not capture dependence between overlapping views. The mean pairwise correlation of sleeve forecast errors is approximately 0.54, and several broad, regional, factor, and sector equity pairs exceed 0.95. Treating these errors as independent may overstate the joint information contained in economically similar views. The resulting portfolios tend, on average, to over-

weight sector sleeves and underweight broad equity and Treasury sleeves relative to Policy. In the realized 2012–2026 sample, those active positions did not provide sufficient return compensation for their concentration and instability.

The comparison therefore reconciles the initial optimism drawn from Bessler *et al.* (2017) with the present evidence. Black–Litterman performs favourably in their low-dimensional, volatility-constrained, bond-heavy setting and improves materially in the present 12-month robustness specification. The frozen baseline nevertheless performs poorly because weak and highly correlated sleeve forecasts are translated into large ETF-level reallocations without a portfolio-volatility ceiling or posterior-covariance risk adjustment. This is a design- and sample-specific result, not evidence that the Black–Litterman framework is generally ineffective.

E.5 Three-State Hidden Markov Model

The regime engine uses the standardized four-variable vector defined in (2.31) and (2.32). At portfolio date t , all parameters below are estimated from the expanding sample ending at $t - 1$.

For state $k \in \{1, 2, 3\}$, the conditional density is

$$f_k(Z_\tau^{(t)}) = \phi(Z_\tau^{(t)}; m_{k,t}, V_{k,t}), \quad V_{k,t} = \text{diag}(v_{1k,t}, \dots, v_{4k,t}), \quad (\text{E.18})$$

where $\phi(\cdot)$ is the multivariate Gaussian density. Let $q_{0,t}$ be the initial state-probability vector and $A_t = [a_{jk,t}]$ the transition matrix.

E.5.1 Scaled Forward Recursion and Likelihood

The unnormalized forward probability is

$$\tilde{\alpha}_{k,\tau} = f_k(Z_\tau^{(t)}) \sum_{j=1}^3 \alpha_{j,\tau-1} a_{jk,t}. \quad (\text{E.19})$$

Define the scaling factor

$$c_\tau = \sum_{h=1}^3 \tilde{\alpha}_{h,\tau}. \quad (\text{E.20})$$

The normalized filtered probability is

$$\alpha_{k,\tau} = \frac{\tilde{\alpha}_{k,\tau}}{c_\tau}, \quad (\text{E.21})$$

and the log likelihood is

$$\log \mathcal{L}_t = \sum_{\tau} \log(c_\tau). \quad (\text{E.22})$$

Scaling avoids numerical underflow when many small probabilities are multiplied over a long sample.

E.5.2 Backward Recursion and Smoothed Probabilities

Starting from a terminal vector of ones, the scaled backward recursion is

$$\beta_{j,\tau} = \frac{1}{c_{\tau+1}} \sum_{k=1}^3 a_{jk,t} f_k \left(Z_{\tau+1}^{(t)} \right) \beta_{k,\tau+1}. \quad (\text{E.23})$$

The smoothed state probability is

$$\eta_{\tau k}^{(t)} = \frac{\alpha_{k,\tau} \beta_{k,\tau}}{\sum_{h=1}^3 \alpha_{h,\tau} \beta_{h,\tau}}. \quad (\text{E.24})$$

The smoothed transition probability from state j to state k is

$$\xi_{jk,\tau}^{(t)} = \frac{\alpha_{j,\tau} a_{jk,t} f_k \left(Z_{\tau+1}^{(t)} \right) \beta_{k,\tau+1}}{\sum_{r=1}^3 \sum_{s=1}^3 \alpha_{r,\tau} a_{rs,t} f_s \left(Z_{\tau+1}^{(t)} \right) \beta_{s,\tau+1}}. \quad (\text{E.25})$$

In the recursive backtest, the terminal observation in these formulas is always month $t-1$. No observation after the formation date enters the smoothing step.

E.5.3 Expectation–Maximization Updates

Given the smoothed probabilities, the M-step updates are

$$q_{0,k,t}^{new} = \eta_{1k}^{(t)}, \quad (\text{E.26})$$

$$a_{jk,t}^{new} = \frac{\sum_{\tau} \xi_{jk,\tau}^{(t)}}{\sum_{\tau} \eta_{\tau j}^{(t)}}, \quad (\text{E.27})$$

$$m_{k,t}^{new} = \frac{\sum_{\tau} \eta_{\tau k}^{(t)} Z_{\tau}^{(t)}}{\sum_{\tau} \eta_{\tau k}^{(t)}}, \quad (\text{E.28})$$

$$v_{jk,t}^{new} = \frac{\sum_{\tau} \eta_{\tau k}^{(t)} (Z_{j,\tau}^{(t)} - m_{jk,t}^{new})^2}{\sum_{\tau} \eta_{\tau k}^{(t)}}. \quad (\text{E.29})$$

A small positive floor is applied to every state variance before the next density evaluation. Each transition row is normalized to sum to one. Iteration continues until the log-likelihood improvement is below the pre-specified tolerance or the maximum iteration count is reached.

At each formation date, 20 initial parameterizations are evaluated. Initial state means are generated from partitions or clustered versions of the standardized observations, transition matrices begin with positive diagonal persistence, and all starts satisfy valid probability constraints. The converged solution with the highest finite log likelihood is retained. The state labels are then assigned with (2.35); labels do not affect likelihood estimation.

The one-step-ahead probability is calculated from the filtered final-state vector and the retained transition matrix. The code verifies that all probabilities lie in $[0, 1]$, each probability vector sums to one, each transition row sums to one, and all state variances are positive.

E.6 Regime-Conditioned Sleeve Moments

At date t , the current constituents of each active sleeve are held fixed when reconstructing historical sleeve returns. Expected-return windows take values $W_{\mu} \in \{12, 24, 48\}$, with 24 as the baseline. The covariance window is always $W_{\Sigma} = 60$.

For a generic window W , define

$$G_{k,t}(W) = \sum_{\tau=t-W}^{t-1} \eta_{\tau k}^{(t)}, \quad (\text{E.30})$$

$$M_{k,t}^{eff}(W) = \frac{G_{k,t}(W)^2}{\sum_{\tau=t-W}^{t-1} \left(\eta_{\tau k}^{(t)}\right)^2}. \quad (\text{E.31})$$

The conditional expected return is

$$\hat{\mu}_{k,t}^{(W_\mu)} = \frac{\sum_{\tau=t-W_\mu}^{t-1} \eta_{\tau k}^{(t)} R_{\mathcal{G}_t, \tau}^{(t)}}{G_{k,t}(W_\mu)}. \quad (\text{E.32})$$

For covariance centring, define the 60-month conditional mean

$$\hat{\mu}_{k,t}^{(60,c)} = \frac{\sum_{\tau=t-60}^{t-1} \eta_{\tau k}^{(t)} R_{\mathcal{G}_t, \tau}^{(t)}}{G_{k,t}(60)}. \quad (\text{E.33})$$

The unequal-weight degrees of freedom are

$$D_{k,t}(60) = G_{k,t}(60) - \frac{\sum_{\tau=t-60}^{t-1} \left(\eta_{\tau k}^{(t)}\right)^2}{G_{k,t}(60)}, \quad (\text{E.34})$$

and the conditional covariance is

$$S_{k,t}^r = \frac{1}{D_{k,t}(60)} \sum_{\tau=t-60}^{t-1} \eta_{\tau k}^{(t)} \left(R_{\mathcal{G}_t, \tau}^{(t)} - \hat{\mu}_{k,t}^{(60,c)}\right) \left(R_{\mathcal{G}_t, \tau}^{(t)} - \hat{\mu}_{k,t}^{(60,c)}\right)^\top. \quad (\text{E.35})$$

Let $\hat{\mu}_t^{U, (W_\mu)}$ be the unconditional sleeve mean and $\hat{\Sigma}_t^{U, \text{QIS}}$ the 60-month unconditional sleeve QIS covariance. With $\kappa = 36$,

$$\lambda_{k,t}^\mu = \frac{M_{k,t}^{eff}(W_\mu)}{M_{k,t}^{eff}(W_\mu) + 36}, \quad (\text{E.36})$$

$$\lambda_{k,t}^\Sigma = \frac{M_{k,t}^{eff}(60)}{M_{k,t}^{eff}(60) + 36}. \quad (\text{E.37})$$

The pooled moments are

$$\tilde{\mu}_{k,t}^{(W_\mu)} = \lambda_{k,t}^\mu \hat{\mu}_{k,t}^{(W_\mu)} + (1 - \lambda_{k,t}^\mu) \hat{\mu}_t^{U,(W_\mu)}, \quad (\text{E.38})$$

$$\tilde{\Sigma}_{k,t} = \lambda_{k,t}^\Sigma S_{k,t}^r + (1 - \lambda_{k,t}^\Sigma) \hat{\Sigma}_t^{U,\text{QIS}}. \quad (\text{E.39})$$

The predictive moments are

$$\hat{\mu}_t^{RS,(W_\mu)} = \sum_{k=1}^3 \pi_{k,t|t-1} \tilde{\mu}_{k,t}^{(W_\mu)}, \quad (\text{E.40})$$

$$\hat{\Sigma}_t^{RS} = \sum_{k=1}^3 \pi_{k,t|t-1} \left[\tilde{\Sigma}_{k,t} + \left(\tilde{\mu}_{k,t}^{(W_\mu)} - \hat{\mu}_t^{RS,(W_\mu)} \right) \left(\tilde{\mu}_{k,t}^{(W_\mu)} - \hat{\mu}_t^{RS,(W_\mu)} \right)^\top \right]. \quad (\text{E.41})$$

The excess-return version used for macro Black–Litterman applies the same formulas to $R_{g,\tau}^{(t)} - r_{f,\tau}$. The regime-conditioned sleeve allocation then solves (2.64), and the unconditional control replaces the predictive inputs with the corresponding unconditional mean and QIS covariance.

For every date and state, the code stores $G_{k,t}(W)$, $M_{k,t}^{eff}(W)$, both reliability weights, the minimum eigenvalue of each pooled covariance, and the contribution of the between-state term to total predictive variance.

E.7 Random DCA and Retail Wealth Accounting

Random DCA is simulated only in H3. The first review is anchored at the common evaluation start, 31 March 2012, and subsequent reviews occur every six holding months. Let b index these semiannual review dates and let $N_b = |\mathcal{U}_b|$. The number of selected ETFs is

$$m_b = \min \left\{ N_b, \max \left\{ 5, \left\lfloor \frac{N_b}{3} + \frac{1}{2} \right\rfloor \right\} \right\}. \quad (\text{E.42})$$

For simulation path q , the subset $\mathcal{S}_b^{(q)} \subseteq \mathcal{U}_b$ is drawn uniformly without replacement, with $|\mathcal{S}_b^{(q)}| = m_b$. Review-date target weights are

$$w_{i,b}^{RDA,q} = \begin{cases} 1/m_b, & i \in \mathcal{S}_b^{(q)}, \\ 0, & i \notin \mathcal{S}_b^{(q)}. \end{cases} \quad (\text{E.43})$$

At the review date, holdings are traded to these equal weights after the new

contribution is added. During the following five non-review months, existing holdings drift and no sale is made merely to restore equal weights. Because the contribution C is gross cash and purchases incur proportional cost c , net ETF purchases equal $C/(1+c)$ and are divided equally among the selected ETFs:

$$C_{i,t}^{RDA,q} = \begin{cases} C/[m_b(1+c)], & i \in \mathcal{S}_b^{(q)}, \\ 0, & i \notin \mathcal{S}_b^{(q)}. \end{cases} \quad (\text{E.44})$$

The corresponding purchase cost is $cC/(1+c)$, so net purchases and cost exhaust the gross contribution exactly. Contribution purchases and review-date trades incur the five-basis-point cost. A new subset is drawn independently at the next review. A root seed of 1508200 is passed to `numpy.random.SeedSequence`, which generates one deterministic child seed per path. All 1,000 paths use the same calendar, eligible universes, return panel, contribution schedule, and cost coefficient.

For a monthly target-weight strategy, post-cost wealth solves

$$V_{s,t}^+ = A_{s,t} - 0.0005 \sum_{i \in \mathcal{U}_t} |w_{i,t}^s V_{s,t}^+ - H_{i,s,t}^-|. \quad (\text{E.45})$$

Fixed-point iteration begins at $V_{s,t}^{+(0)} = A_{s,t}$ and stops when successive values differ by less than 10^{-10} contribution units or after 100 iterations. If convergence is not achieved, a bisection solution is calculated on $[0, A_{s,t}]$. Final holdings plus transaction cost must equal available wealth within tolerance.

The gross contribution-based counterfactual applies the same target weights and cash flows with the cost coefficient set to zero. Random DCA is not forced into a monthly target-weight recursion between reviews; costs are applied only to its actual purchases and review trades. This difference is intentional because the benchmark represents a lower-maintenance investment rule.

E.7.1 External VT Reference

The Vanguard Total World Stock ETF (VT) The Vanguard Group (2026) is analyzed separately from the 71-ETF multi-asset universe. The reference uses a frozen USD adjusted-price series and is reported over the standalone H1 and common evaluation periods. Its frictionless path holds one unit of VT continuously. Its retail path uses the same initial wealth, beginning-of-month contributions, fractional-share assumption, and proportional purchase cost as H3, but it has no strategic rebalancing and no terminal liquidation. A dedicated mani-

fest records the source, download timestamp, adjusted-price convention, input hashes, exact period coverage, and output hashes. VT is external contextual evidence and does not enter the formal assessment of Hypotheses 1–3.

E.8 Stationary-Bootstrap Inference Results

Table E.3 reports all eight H1 and two H2 paired Sharpe-ratio intervals under the primary 12-month expected block length. Table E.4 reports the same intervals under expected block lengths of 6, 12, and 24 months. These exhibits describe conditional sampling uncertainty in the frozen strategy-return paths and do not re-estimate the underlying portfolio models.

Table E.3: Pairwise stationary-bootstrap Sharpe-ratio differences

Family	Comparison	$\hat{\Delta}$	95% interval	Positive	Raw p
H1	GMV – Policy	+0.199	[-0.142, +0.686]	84.63%	0.187
H1	GMV – HEW	+0.202	[-0.151, +0.684]	84.26%	0.189
H1	ERC – Policy	+0.127	[+0.041, +0.209]	99.81%	0.001
H1	ERC – HEW	+0.130	[+0.026, +0.231]	99.34%	0.008
H1	MD – Policy	+0.128	[-0.137, +0.490]	79.98%	0.220
H1	MD – HEW	+0.131	[-0.140, +0.491]	80.26%	0.217
H1	MVO – Policy	-0.041	[-0.285, +0.200]	35.26%	0.623
H1	MVO – HEW	-0.039	[-0.291, +0.213]	36.00%	0.613
H2	BL-M – BL-H	+0.016	[-0.020, +0.065]	74.25%	0.206
H2	Sleeve-MVO-RC – Sleeve-MVO-U	+0.058	[+0.006, +0.121]	98.88%	0.038

Note: Expected block length of 12 months and 10,000 joint replications. Positive is the share of pseudo-histories with a positive paired difference. The reported intervals are percentile intervals; raw p -values use the centred one-sided bootstrap formula. *Source:* Author’s calculations from the frozen H1 and H2 strategy-return paths.

Table E.4: Stationary-bootstrap sensitivity to expected block length

Family	Comparison	$\hat{\Delta}$	$L = 6$	$L = 12$	$L = 24$
H1	GMV – Policy	+0.199	[-0.132, +0.638]	[-0.142, +0.686]	[-0.156, +0.710]
H1	GMV – HEW	+0.202	[-0.131, +0.640]	[-0.151, +0.684]	[-0.169, +0.722]
H1	ERC – Policy	+0.127	[+0.028, +0.221]	[+0.041, +0.209]	[+0.052, +0.199]
H1	ERC – HEW	+0.130	[+0.016, +0.239]	[+0.026, +0.231]	[+0.035, +0.223]
H1	MD – Policy	+0.128	[-0.148, +0.469]	[-0.137, +0.490]	[-0.120, +0.483]
H1	MD – HEW	+0.131	[-0.145, +0.463]	[-0.140, +0.491]	[-0.127, +0.488]
H1	MVO – Policy	-0.041	[-0.318, +0.227]	[-0.285, +0.200]	[-0.270, +0.171]
H1	MVO – HEW	-0.039	[-0.325, +0.241]	[-0.291, +0.213]	[-0.269, +0.185]
H2	BL-M – BL-H	+0.016	[-0.021, +0.065]	[-0.020, +0.065]	[-0.017, +0.061]
H2	RC – U	+0.058	[+0.007, +0.121]	[+0.006, +0.121]	[+0.008, +0.114]

Note: Cells report 95% percentile intervals from 10,000 joint stationary-bootstrap replications. Block lengths were fixed before inspecting the results. *Source:* Author’s calculations from the frozen H1 and H2 strategy-return paths.

E.9 Transaction-Cost Sensitivity

Table E.5 reports deterministic contribution-based outcomes under proportional trading costs of 5, 25, and 50 basis points. All scenarios use the same frozen strategy targets, return history, contribution schedule, and fixed-point accounting procedure. The additional rates are deliberately conservative scenarios rather than estimates of one universal historical retail tariff. Random DCA is supplementary to the H3 support rule and is not rerun for these cost scenarios.

Table E.5: Sensitivity of retail outcomes to proportional transaction costs

<i>Panel A: terminal net wealth</i>				
Strategy	5 bp	25 bp	50 bp	$\Delta(50 \text{ bp} - 5 \text{ bp})$
Policy	381.464	378.329	374.459	-7.005
HEW	393.718	390.505	386.537	-7.181
GMV	305.868	299.358	291.471	-14.398
ERC	421.342	415.144	407.570	-13.772
MD	316.497	309.816	301.720	-14.777
MVO	421.963	388.390	351.082	-70.881
BL-H	312.193	283.290	251.839	-60.354
BL-M	319.459	289.106	256.208	-63.251
Sleeve-MVO-U	386.221	361.187	332.608	-53.613
Sleeve-MVO-RC	419.592	391.665	359.843	-59.749

<i>Panel B: net time-weighted return and terminal cost drag</i>						
Strategy	Net TWR			Cost drag		
	5 bp	25 bp	50 bp	5 bp	25 bp	50 bp
Policy	7.13%	7.01%	6.86%	0.789	3.924	7.794
HEW	7.30%	7.19%	7.04%	0.809	4.022	7.990
GMV	5.77%	5.50%	5.17%	1.656	8.167	16.054
ERC	8.52%	8.30%	8.02%	1.570	7.768	15.342
MD	5.70%	5.43%	5.09%	1.699	8.379	16.476
MVO	8.16%	7.14%	5.87%	8.963	42.536	79.844
BL-H	4.39%	3.19%	1.72%	7.808	36.712	68.162
BL-M	4.62%	3.38%	1.87%	8.218	38.571	71.469
Sleeve-MVO-U	7.61%	6.83%	5.87%	6.583	31.616	60.195
Sleeve-MVO-RC	8.50%	7.71%	6.72%	7.351	35.278	67.100

Note: March 2012–June 2026. All scenarios replay identical frozen target weights and monthly returns. The 25- and 50-basis-point rates are conservative proportional-cost scenarios, not estimates of one universal retail tariff. Random DCA is not rerun. *Source:* Author’s calculations from the frozen common-period target weights and return paths.

E.10 Numerical Validation and Reproducibility

The numerical diagnostics in Table E.6 distinguish economic weakness from computational failure. Every stored H1 and H2 optimizer solution passes its budget and bound checks, and no optimizer fallback enters the baseline results. All stored QIS estimates are finite, symmetric, and positive definite.

The recursive HMM produces 184 valid formation-date fits, each evaluated from 20 deterministic starts, without a baseline fallback. H3, the 1,000 Random DCA paths, and VT all satisfy their post-cost budget identities within numerical tolerance and contain no accounting or MWR fallback. The weak Black–Litterman and selected H3 outcomes therefore reflect empirical performance rather than omitted months or numerical substitution.

Table E.6: Numerical diagnostics for the baseline empirical results

Component	Scope	Failures/fallbacks	Extremal diagnostic
H1 optimizers	796	0	max budget residual 1.74e-13
H2 optimizers	736	0	max budget residual 5.18e-13
QIS covariances	751	0	min eigenvalue 1.20e-08
Recursive HMM	184	0	3 states; 20 starts per formation
H3 target strategies	10	0	max budget residual 5.68e-14
Random DCA	1000	0	max budget residual 2.27e-13
VT external reference	172	0	max budget residual 0.00e+00

Note: Scope denotes optimizer attempts, covariance estimates, HMM formation dates, strategies, paths, or months as appropriate. *Source:* Author’s calculations from the frozen optimizer, covariance, HMM, H3, Random DCA, and VT diagnostic contracts.

Every monthly run stores the data boundary, eligible universe, active sleeves, expected-return window, covariance window, ETF and sleeve input matrices, covariance eigenvalues, optimization status, constraint residuals, HMM likelihood, transition matrix, state parameters, filtered and predicted probabilities, effective regime samples, pooling weights, target weights, trades, and transaction costs.

The following checks are completed before a monthly result is accepted:

1. the ETF and sleeve return matrices match the intended dates, contain finite values, and have the dimensions implied by $\mathcal{U}_t$ and $\mathcal{G}_t$;
2. every macro input is associated with a release or vintage available by the formation date;
3. each QIS covariance is symmetric, finite, and strictly positive definite;
4. portfolio weights sum to one within 10^{-8} and satisfy every ETF or sleeve bound;
5. the HMM transition rows and probability vectors sum to one and all probabilities lie in $[0, 1]$;

6. HMM variances are positive, the retained likelihood is finite, and the economic state labels are assigned from the current estimated state means;
7. regime effective sample sizes and reliability weights are finite and lie in their theoretical ranges;
8. stored Black–Litterman forecast errors were generated before the current formation date and use the sleeve composition and macro vintage available at their original date;
9. post-trade holdings plus transaction costs equal available wealth within numerical tolerance; and
10. every fallback month is retained in the output rather than omitted from performance calculations.

These checks are part of the pre-specified empirical procedure. They prevent data-alignment and solver failures from being silently converted into apparent strategy performance and provide an audit trail for the results reported in Chapter 3.